Yiyang Liu,James Chenhao Liang,Ruixiang Tang,Yugyung Lee,MAJID RABBANI,Sohail Dianat,Raghuveer Rao,Lifu Huang,Dongfang Liu,Qifan Wang,Cheng Han
Re-Imagining Multimodal Instruction Tuning: A Representation View
Multimodal instruction tuning has proven to be an effective strategy for achieving zero-shot generalization by fine-tuning pre-trained Large Multimodal Models (LMMs) with instruction-following data. However, as the scale of LMMs continues to grow, fully fine-tuning these models has become highly parameter-intensive. Although Parameter-Efficient Fine-Tuning (PEFT) methods have been introduced to reduce the number of tunable parameters, a significant performance gap remains compared to full fine-tuning. Furthermore, existing PEFT approaches are often highly parameterized, making them difficult to interpret and control. In light of this, we introduce Multimodal Representation Tuning (MRT), a novel approach that focuses on directly editing semantically rich multimodal representations to achieve strong performance and provide intuitive control over LMMs. Empirical results show that our method surpasses current state-of-the-art baselines with significant performance gains (e.g., 1580.40 MME score) while requiring substantially fewer tunable parameters (e.g., 0.03% parameters). Additionally, we conduct experiments on editing instrumental tokens within multimodal representations, demonstrating that direct manipulation of these representations enables simple yet effective control over network behavior.
****
Dominik Straub,Tobias F. Niehues,Jan Peters,Constantin A. Rothkopf
Inverse decision-making using neural amortized Bayesian actors
Bayesian observer and actor models have provided normative explanations for many behavioral phenomena in perception, sensorimotor control, and other areas of cognitive science and neuroscience. They attribute behavioral variability and biases to interpretable entities such as perceptual and motor uncertainty, prior beliefs, and behavioral costs. However, when extending these models to more naturalistic tasks with continuous actions, solving the Bayesian decision-making problem is often analytically intractable. Inverse decision-making, i.e. performing inference over the parameters of such models given behavioral data, is computationally even more difficult. Therefore, researchers typically constrain their models to easily tractable components, such as Gaussian distributions or quadratic cost functions, or resort to numerical approximations. To overcome these limitations, we amortize the Bayesian actor using a neural network trained on a wide range of parameter settings in an unsupervised fashion. Using the pre-trained neural network enables performing efficient gradient-based Bayesian inference of the Bayesian actor model's parameters. We show on synthetic data that the inferred posterior distributions are in close alignment with those obtained using analytical solutions where they exist. Where no analytical solution is available, we recover posterior distributions close to the ground truth. We then show how our method allows for principled model comparison and how it can be used to disentangle factors that may lead to unidentifiabilities between priors and costs. Finally, we apply our method to empirical data from three sensorimotor tasks and compare model fits with different cost functions to show that it can explain individuals' behavioral patterns.
****
Ashish Kumar,Jaesik Park
Designing Concise ConvNets with Columnar Stages
In the era of vision Transformers, the recent success of VanillaNet shows the huge
potential of simple and concise convolutional neural networks (ConvNets). Where such models mainly focus on runtime, it is also crucial to simultaneously focus on other aspects, e.g., FLOPs, parameters, etc, to strengthen their utility further.
To this end, we introduce a refreshing ConvNet macro design called Columnar Stage Network (CoSNet). CoSNet has a systematically developed simple and concise structure, smaller depth, low parameter count, low FLOPs, and attentionless operations, well suited for resource-constrained deployment. The key novelty

of CoSNet is deploying parallel convolutions with fewer kernels fed by input replication, using columnar stacking of these convolutions, and minimizing the use of 1×1 convolution layers. Our comprehensive evaluations show that CoSNet rivals many renowned ConvNets and Transformer designs under resource-constrained scenarios. Pretrained models shall be open-sourced.
****
Hanzhuo Huang,Yuan Liu,Ge Zheng,Jiepeng Wang,Zhiyang Dou,Sibei Yang
MVTokenFlow: High-quality 4D Content Generation using Multiview Token Flow
In this paper, we present MVTokenFlow for high-quality 4D content creation from monocular videos. Recent advancements in generative models such as video diffusion models and multiview diffusion models enable us to create videos or 3D models. However, extending these generative models for dynamic 4D content creation is still a challenging task that requires the generated content to be consistent spatially and temporally. To address this challenge, MVTokenFlow utilizes the multiview diffusion model to generate multiview images on different timesteps, which attains spatial consistency across different viewpoints and allows us to reconstruct a reasonable coarse 4D field. Then, MVTokenFlow further regenerates all the multiview images using the rendered 2D flows as guidance. The 2D flows effectively associate pixels from different timesteps and improve the temporal consistency by reusing tokens in the regeneration process. Finally, the regenerated images are spatiotemporally consistent and utilized to refine the coarse 4D field to get a high-quality 4D field. Experiments demonstrate the effectiveness of our design and show significantly improved quality than baseline methods. Project page: https://soolab.github.io/MVTokenFlow.
****
ZhongXiang Sun,Xiaoxue Zang,Kai Zheng,Jun Xu,Xiao Zhang,Weijie Yu,Yang Song,Han Li
ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability
Retrieval-Augmented Generation (RAG) models are designed to incorporate external knowledge, reducing hallucinations caused by insufficient parametric (internal) knowledge. However, even with accurate and relevant retrieved content, RAG models can still produce hallucinations by generating outputs that conflict with the retrieved information. Detecting such hallucinations requires disentangling how Large Language Models (LLMs) balance external and parametric knowledge. Current detection methods often focus on one of these mechanisms or without decoupling their intertwined effects, making accurate detection difficult. In this paper, we investigate the internal mechanisms behind hallucinations in RAG scenarios. We discover hallucinations occur when the **Knowledge FFNs** in LLMs overemphasize parametric knowledge in the residual stream, while **Copying Heads** fail to effectively retain or integrate external knowledge from retrieved content. Based on these findings, we propose **ReDeEP**, a novel method that detects hallucinations by decoupling LLM’s utilization of external context and parametric knowledge. Our experiments show that ReDeEP significantly improves RAG hallucination detection accuracy. Additionally, we introduce AARF, which mitigates hallucinations by modulating the contributions of Knowledge FFNs and Copying Heads.
****
Zhenyu He,Jun Zhang,Shengjie Luo,Jingjing Xu,Zhi Zhang,Di He
Let the Code LLM Edit Itself When You Edit the Code
In this work, we investigate a typical scenario in code generation where a developer edits existing code in real time and requests a code assistant, e.g., a large language model, to re-predict the next token or next line on the fly. Naively, the LLM needs to re-encode the entire KV cache to provide an accurate prediction. However, this process is computationally expensive, especially when the sequence length is long. Simply encoding the edited subsequence and integrating it to the original KV cache meets the temporal confusion problem, leading to significantly worse performance. We address this efficiency and accuracy trade-off by introducing $\underline{\textbf{P}\text{ositional}\ \textbf{I}\text{ntegrity}\ \textbf{E}\text{ncoding}}$ (PIE). Building upon the rotary positional encoding,

PIE first removes the rotary matrices in the Key cache that introduce temporal c
onfusion and then reapplies the correct rotary matrices. This process ensures th
at positional relationships between tokens are correct and requires only a singl
e round of matrix multiplication. We validate the effectiveness of PIE through e
xtensive experiments on the RepoBench-C-8k dataset, utilizing DeepSeek-Coder mod
els with 1.3B, 6.7B, and 33B parameters. Our evaluation includes three real-worl
d coding tasks: code insertion, code deletion, and multi-place code editing. Res
ults demonstrate that PIE reduces computational overhead by over 85% compared t
o the standard full recomputation approach across all model sizes and tasks whil
e well approximating the model performance.
****

Yiqin Yang,Quanwei Wang,Chenghao Li,Hao Hu,Chengjie Wu,Yuhua Jiang,Dianyu Zhong,
Ziyou Zhang,Qianchuan Zhao,Chongjie Zhang,Bo XU
Fewer May Be Better: Enhancing Offline Reinforcement Learning with Reduced Datas
et
Research in offline reinforcement learning (RL) marks a paradigm shift in RL. Ho
wever, a critical yet under-investigated aspect of offline RL is determining the
 subset of the offline dataset, which is used to improve algorithm performance w
hile accelerating algorithm training. Moreover, the size of reduced datasets can
 uncover the requisite offline data volume essential for addressing analogous ch
allenges. Based on the above considerations, we propose identifying Reduced Data
sets for Offline RL (ReDOR) by formulating it as a gradient approximation optimi
zation problem. We prove that the common actor-critic framework in reinforcemen
t learning can be transformed into a submodular objective. This insight enables
us to construct a subset by adopting the orthogonal matching pursuit (OMP). Spec
ifically, we have made several critical modifications to OMP to enable successfu
l adaptation with Offline RL algorithms. The experimental results indicate that
the data subsets constructed by the ReDOR can significantly improve algorithm pe
rformance with low computational complexity.
****

Changnan Xiao,Bing Liu
Generalizing Reasoning Problems to Longer Lengths
Length generalization (LG) is a challenging problem in learning to reason. It re
fers to the phenomenon that when trained on reasoning problems of smaller length
s/sizes, the model struggles with problems of larger sizes or lengths. Although
it has been proven that reasoning can be learned if the intermediate reasoning s
teps (also known as chain-of-thought (CoT)) are given in the training data, exis
ting studies only apply to within a given length (interpolation), while LG is ab
out extrapolation beyond the given length. This paper begins by presenting a the
orem that identifies the root cause of the LG problem. It then defines a class o
f reasoning problems for which achieving LG with Transformers can be theoretical
ly guaranteed, provided the CoT schemes are constructed to meet a proposed condi
tion called (n,r)-consistency.
****

Tian Ye,Zicheng Xu,Yuanzhi Li,Zeyuan Allen-Zhu
Physics of Language Models: Part 2.2, How to Learn From Mistakes on Grade-School
 Math Problems
Language models have demonstrated remarkable performance in solving reasoning ta
sks; however, even the strongest models still occasionally make reasoning mistak
es. Recently, there has been active research aimed at improving reasoning accura
cy, particularly by using pretrained language models to "self-correct'' their mi
stakes via multi-round prompting. In this paper, we follow this line of work but
 focus on understanding the usefulness of incorporating ``error-correction'' dat
a directly into the pretraining stage. This data consists of erroneous solution
steps immediately followed by their corrections. Using a synthetic math dataset,
 we show promising results: this type of pretrain data can help language models
achieve higher reasoning accuracy directly (i.e., through simple auto-regression
, without multi-round prompting) compared to pretraining on the same amount of e
rror-free data. We also delve into many details, such as (1) how this approach d
iffers from beam search, (2) how such data can be prepared, (3) whether masking

is needed on the erroneous tokens, (4) the amount of error required, (5) whether such data can be deferred to the fine-tuning stage, and many others.
****

Yingqi Liu,Qinglun Li,Jie Tan,Yifan Shi,Li Shen,Xiaochun Cao
Understanding the Stability-based Generalization of Personalized Federated Learning
Despite great achievements in algorithm design for Personalized Federated Learning (PFL), research on the theoretical analysis of generalization is still in its early stages. Some theoretical results have investigated the generalization performance of personalized models under the problem setting and hypothesis in convex conditions, which can not reflect the real iteration performance during non-convex training. To further understand the real performance from a generalization perspective, we propose the first algorithm-dependent generalization analysis with uniform stability for the typical PFL method, Partial Model Personalization, on smooth and non-convex objectives. Specifically, we decompose the generalization errors into aggregation errors and fine-tuning errors, then creatively establish a generalization analysis framework corresponding to the gradient estimation process of the personalized training. This framework builds up the bridge among PFL, FL and Pure Local Training for personalized aims in heterogeneous scenarios, which clearly demonstrates the effectiveness of PFL from the generalization perspective. Moreover, we demonstrate the impact of trivial factors like learning steps, stepsizes and communication topologies and obtain the excess risk analysis with optimization errors for PFL. Promising experiments on CIFAR datasets also corroborate our theoretical insights. Our code can be seen in https://github.com/YingqiLiu1999/Understanding-the-Stability-based-Generalization-of-Personalized-Federated-Learning.
****

Rubo Wang,Fandi Wu,Xingyu Gao,Jiaxiang Wu,Peilin Zhao,Jianhua Yao
IgGM: A Generative Model for Functional Antibody and Nanobody Design
Immunoglobulins are crucial proteins produced by the immune system to identify and bind to foreign substances, playing an essential role in shielding organisms from infections and diseases. Designing specific antibodies opens new pathways for disease treatment. With the rise of deep learning, AI-driven drug design has become possible, leading to several methods for antibody design. However, many of these approaches require additional conditions that differ from real-world scenarios, making it challenging to incorporate them into existing antibody design processes. Here, we introduce IgGM, a generative model for the de novo design of immunoglobulins with functional specificity. IgGM simultaneously generates antibody sequences and structures for a given antigen, consisting of three core components: a pre-trained language model for extracting sequence features, a feature learning module for identifying pertinent features, and a prediction module that outputs designed antibody sequences and the predicted complete antibody-antigen complex structure. IgGM effectively predicts structures and designs novel antibodies and nanobodies. This makes it highly applicable in a wide range of practical situations related to antibody and nanobody design. Code is available at: https://github.com/TencentAI4S/IgGM.
****

Kenneth Enevoldsen,Isaac Chung,Imene Kerboua,Márton Kardos,Ashwin Mathur,David Stap,Jay Gala,Wissam Siblini,Dominik Krzemi■ski,Genta Indra Winata,Saba Sturua,Saiteja Utpala,Mathieu Ciancone,Marion Schaeffer,Diganta Misra,Shreeya Dhakal,Jonathan Rystrøm,Roman Solomatin,Ömer Veysel Ça■atan,Akash Kundu,Martin Bernstorff,Shitao Xiao,Akshita Sukhlecha,Bhavish Pahwa,Rafa■ Po■wiata,Kranthi Kiran GV,Shawon Ashraf,Daniel Auras,Björn Plüster,Jan Philipp Harries,Loïc Magne,Isabelle Mohr,Dawei Zhu,Hippolyte Gisserot-Boukhlef,Tom Aarsen,Jan Kostkan,Konrad Wojtasik,Taemin Lee,Marek Suppa,Crystina Zhang,Roberta Rocca,Mohammed Hamdy,Andrianos Michail,John Yang,Manuel Faysse,Aleksei Vatolin,Nandan Thakur,Manan Dey,Dipam Vasani,Pranjal A Chitale,Simone Tedeschi,Nguyen Tai,Artem Snegirev,Mariya Hendriksen,Michael Günther,Mengzhou Xia,Weijia Shi,Xing Han Lù,Jordan Clive,Gayatri K,Maksimova Anna,Silvan Wehrli,Maria Tikhonova,Henil Shalin Panchal,Aleksandr Abramov,Malte Ostendorff,Zheng Liu,Simon Clematide,Lester James Validad Miranda,Alena Fenog

enova,Guangyu Song,Ruqiya Bin Safi,Wen-Ding Li,Alessia Borghini,Federico Cassano
,Lasse Hansen,Sara Hooker,Chenghao Xiao,Vaibhav Adlakha,Orion Weller,Siva Reddy,
Niklas Muennighoff
MMTEB: Massive Multilingual Text Embedding Benchmark
Text embeddings are typically evaluated on a narrow set of tasks, limited in ter
ms of languages, domains, and task types. To circumvent this limitation and to p
rovide a more comprehensive evaluation, we introduce the Massive Multilingual Te
xt Embedding Benchmark (MMTEB) -- a large-scale community-driven initiative expa
nding MTEB to over 500 quality-controlled evaluation tasks across 1,000+ languag
es. MMTEB includes a wide range of challenging novel tasks such as instruction f
ollowing, long-document retrieval, and code retrieval, and represents the larges
t multilingual collection of evaluation tasks for embedding models to date. We u
se this collection to construct multiple highly multilingual benchmarks. We eval
uate a representative set of models on these benchmarks.
Our findings indicate that, while LLM-based models can achieve state-of-the-art
performance on a subset of languages, the best-performing publicly available mod
el across languages is the notably smaller, multilingual-e5-large-instruct.

Massive benchmarks often impose high computational demands, limiting accessibili
ty, particularly for low-resource communities. To address this, we downsample ta
sks based on inter-task correlation (i.e., selecting only a diverse set of tasks
) while preserving relative rankings.
We further optimize tasks such as retrieval by sampling hard negatives, creating
 smaller but effective splits. These optimizations allow us to introduce benchma
rks at a significantly lower computational cost. For instance, we introduce a ne
w zero-shot English benchmark that maintains a similar ordering at a fraction of
 the cost.
****
Cuong C. Nguyen,Thanh-Toan Do,Gustavo Carneiro
Probabilistic Learning to Defer: Handling Missing Expert Annotations and Control
ling Workload Distribution
Recent progress in machine learning research is gradually shifting its focus tow
ards *human-AI cooperation* due to the advantages of exploiting the reliability
of human experts and the efficiency of AI models. One of the promising approache
s in human-AI cooperation is *learning to defer* (L2D), where the system analyse
s the input data and decides to make its own decision or defer to human experts.
 Although L2D has demonstrated state-of-the-art performance, in its standard set
ting, L2D entails a severe limitation: all human experts must annotate the whole
 training dataset of interest, resulting in a time-consuming and expensive annot
ation process that can subsequently influence the size and diversity of the trai
ning set. Moreover, the current L2D does not have a principled way to control wo
rkload distribution among human experts and the AI classifier, which is critical
 to optimise resource allocation. We, therefore, propose a new probabilistic mo
delling approach inspired by the mixture-of-experts, where the Expectation - Max
imisation algorithm is leverage to address the issue of missing expert's annotat
ions. Furthermore, we introduce a constraint, which can be solved efficiently du
ring the E-step, to control the workload distribution among human experts and th
e AI classifier. Empirical evaluation on synthetic and real-world datasets shows
 that our proposed probabilistic approach performs competitively, or surpasses p
reviously proposed methods assessed on the same benchmarks.
****
Zihao Huang,Qiyang Min,Hongzhi Huang,Yutao Zeng,Defa Zhu,Ran Guo,zhou Xun
Ultra-Sparse Memory Network
It is widely acknowledged that the performance of Transformer models is logarith
mically related to their number of parameters and computational complexity. Whil
e approaches like Mixture of Experts (MoE) decouple parameter count from computa
tional complexity, they still face challenges in inference due to high memory ac
cess costs. This work introduces UltraMem, incorporating large-scale, ultra-spar
se memory layer to address these limitations. Our approach significantly reduces
 inference latency while maintaining model performance. We also investigate the

scaling laws of this new architecture, demonstrating that it not only exhibits f
avorable scaling properties but outperforms MoE. In experiments, the largest Ult
raMem we train has \textbf{20 million} memory slots. The results show that our m
ethod achieves state-of-the-art inference speed and model performance within a g
iven computational budget, paving the way for billions of slots or experts.
****
Raphaël Sarfati,Toni J.B. Liu,Nicolas Boulle,Christopher Earls
Lines of Thought in Large Language Models
Large Language Models achieve next-token prediction by transporting a vectorized
 piece of text (prompt) across an accompanying embedding space under the action
of successive transformer layers. The resulting high-dimensional trajectories re
alize different contextualization, or 'thinking', steps, and fully determine the
 output probability distribution. We aim to characterize the statistical propert
ies of ensembles of these 'lines of thought.' We observe that independent trajec
tories cluster along a low-dimensional, non-Euclidean manifold, and that their p
ath can be well approximated by a stochastic equation with few parameters extrac
ted from data. We find it remarkable that the vast complexity of such large mode
ls can be reduced to a much simpler form, and we reflect on implications.
****
Yeongjae Cho,Keonwoo Kim,Taebaek Hwang,Sungzoon Cho
Do You Keep an Eye on What I Ask? Mitigating Multimodal Hallucination via Attent
ion-Guided Ensemble Decoding
Recent advancements in Large Vision-Language Models (LVLMs) have significantly e
xpanded their utility in tasks like image captioning and visual question answeri
ng. However, they still struggle with object hallucination, where models generat
e descriptions that inaccurately reflect the visual content by including nonexis
tent objects or misrepresenting existing ones. While previous methods, such as d
ata augmentation and training-free approaches, strive to tackle this issue, they
 still encounter scalability challenges and often depend on additional external
modules. In this work, we propose Ensemble Decoding (ED), a novel strategy that
splits the input image into sub-images and combines logit distributions by assig
ning weights through the attention map. Furthermore, we introduce ED adaptive pl
ausibility constraint to calibrate logit distribution and FastED, a variant desi
gned for speed-critical applications. Extensive experiments across hallucination
 benchmarks demonstrate that our proposed method achieves state-of-the-art perfo
rmance, validating the effectiveness of our approach.
****
Julia Gaudio,Nirmit Joshi
Exact Community Recovery under Side Information: Optimality of Spectral Algorith
ms
We study the problem of exact community recovery in general, two-community block
 models, in the presence of node-attributed *side information*. We allow for a v
ery general side information channel for node attributes, and for pairwise (edge
) observations, consider both Bernoulli and Gaussian matrix models, capturing th
e Stochastic Block Model, Submatrix Localization, and $\mathbb{Z}_2$-Synchroniza
tion as special cases. A recent work of Dreveton et al. 2024 characterized the i
nformation-theoretic limit of a very general exact recovery problem with side in
formation. In this paper, we show algorithmic achievability in the above importa
nt cases by designing a simple but optimal spectral algorithm that incorporates
side information (when present) along with the eigenvectors of the pairwise obse
rvation matrix. Using the powerful tool of entrywise eigenvector analysis [Abbe
et al. 2020], we show that our spectral algorithm can mimic the so called *genie
-aided estimators*, where the ith genie-aided estimator optimally
computes the estimate of the ith label, when all remaining labels
are revealed by a genie. This perspective provides a unified understanding of th
e optimality of spectral algorithms for various exact recovery problems in a rec
ent line of work.
****
Michael Wornow,Suhana Bedi,Miguel Angel Fuentes Hernandez,Ethan Steinberg,Jason
Alan Fries,Christopher Re,Sanmi Koyejo,Nigam Shah

Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data
Foundation Models (FMs) trained on Electronic Health Records (EHRs) have achieved state-of-the-art results on numerous clinical prediction tasks. However, prior EHR FMs typically have context windows of $<$1k tokens, which prevents them from modeling full patient EHRs which can exceed 10k's of events. For making clinical predictions, both model performance and robustness to the unique properties of EHR data are crucial. Recent advancements in subquadratic long-context architectures (e.g. Mamba) offer a promising solution. However, their application to EHR data has not been well-studied. We address this gap by presenting the first systematic evaluation of the effect of context length on modeling EHR data. We find that longer context models improve predictive performance -- our Mamba-based model surpasses the prior state-of-the-art on 9/14 tasks on the EHRSHOT prediction benchmark. Additionally, we measure robustness to three unique, previously underexplored properties of EHR data: (1) the prevalence of ``copy-forwarded" diagnoses which create artificial token repetition in EHR sequences; (2) the irregular time intervals between EHR events which can lead to a wide range of timespans within a context window; and (3) the natural increase in disease complexity over time which makes later tokens in the EHR harder to predict than earlier ones. Stratifying our EHRSHOT results, we find that higher levels of each property correlate negatively with model performance (e.g., a 14% higher Brier loss between the least and most irregular patients), but that longer context models are more robust to more extreme levels of these properties. Our work highlights the potential for using long-context architectures to model EHR data, and offers a case study on how to identify and quantify new challenges in modeling sequential data motivated by domains outside of natural language. We release all of our model checkpoints and code.
****
Rosie Zhao,Depen Morwani,David Brandfonbrener,Nikhil Vyas,Sham M. Kakade
Deconstructing What Makes a Good Optimizer for Autoregressive Language Models
Training language models becomes increasingly expensive with scale, prompting numerous attempts to improve optimization efficiency. Despite these efforts, the Adam optimizer remains the most widely used, due to a prevailing view that it is the most effective approach. We aim to compare several optimization algorithms, including SGD, Adafactor, Adam, Lion, and Sophia in the context of autoregressive language modeling across a range of model sizes, hyperparameters, and architecture variants. Our findings indicate that, except for SGD, these algorithms all perform comparably both in their optimal performance and also in terms of how they fare across a wide range of hyperparameter choices. Our results suggest to practitioners that the choice of optimizer can be guided by practical considerations like memory constraints and ease of implementation, as no single algorithm emerged as a clear winner in terms of performance or stability to hyperparameter misspecification. Given our findings, we further dissect these approaches, examining two simplified versions of Adam: a) signed momentum (Signum) which we see recovers both the performance and hyperparameter stability of Adam and b) Adalayer, a layerwise variant of Adam which we introduce to study the impact on Adam's preconditioning for different layers of the network. Examining Adalayer leads us to the conclusion that, perhaps surprisingly, adaptivity on *both* the last layer and LayerNorm parameters in particular are necessary for retaining performance and stability to learning rate.
****
Yuechen Xie,Jie Song,Mengqi Xue,Haofei Zhang,Xingen Wang,Bingde Hu,Genlang Chen,Mingli Song
Dataset Ownership Verification in Contrastive Pre-trained Models
High-quality open-source datasets, which necessitate substantial efforts for curation, has become the primary catalyst for the swift progress of deep learning. Concurrently, protecting these datasets is paramount for the well-being of the data owner. Dataset ownership verification emerges as a crucial method in this domain, but existing approaches are often limited to supervised models and cannot be directly extended to increasingly popular unsupervised pre-trained models. In

this work, we propose the first dataset ownership verification method tailored specifically for self-supervised pre-trained models by contrastive learning. Its primary objective is to ascertain whether a suspicious black-box backbone has been pre-trained on a specific unlabeled dataset, aiding dataset owners in upholding their rights. The proposed approach is motivated by our empirical insights that when models are trained with the target dataset, the unary and binary instance relationships within the embedding space exhibit significant variations compared to models trained without the target dataset. We validate the efficacy of this approach across multiple contrastive pre-trained models including SimCLR, BYOL, SimSiam, MOCO v3, and DINO. The results demonstrate that our method rejects the null hypothesis with a p-value markedly below 0.05, surpassing all previous methodologies. Our code is available at https://github.com/xieyc99/DOV4CL.
****

Swarnadeep Saha,Archiki Prasad,Justin Chen,Peter Hase,Elias Stengel-Eskin,Mohit Bansal
System 1.x: Learning to Balance Fast and Slow Planning with Language Models
Language models can be used to solve long-horizon planning problems in two distinct modes. In a fast 'System-1' mode, models directly generate plans without any explicit search or backtracking, and in a slow 'System-2' mode, they plan step-by-step by explicitly searching over possible actions. System-2 planning, while typically more effective, is also computationally more expensive and often infeasible for long plans or large action spaces. Moreover, isolated System-1 or System-2 planning ignores the user's end goals and constraints (e.g., token budget), failing to provide ways for the user to control the model's behavior. To this end, we propose the System-1.x Planner, a framework for controllable planning with language models that is capable of generating hybrid plans and balancing between the two planning modes based on the difficulty of the problem at hand. System-1.x consists of (i) a controller, (ii) a System-1 Planner, and (iii) a System-2 Planner. Based on a user-specified hybridization factor x governing the degree to which the system uses System-1 vs. System-2, the controller decomposes a planning problem into subgoals, and classifies them as easy or hard to be solved by either System-1 or System-2, respectively. We fine-tune all three components on top of a single base LLM, requiring only search traces as supervision. Experiments with two diverse planning tasks -- Maze Navigation and Blocksworld -- show that our System-1.x Planner outperforms a System-1 Planner, a System-2 Planner trained to approximate A* search, and also a symbolic planner (A* search), given a state exploration budget. We also demonstrate the following key properties of our planner: (1) controllability: by adjusting the hybridization factor x (e.g., System-1.75 vs. System-1.5) we can perform more (or less) search, improving performance, (2) flexibility: by building a neuro-symbolic variant composed of a neural System-1 planner and a symbolic System-2 planner, we can take advantage of existing symbolic methods, and (3) generalizability: by learning from different search algorithms (BFS, DFS, A*), we show that our method is robust to the choice of search algorithm used for training.
****

Zepeng Frazier Huo,Jason Alan Fries,Alejandro Lozano,Jeya Maria Jose Valanarasu,Ethan Steinberg,Louis Blankemeier,Akshay S Chaudhari,Curtis Langlotz,Nigam Shah
Time-to-Event Pretraining for 3D Medical Imaging
With the rise of medical foundation models and the growing availability of imaging data, scalable pretraining techniques offer a promising way to identify imaging biomarkers predictive of future disease risk. While current self-supervised methods for 3D medical imaging models capture local structural features like organ morphology, they fail to link pixel biomarkers with long-term health outcomes due to a missing context problem. Current approaches lack the temporal context necessary to identify biomarkers correlated with disease progression, as they rely on supervision derived only from images and concurrent text descriptions. To address this, we introduce time-to-event pretraining, a pretraining framework for 3D medical imaging models that leverages large-scale temporal supervision from paired, longitudinal electronic health records (EHRs). Using a dataset of 18,945 CT scans (4.2 million 2D images) and time-to-event distributions across thousan

ds of EHR-derived tasks, our method improves outcome prediction, achieving an av
erage AUROC increase of 23.7% and a 29.4% gain in Harrell’s C-index across 8 ben
chmark tasks. Importantly, these gains are achieved without sacrificing diagnost
ic classification performance. This study lays the foundation for integrating lo
ngitudinal EHR and 3D imaging data to advance clinical risk prediction.
****

S David Mis,Matti Lassas,Maarten V. de Hoop
Semialgebraic Neural Networks: From roots to representations
Many numerical algorithms in scientific computing—particularly in areas like num
erical linear algebra, PDE simulation, and inverse problems—produce outputs that
 can be represented by semialgebraic functions; that is, the graph of the comput
ed function can be described by finitely many polynomial equalities and inequali
ties. In this work, we introduce Semialgebraic Neural Networks (SANNs), a neural
 network architecture capable of representing any bounded semialgebraic function
, and computing such functions up to the accuracy of a numerical ODE solver chos
en by the programmer. Conceptually, we encode the graph of the learned function
as the kernel of a piecewise polynomial selected from a class of functions whose
 roots can be evaluated using a particular homotopy continuation method. We show
 by construction that the SANN architecture is able to execute this continuation
 method, thus evaluating the learned semialgebraic function. Furthermore, the ar
chitecture can exactly represent even discontinuous semialgebraic functions by e
xecuting a continuation method on each connected component of the target functio
n. Lastly, we provide example applications of these networks and show they can b
e trained with traditional deep-learning techniques.
****

Moussa Koulako Bala Doumbouya,Ananjan Nandi,Gabriel Poesia,Davide Ghilardi,Anna
Goldie,Federico Bianchi,Dan Jurafsky,Christopher D Manning
h4rm3l: A Language for Composable Jailbreak Attack Synthesis
Despite their demonstrated valuable capabilities, state-of-the-art (SOTA) widely
 deployed large language models (LLMs) still have the potential to cause harm to
 society due to the ineffectiveness of their safety filters, which can be bypass
ed by prompt transformations called jailbreak attacks. Current approaches to LLM
 safety assessment, which employ datasets of templated prompts and benchmarking
pipelines, fail to cover sufficiently large and diverse sets of jailbreak attack
s, leading to the widespread deployment of unsafe LLMs. Recent research showed t
hat novel jailbreak attacks could be derived by composition; however, a formal c
omposable representation for jailbreak attacks, which, among other benefits, cou
ld enable the exploration of a large compositional space of jailbreak attacks th
rough program synthesis methods, has not been previously proposed. We introduce
h4rm3l, a novel approach that addresses this gap with a human-readable domain-sp
ecific language (DSL). Our framework comprises: (1) The h4rm3l DSL, which formal
ly expresses jailbreak attacks as compositions of parameterized string transform
ation primitives. (2) A synthesizer with bandit algorithms that efficiently gene
rates jailbreak attacks optimized for a target black box LLM. (3) The h4rm3l red
-teaming software toolkit that employs the previous two components and an automa
ted harmful LLM behavior classifier that is strongly aligned with human judgment
. We demonstrate h4rm3l's efficacy by synthesizing a dataset of 2656 successful
novel jailbreak attacks targeting 6 SOTA open-source and proprietary LLMs (GPT-3
.5, GPT-4o, Claude-3-Sonnet, Claude-3-Haiku, Llama-3-8B, and Llama-3-70B), and b
y benchmarking those models against a subset of these synthesized attacks. Our r
esults show that h4rm3l's synthesized attacks are diverse and more successful th
an existing jailbreak attacks in literature, with success rates exceeding 90% on
 SOTA LLMs. Warning: This paper and related research artifacts contain offensive
 and potentially disturbing prompts and model-generated content.
****

Yufan Hu,Yujiang,Wuyang Li,Yixuan Yuan
Synthesizing Realistic fMRI: A Physiological Dynamics-Driven Hierarchical Diffus
ion Model for Efficient fMRI Acquisition
Functional magnetic resonance imaging (fMRI) is essential for mapping brain acti
vity but faces challenges like lengthy acquisition time and sensitivity to patie

nt movement, limiting its clinical and machine learning applications. While gene
rative models such as diffusion models can synthesize fMRI signals to alleviate
these issues, they often underperform due to neglecting the brain's complex stru
ctural and dynamic properties.
To address these limitations, we propose the Physiological Dynamics-Driven Hiera
rchical Diffusion Model, a novel framework integrating two key brain physiologic
al properties into the diffusion process: brain hierarchical regional interactio
ns and multifractal dynamics.
To model complex interactions among brain regions, we construct hypergraphs base
d on the prior knowledge of brain functional parcellation reflected by resting-s
tate functional connectivity (rsFC). This enables the aggregation of fMRI signal
s across multiple scales and generates hierarchical signals.
Additionally, by incorporating the prediction of two key dynamics properties of
fMRI—the multifractal spectrum and generalized Hurst exponent—our framework effe
ctively guides the diffusion process, ensuring the preservation of the scale-inv
ariant characteristics inherent in real fMRI data.
Our framework employs progressive diffusion generation, with signals representin
g broader brain region information conditioning those that capture localized det
ails, and unifies multiple inputs during denoising for balanced integration.
Experiments demonstrate that our model generates physiologically realistic fMRI
signals, potentially reducing acquisition time and enhancing data quality, benef
iting clinical diagnostics and machine learning in neuroscience.
****

Tian-Shuo Liu,Xu-Hui Liu,Ruifeng Chen,Lixuan Jin,Pengyuan Wang,Zhilong Zhang,Yan
g Yu
Semantic Temporal Abstraction via Vision-Language Model Guidance for Efficient R
einforcement Learning
Extracting temporally extended skills can significantly improve the efficiency o
f reinforcement learning (RL) by breaking down complex decision-making problems
with sparse rewards into simpler subtasks and enabling more effective credit ass
ignment. However, existing abstraction methods either discover skills in an unsu
pervised manner, which often lacks semantic information and leads to erroneous o
r scattered skill extraction results, or require substantial human intervention.
 In this work, we propose to leverage the extensive knowledge in pretrained Visi
on-Language Models (VLMs) to progressively guide the latent space after vector q
uantization to be more semantically meaningful through relabeling each skill. Th
is approach, termed **V**ision-l**an**guage model guided **T**emporal **A**bstra
ction (**VanTA**), facilitates the discovery of more interpretable and task-rele
vant temporal segmentations from offline data without the need for extensive man
ual intervention or heuristics. By leveraging the rich information in VLMs, our
method can significantly outperform existing offline RL approaches that depend o
nly on limited training data. From a theory perspective, we demonstrate that str
onger internal sequential correlations within each sub-task, induced by VanTA, e
ffectively reduces suboptimality in policy learning. We validate the effectivene
ss of our approach through extensive experiments on diverse environments, includ
ing Franka Kitchen, Minigrid, and Crafter. These experiments show that our metho
d surpasses existing approaches in long-horizon offline reinforcement learning s
cenarios with both proprioceptive and visual observations.
****

Daiyao Yi,Hao Dong,Michael James Higley,Anne Churchland,Shreya Saxena
Shared-AE: Automatic Identification of Shared Subspaces in High-dimensional Neur
al and Behavioral Activity
Understanding the relationship between behavior and neural activity is crucial f
or understanding brain function. An effective method is to learn embeddings for
interconnected modalities. For simple behavioral tasks, neural features can be l
earned based on labels. However, complex behaviors, such as social interactions,
 require the joint extraction of behavioral and neural characteristics. In this
paper, we present an autoencoder (AE) framework, called Shared-AE, which include
s a novel regularization term that automatically identifies features shared betw
een neural activity and behavior, while simultaneously capturing the unique priv

ate features specific to each modality. We apply Shared-AE to large-scale neural activity recorded across the entire dorsal cortex of the mouse, during two very different behaviors: (i) head-fixed mice performing a self-initiated decision-making task, and (ii) freely-moving social behavior amongst two mice. Our model successfully captures both `shared features', shared across neural and behavioral activity, and `private features', unique to each modality, significantly enhancing our understanding of the alignment between neural activity and complex behaviors. The original code for the entire Shared-AE framework on Pytorch has been made publicly available at: \url{https://github.com/saxenalab-neuro/Shared-AE}.
****

Viraat Aryabumi,Yixuan Su,Raymond Ma,Adrien Morisot,Ivan Zhang,Acyr Locatelli,Marzieh Fadaee,Ahmet Üstün,Sara Hooker
To Code or Not To Code? Exploring Impact of Code in Pre-training
Including code in the pre-training data mixture, even for models not specifically designed for code, has become a common practice in LLMs pre-training. While there has been anecdotal consensus among practitioners that code data plays a vital role in general LLMs' performance, there is only limited work analyzing the precise impact of code on non-code tasks. In this work, we systematically investigate the impact of code data on general performance. We ask “what is the impact of code data used in pre-training on a large variety of downstream tasks beyond code generation”. We conduct extensive ablations and evaluate across a broad range of natural language reasoning tasks, world knowledge tasks, code benchmarks, and LLM-as-a-judge win-rates for models with sizes ranging from 470M to 2.8B parameters. Across settings, we find a consistent results that code is a critical building block for generalization far beyond coding tasks and improvements to code quality have an outsized impact across all tasks. In particular, compared to text-only pre-training, the addition of code results in up to relative increase of 8.2% in natural language (NL) reasoning, 4.2% in world knowledge, 6.6% improvement in generative win-rates, and a 12x boost in code performance respectively. Our work suggests investments in code quality and preserving code during pre-training have positive impacts.
****

Binghui Li,Zhixuan Pan,Kaifeng Lyu,Jian Li
Feature Averaging: An Implicit Bias of Gradient Descent Leading to Non-Robustness in Neural Networks
In this work, we investigate a particular implicit bias in gradient descent training, which we term “Feature Averaging,” and argue that it is one of the principal factors contributing to the non-robustness of deep neural networks. We show that, even when multiple discriminative features are present in the input data, neural networks trained by gradient descent tend to rely on an average (or a certain combination) of these features for classification, rather than distinguishing and leveraging each feature individually. Specifically, we provide a detailed theoretical analysis of the training dynamics of two-layer ReLU networks on a binary classification task, where the data distribution consists of multiple clusters with mutually orthogonal centers. We rigorously prove that gradient descent biases the network towards feature averaging, where the weights of each hidden neuron represent an average of the cluster centers (each corresponding to a distinct feature), thereby making the network vulnerable to input perturbations aligned with the negative direction of the averaged features. On the positive side, we demonstrate that this vulnerability can be mitigated through more granular supervision. In particular, we prove that a two-layer ReLU network can achieve optimal robustness when trained to classify individual features rather than merely the original binary classes. Finally, we validate our theoretical findings with experiments on synthetic datasets, MNIST, and CIFAR-10, and confirm the prevalence of feature averaging and its impact on adversarial robustness. We hope these theoretical and empirical insights deepen the understanding of how gradient descent shapes feature learning and adversarial robustness, and how more detailed supervision can enhance robustness.
****

Yongxin Guo,Xiaoying Tang,Tao Lin

Enhancing Clustered Federated Learning: Integration of Strategies and Improved Methodologies

Federated Learning (FL) is an evolving distributed machine learning approach that safeguards client privacy by keeping data on edge devices. However, the variation in data among clients poses challenges in training models that excel across all local distributions. Recent studies suggest clustering as a solution to address client heterogeneity in FL by grouping clients with distribution shifts into distinct clusters. Nonetheless, the diverse learning frameworks used in current clustered FL methods create difficulties in integrating these methods, leveraging their advantages, and making further enhancements.

To this end, this paper conducts a thorough examination of existing clustered FL methods and introduces a four-tier framework, named HCFL, to encompass and extend the existing approaches. Utilizing the HCFL, we identify persistent challenges associated with current clustering methods in each tier and propose an enhanced clustering method called HCFL^{+} to overcome these challenges. Through extensive numerical evaluations, we demonstrate the effectiveness of our clustering framework and the enhanced components. Our code is available at \url{https://github.com/LINs-lab/HCFL}.

****

Chenlu Ding,Jiancan Wu,Yancheng Yuan,Jinda Lu,Kai Zhang,Alex Su,Xiang Wang,Xiangnan He

Unified Parameter-Efficient Unlearning for LLMs

The advent of Large Language Models (LLMs) has revolutionized natural language processing, enabling advanced understanding and reasoning capabilities across a variety of tasks. Fine-tuning these models for specific domains, particularly through Parameter-Efficient Fine-Tuning (PEFT) strategies like LoRA, has become a prevalent practice due to its efficiency. However, this raises significant privacy and security concerns, as models may inadvertently retain and disseminate sensitive or undesirable information. To address these issues, we introduce a novel instance-wise unlearning framework, LLMEraser, which systematically categorizes unlearning tasks and applies precise parameter adjustments using influence functions. Unlike traditional unlearning techniques that are often limited in scope and require extensive retraining, LLMEraser is designed to handle a broad spectrum of unlearning tasks without compromising model performance. Extensive experiments on benchmark datasets demonstrate that LLMEraser excels in efficiently managing various unlearning scenarios while maintaining the overall integrity and efficacy of the models.

****

Adam Kania,Marko Mihajlovic,Sergey Prokudin,Jacek Tabor,Przemys■aw Spurek

FreSh: Frequency Shifting for Accelerated Neural Representation Learning

Implicit Neural Representations (INRs) have recently gained attention as a powerful approach for continuously representing signals such as images, videos, and 3D shapes using multilayer perceptrons (MLPs). However, MLPs are known to exhibit a low-frequency bias, limiting their ability to capture high-frequency details accurately. This limitation is typically addressed by incorporating high-frequency input embeddings or specialized activation layers. In this work, we demonstrate that these embeddings and activations are often configured with hyperparameters that perform well on average but are suboptimal for specific input signals under consideration, necessitating a costly grid search to identify optimal settings. Our key observation is that the initial frequency spectrum of an untrained model's output correlates strongly with the model's eventual performance on a given target signal. Leveraging this insight, we propose frequency shifting (or FreSh), a method that selects embedding hyperparameters to align the frequency spectrum of the model’s initial output with that of the target signal. We show that this simple initialization technique improves performance across various neural representation methods and tasks, achieving results comparable to extensive hyperparameter sweeps but with only marginal computational overhead compared to training a single model with default hyperparameters.

****

Qiyuan He,Kai Xu,Angela Yao

EvA: Erasing Spurious Correlations with Activations
Spurious correlations often arise when models associate features strongly correl
ated with, but not causally related to, the label e.g. an image classifier assoc
iates bodies of water with ducks. To mitigate spurious correlations, existing me
thods focus on learning unbiased representation or incorporating additional info
rmation about the correlations during training. This work removes spurious corre
lations by ``**E**rasing **wi**th **A**ctivations'' (EvA). EvA learns class-spe
cific spurious indicator on each channel for the fully connected layer of pretra
ined networks. By erasing spurious connections during re-weighting, EvA achieves
 state-of-the-art performance across diverse datasets (6.2\% relative gain on BA
R and achieves 4.1\% on Waterbirds). For biased datasets without any information
 about the spurious correlations, EvA can outperform previous methods (4.8\% rel
ative gain on Waterbirds) with 6 orders of magnitude less compute, highlighting
its data and computational efficiency.
****
Enshu Liu,Xuefei Ning,Yu Wang,Zinan Lin
Distilled Decoding 1: One-step Sampling of Image Auto-regressive Models with Flo
w Matching
Autoregressive (AR) models have recently achieved state-of-the-art performance i
n text and image generation. However, their primary limitation is slow generatio
n speed due to the token-by-token process. We ask an ambitious question: can a p
re-trained AR model be adapted to generate outputs in just one or two steps? If
successful, this would significantly advance the development and deployment of A
R models. We notice that existing works that attempt to speed up AR generation b
y generating multiple tokens at once fundamentally cannot capture the output dis
tribution due to the conditional dependencies between tokens, limiting their eff
ectiveness for few-step generation. To overcome this, we propose Distilled Decod
ing (DD), which leverages flow matching to create a deterministic mapping from G
aussian distribution to the output distribution of the pre-trained AR model. We
then train a network to distill this mapping, enabling few-step generation. The
entire training process of DD does not need the training data of the original AR
 model (as opposed to some other methods), thus making DD more practical. We eva
luate DD on state-of-the-art image AR models and present promising results. For
VAR, which requires 10-step generation (680 tokens), DD enables one-step generat
ion (6.3$\times$ speed-up), with an acceptable increase in FID from 4.19 to 9.96
. Similarly, for LlamaGen, DD reduces generation from 256 steps to 1, achieving
an 217.8$\times$ speed-up with a comparable FID increase from 4.11 to 11.35. In
both cases, baseline methods completely fail with FID scores $>$100. As the firs
t work to demonstrate the possibility of one-step generation for image AR models
, DD challenges the prevailing notion that AR models are inherently slow, and op
ens up new opportunities for efficient AR generation. The code and the pre-train
ed models will be released at https://github.com/imagination-research/distilled-
decoding. The project website is at https://imagination-research.github.io/disti
lled-decoding.
****
Tianjun Wei,Wei Wen,Ruizhi Qiao,Xing Sun,Jianghong Ma
RocketEval: Efficient automated LLM evaluation via grading checklist
Evaluating large language models (LLMs) in diverse and challenging scenarios is
essential to align them with human preferences. To mitigate the prohibitive cost
s associated with human evaluations, utilizing a powerful LLM as a judge has eme
rged as a favored approach. Nevertheless, this methodology encounters several ch
allenges, including substantial expenses, concerns regarding privacy and securit
y, and reproducibility. In this paper, we propose a straightforward, replicable,
 and accurate automated evaluation method by leveraging a lightweight LLM as the
 judge, named RocketEval. Initially, we identify that the performance disparity
between lightweight and powerful LLMs in evaluation tasks primarily stems from t
heir ability to conduct comprehensive analyses, which is not easily enhanced thr
ough techniques such as chain-of-thought reasoning. By reframing the evaluation
task as a multi-faceted Q\&A using an instance-specific checklist, we demonstrat
e that the limited judgment accuracy of lightweight LLMs is largely attributes t

o high uncertainty and positional bias. To address these challenges, we introduc
e an automated evaluation process grounded in checklist grading, which is design
ed to accommodate a variety of scenarios and questions. This process encompasses
 the creation of checklists, the grading of these checklists by lightweight LLMs
, and the reweighting of checklist items to align with the supervised annotation
s. Our experiments carried out on the automated evaluation benchmarks, MT-Bench
and WildBench datasets, reveal that RocketEval, when using $\textit{Gemma-2-2B}$
 as the judge, achieves a high correlation (0.965) with human preferences, which
 is comparable to $\textit{GPT-4o}$. Moreover, RocketEval provides a cost reduct
ion exceeding 50-fold for large-scale evaluation and comparison scenarios. Our c
ode is available at https://github.com/Joinn99/RocketEval-ICLR.
****
Yuqi Feng,Yuwei Ou,Jiahao Fan,Yanan Sun
Zero-cost Proxy for Adversarial Robustness Evaluation
Deep neural networks (DNNs) easily cause security issues due to the lack of adve
rsarial robustness. An emerging research topic for this problem is to design adv
ersarially robust architectures via neural architecture search (NAS), i.e., robu
st NAS. However, robust NAS needs to train numerous DNNs for robustness estimati
on, making the search process prohibitively expensive. In this paper, we propose
 a zero-cost proxy to evaluate the adversarial robustness without training. Spec
ifically, the proposed zero-cost proxy formulates the upper bound of adversarial
 loss, which can directly reflect the adversarial robustness. The formulation in
volves only the initialized weights of DNNs, thus the training process is no lon
ger needed. Moreover, we theoretically justify the validity of the proposed prox
y based on the theory of neural tangent kernel and input loss landscape. Experim
ental results show that the proposed zero-cost proxy can bring more than $20\tim
es$ speedup compared with the state-of-the-art robust NAS methods, while the sea
rched architecture has superior robustness and transferability under white-box a
nd black-box attacks. Furthermore, compared with the state-of-the-art zero-cost
proxies, the calculation of the proposed method has the strongest correlation wi
th adversarial robustness. Our source code is available at https://github.com/fy
qsama/Robust_ZCP.
****
Yingyu Lin,Yuxing Huang,Wenqin Liu,Haoran Deng,Ignavier Ng,Kun Zhang,Mingming Go
ng,Yian Ma,Biwei Huang
A Skewness-Based Criterion for Addressing Heteroscedastic Noise in Causal Discov
ery
Real-world data often violates the equal-variance assumption (homoscedasticity),
 making it essential to account for heteroscedastic noise in causal discovery. I
n this work, we explore heteroscedastic symmetric noise models (HSNMs), where th
e effect Y is modeled as $Y = f(X) + \sigma(X)N$, with X as the cause and $N
$ as independent noise following a symmetric distribution. We introduce a novel
criterion for identifying HSNMs based on the skewness of the score (i.e., the gr
adient of the log density) of the data distribution. This criterion establishes
a computationally tractable measurement that is zero in the causal direction but
 nonzero in the anticausal direction, enabling the causal direction discovery. W
e extend this skewness-based criterion to the multivariate setting and propose \
texttt{SkewScore}, an algorithm that handles heteroscedastic noise without requi
ring the extraction of exogenous noise. We also conduct a case study on the robu
stness of \texttt{SkewScore} in a bivariate model with a latent confounder, prov
iding theoretical insights into its performance. Empirical studies further valid
ate the effectiveness of the proposed method.
****
Buu Phan,Brandon Amos,Itai Gat,Marton Havasi,Matthew J. Muckley,Karen Ullrich
Exact Byte-Level Probabilities from Tokenized Language Models for FIM-Tasks and
Model Ensembles
Tokenization is associated with many poorly understood shortcomings in language
models (LMs), yet remains an important component for long sequence scaling purpo
ses. This work studies how tokenization impacts model performance by analyzing
 and comparing the stochastic behavior of tokenized models with their byte-level

, or token-free, counterparts. We discover that, even when the two models are st
atistically equivalent, their predictive distributions over the next byte can be
 substantially different, a phenomenon we term as ``tokenization bias''. To full
y characterize this phenomenon, we introduce the Byte-Token Representation Lemm
a, a framework that establishes a mapping between the learned token distribution
 and its equivalent byte-level distribution. From this result, we develop a nex
t-byte sampling algorithm that eliminates tokenization bias without requiring f
urther training or optimization. In other words, this enables zero-shot conversi
on of tokenized LMs into statistically equivalent token-free ones. We demonstrat
e its broad applicability with two use cases: fill-in-the-middle (FIM) tasks and
 model ensembles. In FIM tasks where input prompts may terminate mid-token, lead
ing to out-of-distribution tokenization, our method mitigates performance degrad
ation and achieves 18\% improvement in FIM coding benchmarks, while consistently
 outperforming the standard token healing fix. For model ensembles where each mo
del employs a distinct vocabulary, our approach enables seamless integration, re
sulting in improved performance up to 3.7\% over individual models across variou
s standard baselines in reasoning, knowledge, and coding. Code is available at:h
ttps: //github.com/facebookresearch/Exact-Byte-Level-Probabilities-from-Tokenize
d-LMs.
****
Feilong Tang,Zile Huang,Chengzhi Liu,Qiang Sun,Harry Yang,Ser-Nam Lim
Intervening Anchor Token: Decoding Strategy in Alleviating Hallucinations for ML
LMs
Multimodal large language models (MLLMs) offer a powerful mechanism for interpre
ting visual information. However, they often suffer from hallucinations, which i
mpede the real-world usage of these models. Existing methods attempt to alleviat
e this issue by designing special decoding strategies that penalize the summary
tokens. However, these methods lack analysis of the relationship between halluci
nation and summarization mechanism of LLMs. Interestingly, we find that penalizi
ng summary tokens is not necessary: merely intervening the query-key parameters
variance, without costing extra inference time, still alleviates hallucinations.
 Specifically, we explore the causes of hallucinations by analyzing localized se
lf-attention patterns called ``anchor" tokens and define the attention localizat
ion degree of the model as token propagation probabilities. Our analysis reveals
 that over-propagation of anchor tokens occurs when the distribution of eigenval
ues of the query and key matrices has a non-zero mean and a polarized variance,
leading to excessive dependence on anchor tokens while neglecting vision informa
tion and describes the image content with hallucination. Based on the observatio
n, we propose a versatile plug-and-play decoding strategy, Dynamic Token Propaga
tion Mechanism (TAME), to alleviate excessive propagation by dynamically interve
ning the eigenspectrum variance of the attention weight, thereby alleviating hal
lucinations without relying on complex decoding strategies. Extensive experiment
s reveal a correlation between the eigenspectrum and hallucinations across vario
us MLLMs, and show that TAME reduces the percentage of hallucinated objects.
****
Shi Yu,Chaoyue Tang,Bokai Xu,Junbo Cui,Junhao Ran,Yukun Yan,Zhenghao Liu,Shuo Wa
ng,Xu Han,Zhiyuan Liu,Maosong Sun
VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents
Retrieval-augmented generation (RAG) is an effective technique that enables larg
e language models (LLMs) to utilize external knowledge sources for generation.
However, current RAG systems are solely based on text, rendering it impossible t
o utilize vision information like layout and images that play crucial roles in r
eal-world multi-modality documents. In this paper, we introduce VisRAG, which t
ackles this issue by establishing a vision-language model (VLM)-based RAG pipeli
ne. In this pipeline, instead of first parsing the document to obtain text, the
 document is directly embedded using a VLM as an image and then retrieved to enh
ance the generation of a VLM. Compared to traditional text-based RAG, VisRAG max
imizes the retention and utilization of the data information in the original doc
uments, eliminating the information loss introduced during the parsing process.
We collect both open-source and synthetic data to train the retriever in VisRAG

and explore a variety of generation methods. Experiments demonstrate that VisRAG outperforms traditional RAG in both the retrieval and generation stages, achieving a 20–40% end-to-end performance gain over traditional text-based RAG pipeline. Further analysis reveals that VisRAG is efficient in utilizing training data and demonstrates strong generalization capability, positioning it as a promising solution for RAG on multi-modality documents. Our code and data are available at https://github.com/openbmb/visrag.

****

Kun Gao,Katsumi Inoue,Yongzhi Cao,Hanpin Wang,Yang Feng

Differentiable Rule Induction from Raw Sequence Inputs

Rule learning-based models are widely used in highly interpretable scenarios due to their transparent structures. Inductive logic programming (ILP), a form of machine learning, induces rules from facts while maintaining interpretability. Differentiable ILP models enhance this process by leveraging neural networks to improve robustness and scalability. However, most differentiable ILP methods rely on symbolic datasets, facing challenges when learning directly from raw data. Specifically, they struggle with explicit label leakage: The inability to map continuous inputs to symbolic variables without explicit supervision of input feature labels. In this work, we address this issue by integrating a self-supervised differentiable clustering model with a novel differentiable ILP model, enabling rule learning from raw data without explicit label leakage. The learned rules effectively describe raw data through its features. We demonstrate that our method intuitively and precisely learns generalized rules from time series and image data.

****

Meng YOU,Zhiyu Zhu,Hui LIU,Junhui Hou

NVS-Solver: Video Diffusion Model as Zero-Shot Novel View Synthesizer

By harnessing the potent generative capabilities of pre-trained large video diffusion models, we propose a new novel view synthesis paradigm that operates without the need for training. The proposed method adaptively modulates the diffusion sampling process with the given views to enable the creation of visually pleasing results from single or multiple views of static scenes or monocular videos of dynamic scenes. Specifically, built upon our theoretical modeling, we iteratively modulate the score function with the given scene priors represented with warped input views to control the video diffusion process. Moreover, by theoretically exploring the boundary of the estimation error, we achieve the modulation in an adaptive fashion according to the view pose and the number of diffusion steps. Extensive evaluations on both static and dynamic scenes substantiate the significant superiority of our method over state-of-the-art methods both quantitatively and qualitatively. The source code can be found on https://github.com/ZHU-Zhiyu/NVS_Solver.

****

Jiaqi Guan,Jiahan Li,Xiangxin Zhou,Xingang Peng,Sheng Wang,Yunan Luo,Jian Peng,Jianzhu Ma

Group Ligands Docking to Protein Pockets

Molecular docking is a key task in computational biology that has attracted increasing interest from the machine learning community. While existing methods have achieved success, they generally treat each protein-ligand pair in isolation. Inspired by the biochemical observation that ligands binding to the same target protein tend to adopt similar poses, we propose \textsc{GroupBind}, a novel molecular docking framework that simultaneously considers multiple ligands docking to a protein. This is achieved by introducing an interaction layer for the group of ligands and a triangle attention module for embedding protein-ligand and group-ligand pairs. By integrating our approach with diffusion based docking model, we set a new state-of-the-art performance on the PDBBind blind docking benchmark, demonstrating the effectiveness of our paradigm in enhancing molecular docking accuracy.

****

Michael Matthews,Michael Beukman,Chris Lu,Jakob Nicolaus Foerster

Kinetix: Investigating the Training of General Agents through Open-Ended Physics

-Based Control Tasks
While large models trained with self-supervised learning on offline datasets hav
e shown remarkable capabilities in text and image domains, achieving the same ge
neralisation for agents that act in sequential decision problems remains an open
 challenge.
In this work, we take a step towards this goal by procedurally generating tens o
f millions of 2D physics-based tasks and using these to train a general reinforc
ement learning (RL) agent for physical control.
To this end, we introduce Kinetix: an open-ended space of physics-based RL envir
onments that can represent tasks ranging from robotic locomotion and grasping to
 video games and classic RL environments, all within a unified framework.
Kinetix makes use of our novel hardware-accelerated physics engine Jax2D that al
lows us to cheaply simulate billions of environment steps during training.
Our trained agent exhibits strong physical reasoning capabilities in 2D space, b
eing able to zero-shot solve unseen human-designed environments. Furthermore, f
ine-tuning this general agent on tasks of interest shows significantly stronger
performance than training an RL agent *tabula rasa*. This includes solving some
 environments that standard RL training completely fails at.
We believe this demonstrates the feasibility of large scale, mixed-quality pre-t
raining for online RL and we hope that Kinetix will serve as a useful framework
to investigate this further.
****
Tehila Dahan,Kfir Yehuda Levy
Do Stochastic, Feel Noiseless: Stable Stochastic Optimization via a Double Momen
tum Mechanism
Optimization methods are crucial to the success of machine learning, with Stocha
stic Gradient Descent (SGD) serving as a foundational algorithm for training mod
els. However, SGD is often sensitive to the choice of the learning rate, which n
ecessitates extensive hyperparameter tuning. In this work, we introduce a new va
riant of SGD that brings enhanced stability in two key aspects. First, our metho
d allows the use of the same fixed learning rate to attain optimal convergence r
ates regardless of the noise magnitude, eliminating the need to adjust learning
rates between noiseless and noisy settings. Second, our approach achieves these
optimal rates over a wide range of learning rates, significantly reducing sensit
ivity compared to standard SGD, which requires precise learning rate selection.
Our key innovation is a novel gradient estimator based on a double-momentum mech
anism that combines two recent momentum-based techniques. Utilizing this estimat
or, we design both standard and accelerated algorithms that are robust to the ch
oice of learning rate. Specifically, our methods attain optimal convergence rate
s in both noiseless and noisy stochastic convex optimization scenarios without t
he need for learning rate decay or fine-tuning. We also prove that our approach
maintains optimal performance across a wide spectrum of learning rates, undersco
ring its stability and practicality. Empirical studies further validate the rob
ustness and enhanced stability of our approach.
****
Lun Huang,Qiang Qiu,Guillermo Sapiro
SSOLE: Rethinking Orthogonal Low-rank Embedding for Self-Supervised Learning
Self-supervised learning (SSL) aims to learn meaningful representations from unl
abeled data. Orthogonal Low-rank Embedding (OLE) shows promise for SSL by enhanc
ing intra-class similarity in a low-rank subspace and promoting inter-class diss
imilarity in a high-rank subspace, making it particularly suitable for multi-vie
w learning tasks. However, directly applying OLE to SSL poses significant challe
nges: (1) the virtually infinite number of "classes" in SSL makes achieving the
OLE objective impractical, leading to representational collapse; and (2) low-ran
k constraints may fail to distinguish between positively and negatively correlat
ed features, further undermining learning. To address these issues, we propose S
SOLE (Self-Supervised Orthogonal Low-rank Embedding), a novel framework that int
egrates OLE principles into SSL by (1) decoupling the low-rank and high-rank enf
orcement to align with SSL objectives; and (2) applying low-rank constraints to
feature deviations from their mean, ensuring better alignment of positive pairs

by accounting for the signs of cosine similarities. Our theoretical analysis and empirical results demonstrate that these adaptations are crucial to SSOLE's effectiveness. Moreover, SSOLE achieves competitive performance across SSL benchmarks without relying on large batch sizes, memory banks, or dual-encoder architectures, making it an efficient and scalable solution for self-supervised tasks. Code is available at https://github.com/husthuaan/ssole.
****

Jonas Linkerhägner,Cheng Shi,Ivan Dokmani■
Joint Graph Rewiring and Feature Denoising via Spectral Resonance
When learning from graph data, the graph and the node features both give noisy information about the node labels. In this paper we propose an algorithm to **j**ointly **d**enoise the features and **r**ewire the graph (JDR), which improves the performance of downstream node classification graph neural nets (GNNs). JDR works by aligning the leading spectral spaces of graph and feature matrices. It approximately solves the associated non-convex optimization problem in a way that handles graphs with multiple classes and different levels of homophily or heterophily. We theoretically justify JDR in a stylized setting and show that it consistently outperforms existing rewiring methods on a wide range of synthetic and real-world node classification tasks.
****

Wei Yu,Songheng Yin,Steve Easterbrook,Animesh Garg
EgoSim: Egocentric Exploration in Virtual Worlds with Multi-modal Conditioning
Recent advancements in video diffusion models have established a strong foundation for developing world models with practical applications. The next challenge lies in exploring how an agent can leverage these foundation models to understand, interact with, and plan within observed environments. This requires adding more controllability to the model, transforming it into a versatile game engine capable of dynamic manipulation and control. To address this, we investigated three key conditioning factors: camera, context frame, and text, identifying limitations in current model designs. Specifically, the fusion of camera embeddings with video features leads to camera control being influenced by those features. Additionally, while textual information compensates for necessary spatiotemporal structures, it often intrudes into already observed parts of the scene. To tackle these issues, we designed the Spacetime Epipolar Attention Layer, which ensures that egomotion generated by the model strictly aligns with the camera's movement through rigid constraints. Moreover, we propose the CI2V-adapter, which uses camera information to better determine whether to prioritize textual or visual embeddings, thereby alleviating the issue of textual intrusion into observed areas. Through extensive experiments, we demonstrate that our new model EgoSim achieves excellent results on both the RealEstate and newly repurposed Epic-Field datasets. For more results, please refer to https://egosim.github.io/EgoSim/.
****

Yiliao Song,Zhenqiao Yuan,Shuhai Zhang,Zhen Fang,Jun Yu,Feng Liu
Deep Kernel Relative Test for Machine-generated Text Detection
Recent studies demonstrate that two-sample test can effectively detect machine-generated texts (MGTs) with excellent adaptation ability to texts generated by newer LLMs. However, two-sample test-based detection relies on the assumption that human-written texts (HWTs) must follow the distribution of seen HWTs. As a result, it tends to make mistakes in identifying HWTs that deviate from the seen HWT distribution, limiting their use in sensitive areas like academic integrity verification. To address this issue, we propose to employ non-parametric kernel relative test to detect MGTs by testing whether it is statistically significant that the distribution of a text to be tested is closer to the distribution of HWTs than to the MGTs' distribution. We further develop a kernel optimisation algorithm in relative test to select the best kernel that can enhance the testing capability for MGT detection. As relative test does not assume that a text to be tested must belong exclusively to either MGTs or HWTs, relative test can largely reduce the false positive error compared to two-sample test, offering significant advantages in practice. Extensive experiments demonstrate the superior performance of our method, compared to state-of-the-art non-parametric and parametric de

tectors. The code and demo are available: https://github.com/xLearn-AU/R-Detect.
****
Junyan Ye,Baichuan Zhou,Zilong Huang,Junan Zhang,Tianyi Bai,Hengrui Kang,Jun He,Honglin Lin,Zihao Wang,Tong Wu,Zhizheng Wu,Yiping Chen,Dahua Lin,Conghui He,Weijia Li
LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models
With the rapid development of AI-generated content, the future internet may be inundated with synthetic data, making the discrimination of authentic and credible multimodal data increasingly challenging. Synthetic data detection has thus garnered widespread attention, and the performance of large multimodal models (LMMs) in this task has attracted significant interest. LMMs can provide natural language explanations for their authenticity judgments, enhancing the explainability of synthetic content detection. Simultaneously, the task of distinguishing between real and synthetic data effectively tests the perception, knowledge, and reasoning capabilities of LMMs. In response, we introduce LOKI, a novel benchmark designed to evaluate the ability of LMMs to detect synthetic data across multiple modalities. LOKI encompasses video, image, 3D, text, and audio modalities, comprising 18K carefully curated questions across 26 subcategories with clear difficulty levels. The benchmark includes coarse-grained judgment and multiple-choice questions, as well as fine-grained anomaly selection and explanation tasks, allowing for a comprehensive analysis of LMMs. We evaluated 22 open-source LMMs and 6 closed-source models on LOKI, highlighting their potential as synthetic data detectors and also revealing some limitations in the development of LMM capabilities. More information about LOKI can be found at https://opendatalab.github.io/LOKI/.
****
Zhengwei Yin,Hongjun Wang,Guixu Lin,Weihang Ran,Yinqiang Zheng
Random Is All You Need: Random Noise Injection on Feature Statistics for Generalizable Deep Image Denoising
Recent advancements in generalizable deep image denoising have catalyzed the development of robust noise-handling models. The current state-of-the-art, Masked Training (MT), constructs a masked swinir model which is trained exclusively on Gaussian noise (σ=15) but can achieve commendable denoising performance across various noise types (*i.e.* speckle noise, poisson noise). However, this method, while focusing on content reconstruction, often produces over-smoothed images and poses challenges in mask ratio optimization, complicating its integration with other methodologies. In response, this paper introduces RNINet, a novel architecture built on a streamlined encoder-decoder framework to enhance both efficiency and overall performance. Initially, we train a pure RNINet (only simple encoder-decoder) on individual noise types, observing that feature statistics such as mean and variance shift in response to different noise conditions. Leveraging these insights, we incorporate a noise injection block that injects random noise into feature statistics within our framework, significantly improving generalization across unseen noise types. Our framework not only simplifies the architectural complexity found in MT but also delivers superior performance. Comprehensive experimental evaluations demonstrate that our method outperforms MT in various unseen noise conditions in terms of denoising effectiveness and computational efficiency (lower MACs and GPU memory usage), achieving up to 10 times faster inference speeds and underscoring it's capability for large scale deployments.
****
Jiayi Zhang,Jinyu Xiang,Zhaoyang Yu,Fengwei Teng,Xiong-Hui Chen,Jiaqi Chen,Mingchen Zhuge,Xin Cheng,Sirui Hong,Jinlin Wang,Bingnan Zheng,Bang Liu,Yuyu Luo,Chenglin Wu
AFlow: Automating Agentic Workflow Generation
Large language models (LLMs) have demonstrated remarkable potential in solving complex tasks across diverse domains, typically by employing agentic workflows that follow detailed instructions and operational sequences. However, constructing these workflows requires significant human effort, limiting scalability and generalizability. Recent research has sought to automate the generation and optimiz

ation of these workflows, but existing methods still rely on initial manual setu
p and fall short of achieving fully automated and effective workflow generation.
To address this challenge, we reformulate workflow optimization as a search pro
blem over code-represented workflows, where LLM-invoking nodes are connected by
edges. We introduce AFLOW, an automated framework that efficiently explores this
space using Monte Carlo Tree Search, iteratively refining workflows through cod
e modification, tree-structured experience, and execution feedback. Empirical ev
aluations across six benchmark datasets demonstrate AFLOW's efficacy, yielding a
5.7% average improvement over state-of-the-art baselines. Furthermore, AFLOW en
ables smaller models to outperform GPT-4o on specific tasks at 4.55% of its infe
rence cost in dollars. The code is available at https://github.com/geekan/MetaGP
T.
****

Darko Drakulic,Sofia Michel,Jean-Marc Andreoli

GOAL: A Generalist Combinatorial Optimization Agent Learner

Machine Learning-based heuristics have recently shown impressive performance in
solving a variety of hard combinatorial optimization problems (COPs). However th
ey generally rely on a separate neural model, specialized and trained for each s
ingle problem. Any variation of a problem requires adjustment of its model and r
e-training from scratch. In this paper, we propose GOAL (for Generalist combinat
orial Optimization Agent Learner), a generalist model capable of efficiently sol
ving multiple COPs and which can be fine-tuned to solve new COPs. GOAL consists
of a single backbone plus light-weight problem-specific adapters for input and o
utput processing. The backbone is based on a new form of mixed-attention blocks
which allows to handle problems defined on graphs with arbitrary combinations of
node, edge and instance-level features. Additionally, problems which involve he
terogeneous types of nodes or edges are handled through a novel multi-type trans
former architecture, where the attention blocks are duplicated to attend the mea
ningful combinations of types while relying on the same shared parameters. We tr
ain GOAL on a set of routing, scheduling and classic graph problems and show tha
t it is only slightly inferior to the specialized baselines while being the firs
t multi-task model that solves a wide range of COPs. Finally we showcase the str
ong transfer learning capacity of GOAL by fine-tuning it on several new problems
. Our code is available at https://github.com/naver/goal-co .
****

Brent A. Wallace,Jennie Si

Integral Performance Approximation for Continuous-Time Reinforcement Learning Co
ntrol

We introduce integral performance approximation (IPA), a new continuous-time rei
nforcement learning (CT-RL) control method. It leverages an affine nonlinear dyn
amic model, which partially captures the dynamics of the physical environment, a
longside state-action trajectory data to enable optimal control with great data
efficiency and robust control performance. Utilizing Kleinman algorithm structur
es allows IPA to provide theoretical guarantees of learning convergence, solutio
n optimality, and closed-loop stability. Furthermore, we demonstrate the effecti
veness of IPA on three CT-RL environments including hypersonic vehicle (HSV) con
trol, which has additional challenges caused by unstable and nonminimum phase dy
namics. As a result, we demonstrate that the IPA method leads to new, SOTA contr
ol design and performance in CT-RL.
****

Tao Ren,Zishi Zhang,Jinyang Jiang,Guanghao Li,Zeliang Zhang,Mingqian Feng,Yijie
Peng

FLOPS: Forward Learning with OPtimal Sampling

Given the limitations of backpropagation, perturbation-based gradient computatio
n methods have recently gained focus for learning with only forward passes, also
referred to as queries. Conventional forward learning consumes enormous queries
on each data point for accurate gradient estimation through Monte Carlo samplin
g, which hinders the scalability of those algorithms. However, not all data poin
ts deserve equal queries for gradient estimation. In this paper, we study the pr
oblem of improving the forward learning efficiency from a novel perspective: how

to reduce the gradient estimation variance with minimum cost? For this, we allo
cate the optimal number of queries within a set budget during training to balanc
e estimation accuracy and computational efficiency. Specifically, with a simplif
ied proxy objective and a reparameterization technique, we derive a novel plug-a
nd-play query allocator with minimal parameters. Theoretical results are carried
 out to verify its optimality. We conduct extensive experiments for fine-tuning
Vision Transformers on various datasets and further deploy the allocator to two
black-box applications: prompt tuning and multimodal alignment for foundation mo
dels. All findings demonstrate that our proposed allocator significantly enhance
s the scalability of forward-learning algorithms, paving the way for real-world
applications. The implementation is available at https://github.com/RTkenny/FLOP
S-Forward-Learning-with-OPtimal-Sampling.
****
Anqi Li,Feng Li,Yuxi Liu,Runmin Cong,Yao Zhao,Huihui Bai
Once-for-All: Controllable Generative Image Compression with Dynamic Granularity
 Adaptation
Although recent generative image compression methods have demonstrated impressiv
e potential in optimizing the rate-distortion-perception trade-off, they still f
ace the critical challenge of flexible rate adaptation to diverse compression ne
cessities and scenarios. To overcome this challenge, this paper proposes a $\tex
tbf{Control}$lable $\textbf{G}$enerative $\textbf{I}$mage $\textbf{C}$ompression
 framework, $\textbf{Control-GIC}$, the first capable of fine-grained bitrate ad
aptation across a broad spectrum while ensuring high-fidelity and generality com
pression. We base Control-GIC on a VQGAN framework representing an image as a se
quence of variable-length codes ($\textit{i.e.}$ VQ-indices), which can be lossl
essly compressed and exhibits a direct positive correlation with bitrates. Drawi
ng inspiration from the classical coding principle, we correlate the information
 density of local image patches with their granular representations. Hence, we c
an flexibly determine a proper allocation of granularity for the patches to achi
eve dynamic adjustment for VQ-indices, resulting in desirable compression rates.
 We further develop a probabilistic conditional decoder capable of retrieving hi
storic encoded multi-granularity representations according to transmitted codes,
 and then reconstruct hierarchical granular features in the formalization of con
ditional probability, enabling more informative aggregation to improve reconstru
ction realism. Our experiments show that Control-GIC allows highly flexible and
controllable bitrate adaptation where the results demonstrate its superior perfo
rmance over recent state-of-the-art methods.
****
Shijie Chen,Bernal Jimenez Gutierrez,Yu Su
Attention in Large Language Models Yields Efficient Zero-Shot Re-Rankers
Information retrieval (IR) systems have played a vital role in modern digital li
fe and have cemented their continued usefulness in this new era of generative AI
 via retrieval-augmented generation. With strong language processing capabilitie
s and remarkable versatility, large language models (LLMs) have become popular c
hoices for zero-shot re-ranking in IR systems. So far, LLM-based re-ranking meth
ods rely on strong generative capabilities, which restricts their use to either
specialized or powerful proprietary models. Given these restrictions, we ask: is
 autoregressive generation necessary and optimal for LLMs to perform re-ranking?
 We hypothesize that there are abundant signals relevant to re-ranking within LL
Ms that might not be used to their full potential via generation. To more direc
tly leverage such signals, we propose in-context re-ranking (ICR), a novel metho
d that leverages the change in attention pattern caused by the search query for
 accurate and efficient re-ranking. We assume that more relevant documents shoul
d receive more attention weights when an LLM is processing the query tokens, and
 leverage such signals for re-ranking. To mitigate the intrinsic biases in LLMs,
 we propose a calibration method using a content-free query. Due to the absence
of generation, ICR only requires two ($O(1)$) forward passes to re-rank N docu
ments, making it substantially more efficient than generative re-ranking methods
 that require at least $O(N)$ forward passes. Our novel design also enables ICR
to be applied to any LLM without specialized training while guaranteeing a well-

formed ranking. Extensive experiments with two popular open-weight LLMs on stand
ard single-hop and multi-hop information retrieval benchmarks show that ICR outp
erforms RankGPT while cutting the latency by more than 60% in practice. Through
detailed analyses, we show that ICR's performance is specially strong on tasks t
hat require more complex re-ranking signals, such as handling contextualization
and contradiction between the query and passages, as well as information integra
tion across multiple passages. Our findings call for further exploration on nove
l ways of utilizing open-weight LLMs beyond text generation.
****

Shaocong Ma,Heng Huang

Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and
  Directionally Aligned Perturbations

In this paper, we explore the two-point zeroth-order gradient estimator and iden
tify the distribution of random perturbations that minimizes the estimator's asy
mptotic variance as the perturbation stepsize tends to zero. We formulate it as
a constrained functional optimization problem over the space of perturbation dis
tributions. Our findings reveal that such desired perturbations can align direct
ionally with the true gradient, instead of maintaining a fixed length. While exi
sting research has largely focused on fixed-length perturbations, the potential
advantages of directional alignment have been overlooked. To address this gap, w
e delve into the theoretical and empirical properties of the directionally align
ed perturbation (DAP) scheme, which adaptively offers higher accuracy along crit
ical directions. Additionally, we provide a convergence analysis for stochastic
gradient descent using δ-unbiased random perturbations, extending existin
g complexity bounds to a wider range of perturbations. Through empirical evaluat
ions on both synthetic problems and practical tasks, we demonstrate that DAPs ou
tperform traditional methods under specific conditions.
****

Tung Nguyen,Aditya Grover

LICO: Large Language Models for In-Context Molecular Optimization

Optimizing black-box functions is a fundamental problem in science and engineeri
ng. To solve this problem, many approaches learn a surrogate function that estim
ates the underlying objective from limited historical evaluations. Large Languag
e Models (LLMs), with their strong pattern-matching capabilities via pretraining
 on vast amounts of data, stand out as a potential candidate for surrogate model
ing. However, directly prompting a pretrained language model to produce predicti
ons is not feasible in many scientific domains due to the scarcity of domain-spe
cific data in the pretraining corpora and the challenges of articulating complex
 problems in natural language. In this work, we introduce LICO, a general-purpos
e model that extends arbitrary base LLMs for black-box optimization, with a part
icular application to the molecular domain. To achieve this, we equip the langua
ge model with a separate embedding layer and prediction layer, and train the mod
el to perform in-context predictions on a diverse set of functions defined over
the domain. Once trained, LICO can generalize to unseen molecule properties simp
ly via in-context prompting. LICO performs competitively on PMO, a challenging m
olecular optimization benchmark comprising 23 objective functions, and achieves
state-of-the-art performance on its low-budget version PMO-1K.
****

Nan Zhang,Prafulla Kumar Choubey,Alexander Fabbri,Gabriel Bernadett-Shapiro,Rui
Zhang,Prasenjit Mitra,Caiming Xiong,Chien-Sheng Wu

SiReRAG: Indexing Similar and Related Information for Multihop Reasoning

Indexing is an important step towards strong performance in retrieval-augmented
generation (RAG) systems. However, existing methods organize data based on eithe
r semantic similarity (similarity) or related information (relatedness), but do
not cover both perspectives comprehensively. Our analysis reveals that modeling
only one perspective results in insufficient knowledge synthesis, leading to sub
optimal performance on complex tasks requiring multihop reasoning. In this paper
, we propose SiReRAG, a novel RAG indexing approach that explicitly considers bo
th similar and related information. On the similarity side, we follow existing w
ork and explore some variances to construct a similarity tree based on recursive

summarization. On the relatedness side, SiReRAG extracts propositions and entit
ies from texts, groups propositions via shared entities, and generates recursive
summaries to construct a relatedness tree. We index and flatten both similarity
and relatedness trees into a unified retrieval pool. Our experiments demonstrat
e that SiReRAG consistently outperforms state-of-the-art indexing methods on thr
ee multihop datasets (MuSiQue, 2WikiMultiHopQA, and HotpotQA), with an average 1
.9% improvement in F1 scores. As a reasonably efficient solution, SiReRAG enhanc
es existing reranking methods significantly, with up to 7.8% improvement in aver
age F1 scores. Our code is available at https://github.com/SalesforceAIResearch/
SiReRAG.
****

Xiang Lisa Li,Farzaan Kaiyom,Evan Zheran Liu,Yifan Mai,Percy Liang,Tatsunori Has
himoto
AutoBencher: Towards Declarative Benchmark Construction
We present AutoBencher, a declarative framework for automatic benchmark construc
tion, and use it to scalably discover novel insights and vulnerabilities of exis
ting language models. Concretely, given a few desiderata of benchmarks (e.g., qu
estion difficulty, topic salience), we operationalize each desideratum and cast
benchmark creation as an optimization problem. Specifically, we experiment with
two settings with different optimization objectives: (i) for capability evaluati
on, we declare the goal of finding a salient, difficult dataset that induces nov
el performance patterns; (ii) for safety evaluation, we declare the goal of find
ing a dataset of unsafe prompts that existing LMs fail to decline. To tackle thi
s optimization problem, we use a language model to iteratively propose and refin
e dataset descriptions, which are then used to generate topic-specific questions
and answers. These descriptions are optimized to improve the declared desiderat
a. We use AutoBencher (powered by GPT-4) to create datasets for math, multilingu
ality, knowledge, and safety. The scalability of AutoBencher allows it to test f
ine-grained categories and tail knowledge, creating datasets that elicit 22% mor
e model errors (i.e., difficulty) than existing benchmarks. On the novelty ends,
AutoBencher also helps identify specific gaps not captured by existing benchmar
ks: e.g., Gemini-Pro has knowledge gaps on Permian Extinction and Fordism while
GPT-4o fails to decline harmful requests about cryptocurrency scams.
****

Yurou Liu,Jiahao Chen,Rui Jiao,Jiangmeng Li,Wenbing Huang,Bing Su
DenoiseVAE: Learning Molecule-Adaptive Noise Distributions for Denoising-based 3
D Molecular Pre-training
Denoising learning of 3D molecules learns molecular representations by imposing
noises into the equilibrium conformation and predicting the added noises to reco
ver the equilibrium conformation, which essentially captures the information of
molecular force fields. Due to the specificity of Potential Energy Surfaces, the
probabilities of physically reasonable noises for each atom in different molecu
les are different. However, existing methods apply the shared heuristic hand-cra
fted noise sampling strategy to all molecules, resulting in inaccurate force fie
ld learning. In this paper, we propose a novel 3D molecular pre-training method,
namely DenoiseVAE, which employs a Noise Generator to acquire atom-specific noi
se distributions for different molecules. It utilizes the stochastic reparameter
ization technique to sample noisy conformations from the generated distributions
, which are inputted into a Denoising Module for denoising. The Noise Generator
and the Denoising Module are jointly learned in a manner conforming with the par
adigm of Variational Auto Encoder. Consequently, the sampled noisy conformations
can be more diverse, adaptive, and informative, and thus DenoiseVAE can learn r
epresentations that better reveal the molecular force fields. Extensive experime
nts show that DenoiseVAE outperforms the current state-of-the-art methods on var
ious molecular property prediction tasks, demonstrating the effectiveness of it.
****

Hongjin SU,Howard Yen,Mengzhou Xia,Weijia Shi,Niklas Muennighoff,Han-yu Wang,Liu
Haisu,Quan Shi,Zachary S Siegel,Michael Tang,Ruoxi Sun,Jinsung Yoon,Sercan O Ar
ik,Danqi Chen,Tao Yu
BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval

Existing retrieval benchmarks primarily consist of information-seeking queries (e.g., aggregated questions from search engines) where keyword or semantic-based retrieval is usually sufficient. However, many complex real-world queries requir e in-depth reasoning to identify relevant documents that go beyond surface form matching. For example, finding documentation for a coding question requires unde rstanding the logic and syntax of the functions involved. To better benchmark re trieval on such challenging queries, we introduce BRIGHT, the first text retriev al benchmark that requires intensive reasoning to retrieve relevant documents. O ur dataset consists of 1,398 real-world queries spanning diverse domains such as economics, psychology, mathematics, coding, and more. These queries are drawn f rom naturally occurring or carefully curated human data. Extensive evaluation re veals that even state-of-the-art retrieval models perform poorly on BRIGHT. The leading model on the MTEB leaderboard (Muennighoff et al., 2023), which achieves a score of 59.0 nDCG@10,1 produces a score of nDCG@10 of 18.0 on BRIGHT. We sho w that incorporating explicit reasoning about the query improves retrieval perfo rmance by up to 12.2 points. Moreover, incorporating retrieved documents from th e top-performing retriever boosts question answering performance by over 6.6 poi nts. We believe that BRIGHT paves the way for future research on retrieval syste ms in more realistic and challenging settings.
****
Haizhong Zheng,Elisa Tsai,Yifu Lu,Jiachen Sun,Brian R. Bartoldson,Bhavya Kailkhu ra,Atul Prakash
ELFS: Label-Free Coreset Selection with Proxy Training Dynamics
High-quality human-annotated data is crucial for modern deep learning pipelines, yet the human annotation process is both costly and time-consuming. Given a con strained human labeling budget, selecting an informative and representative data subset for labeling can significantly reduce human annotation effort. Well-perf orming state-of-the-art (SOTA) coreset selection methods require ground truth la bels over the whole dataset, failing to reduce the human labeling burden. Meanwh ile, SOTA label-free coreset selection methods deliver inferior performance due to poor geometry-based difficulty scores. In this paper, we introduce ELFS (Effe ctive Label-Free Coreset Selection), a novel label-free coreset selection method . ELFS significantly improves label-free coreset selection by addressing two cha llenges: 1) ELFS utilizes deep clustering to estimate training dynamics-based da ta difficulty scores without ground truth labels; 2) Pseudo-labels introduce a d istribution shift in the data difficulty scores, and we propose a simple but eff ective double-end pruning method to mitigate bias on calculated scores. We evalu ate ELFS on four vision benchmarks and show that, given the same vision encoder, ELFS consistently outperforms SOTA label-free baselines. For instance, when usi ng SwAV as the encoder, ELFS outperforms D2 by up to 10.2% in accuracy on ImageN et-1K. We make our code publicly available on GitHub.
****
Xiaojuan Wang,Boyang Zhou,Brian Curless,Ira Kemelmacher-Shlizerman,Aleksander Ho lynski,Steve Seitz
Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolati on
We present a method for generating video sequences with coherent motion between a pair of input keyframes. We adapt a pretrained large-scale image-to-video diff usion model (originally trained to generate videos moving forward in time from a single input image) for keyframe interpolation, i.e., to produce a video betwee n two input frames. We accomplish this adaptation through a lightweight fine-tun ing technique that produces a version of the model that instead predicts videos moving backwards in time from a single input image. This model (along with the o riginal forward-moving model) is subsequently used in a dual-directional diffusi on sampling process that combines the overlapping model estimates starting from each of the two keyframes. Our experiments shows that our method outperforms bot h existing diffusion-based methods and traditional frame interpolation technique s.
****
Alexander H. Liu,Sang-gil Lee,Chao-Han Huck Yang,Yuan Gong,Yu-Chiang Frank Wang,

James R. Glass,Rafael Valle,Bryan Catanzaro
UniWav: Towards Unified Pre-training for Speech Representation Learning and Generation
Pre-training and representation learning have been playing an increasingly important role in modern speech processing. Nevertheless, different applications have been relying on different foundation models, since predominant pre-training techniques are either designed for discriminative tasks or generative tasks. In this work, we make the first attempt at building a unified pre-training framework for both types of tasks in speech. We show that with the appropriate design choices for pre-training, one can jointly learn a representation encoder and generative audio decoder that can be applied to both types of tasks. We propose UniWav, an encoder-decoder framework designed to unify pre-training representation learning and generative tasks. On speech recognition, text-to-speech, and speech tokenization, UniWav achieves comparable performance to different existing foundation models, each trained on a specific task. Our findings suggest that a single general-purpose foundation model for speech can be built to replace different foundation models, reducing the overhead and cost of pre-training.
****
Qintong Li,Jiahui Gao,Sheng Wang,Renjie Pi,Xueliang Zhao,Chuan Wu,Xin Jiang,Zhenguo Li,Lingpeng Kong
Forewarned is Forearmed: Harnessing LLMs for Data Synthesis via Failure-induced Exploration
Large language models (LLMs) have significantly benefited from training on diverse, high-quality task-specific data, leading to impressive performance across a range of downstream applications. Current methods often rely on human-annotated data or predefined task templates to direct powerful LLMs in synthesizing task-relevant data for effective model training. However, this dependence on manually designed components may constrain the scope of generated data, potentially overlooking critical edge cases or novel scenarios that could challenge the model. In this paper, we present a novel approach, ReverseGen, designed to automatically generate effective training samples that expose the weaknesses of LLMs. Specifically, we introduce a dedicated proposer trained to produce queries that lead target models to generate unsatisfactory responses. These failure-inducing queries are then used to construct training data, helping to address the models' shortcomings and improve overall performance. Our approach is flexible and can be applied to models of various scales (3B, 7B, and 8B). We evaluate ReverseGen on three key applications—safety, honesty, and math—demonstrating that our generated data is both highly effective and diverse. Models fine-tuned with ReverseGen-generated data consistently outperform those trained on human-annotated or general model-generated data, offering a new perspective on data synthesis for task-specific LLM enhancement.
****
Haotian Wang,Haoxuan Li,Hao Zou,Haoang Chi,Long Lan,Wanrong Huang,Wenjing Yang
Effective and Efficient Time-Varying Counterfactual Prediction with State-Space Models
Time-varying counterfactual prediction (TCP) from observational data supports the answer of when and how to assign multiple sequential treatments, yielding importance in various applications. Despite the progress achieved by recent advances, e.g., LSTM or Transformer based causal approaches, their capability of capturing interactions in long sequences remains to be improved in both prediction performance and running efficiency. In parallel with the development of TCP, the success of the state-space models (SSMs) has achieved remarkable progress toward long-sequence modeling with saved running time. Consequently, studying how Mamba simultaneously benefits the effectiveness and efficiency of TCP becomes a compelling research direction. In this paper, we propose to exploit advantages of the SSMs to tackle the TCP task, by introducing a counterfactual Mamba model with Covariate-based Decorrelation towards Selective Parameters (Mamba-CDSP). Motivated by the over-balancing problem in TCP of the direct covariate balancing methods, we propose to de-correlate between the current treatment and the representation of historical covariates, treatments, and outcomes, which can mitigate the conf

ounding bias while preserve more covariate information. In addition, we show tha
t the overall de-correlation in TCP is equivalent to regularizing the selective
parameters of Mamba over each time step, which leads our approach to be effectiv
e and lightweight. We conducted extensive experiments on both synthetic and real
-world datasets, demonstrating that Mamba-CDSP not only outperforms baselines by
 a large margin, but also exhibits prominent running efficiency.
****
Robert Hönig,Javier Rando,Nicholas Carlini,Florian Tramèr
Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI
Artists are increasingly concerned about advancements in image generation models
 that can closely replicate their unique artistic styles.
In response, several protection tools against style mimicry have been developed
that incorporate small adversarial perturbations into artworks published online.
 In this work, we evaluate the effectiveness of popular protections---with milli
ons of downloads---and show they only provide a false sense of security. We find
 that low-effort and "off-the-shelf" techniques, such as image upscaling, are su
fficient to create robust mimicry methods that significantly degrade existing pr
otections. Through a user study, we demonstrate that **all existing protections
can be easily bypassed**, leaving artists vulnerable to style mimicry. We cauti
on that tools based on adversarial perturbations cannot reliably protect artists
 from the misuse of generative AI, and urge the development of alternative prote
ctive solutions.
****
Yixuan Wang,Jonathan W. Siegel,Ziming Liu,Thomas Y. Hou
On the expressiveness and spectral bias of KANs
Kolmogorov-Arnold Networks (KAN) \cite{liu2024kan} were very recently proposed a
s a potential alternative to the prevalent architectural backbone of many deep l
earning models, the multi-layer perceptron (MLP). KANs have seen success in vari
ous tasks of AI for science, with their empirical efficiency and accuracy demons
trated in function regression, PDE solving, and many more scientific problems.

In this article, we revisit the comparison of KANs and MLPs, with emphasis on a
theoretical perspective. On the one hand, we compare the representation and appr
oximation capabilities of KANs and MLPs. We establish that MLPs can be represent
ed using KANs of a comparable size. This shows that the approximation and repres
entation capabilities of KANs are at least as good as MLPs. Conversely, we show
that KANs can be represented using MLPs, but that in this representation the num
ber of parameters increases by a factor of the KAN grid size. This suggests that
 KANs with a large grid size may be more efficient than MLPs at approximating ce
rtain functions. On the other hand, from the perspective of learning and optimiz
ation, we study the spectral bias of KANs compared with MLPs. We demonstrate tha
t KANs are less biased toward low frequencies than MLPs. We highlight that the m
ulti-level learning feature specific to KANs, i.e. grid extension of splines, im
proves the learning process for high-frequency components. Detailed comparisons
 with different choices of depth, width, and grid sizes of KANs are made, sheddi
ng some light on how to choose the hyperparameters in practice.
****
Milad Khademi Nori,IL MIN KIM,Guanghui Wang
Federated Class-Incremental Learning: A Hybrid Approach Using Latent Exemplars a
nd Data-Free Techniques to Address Local and Global Forgetting
Federated Class-Incremental Learning (FCIL) refers to a scenario where a dynamic
ally changing number of clients collaboratively learn an ever-increasing number
of incoming tasks. FCIL is known to suffer from local forgetting due to class im
balance at each client and global forgetting due to class imbalance across clien
ts. We develop a mathematical framework for FCIL that formulates local and globa
l forgetting. Then, we propose an approach called Hybrid Rehearsal (HR), which u
tilizes latent exemplars and data-free techniques to address local and global fo
rgetting, respectively. HR employs a customized autoencoder designed for both da
ta classification and the generation of synthetic data. To determine the embeddi
ngs of new tasks for all clients in the latent space of the encoder, the server

uses the Lennard-Jones Potential formulations. Meanwhile, at the clients, the de
coder decodes the stored low-dimensional latent space exemplars back to the high
-dimensional input space, used to address local forgetting. To overcome global f
orgetting, the decoder generates synthetic data. Furthermore, our mathematical f
ramework proves that our proposed approach HR can, in principle, tackle the two
local and global forgetting challenges. In practice, extensive experiments demon
strate that while preserving privacy, our proposed approach outperforms the stat
e-of-the-art baselines on multiple FCIL benchmarks with low compute and memory f
ootprints.
****

Adriana Eufrosina Bora,Pierre-Luc St-Charles,Mirko Bronzi,Arsene Fansi Tchango,B
runo Rousseau,Kerrie Mengersen

AIMS.au: A Dataset for the Analysis of Modern Slavery Countermeasures in Corpora
te Statements

Despite over a decade of legislative efforts to address modern slavery in the su
pply chains of large corporations, the effectiveness of government oversight rem
ains hampered by the challenge of scrutinizing thousands of statements annually.
 While Large Language Models (LLMs) can be considered a well established solutio
n for the automatic analysis and summarization of documents, recognizing concret
e modern slavery countermeasures taken by companies and differentiating those fr
om vague claims remains a challenging task. To help evaluate and fine-tune LLMs
for the assessment of corporate statements, we introduce a dataset composed of 5
,731 modern slavery statements taken from the Australian Modern Slavery Register
 and annotated at the sentence level. This paper details the construction steps
for the dataset that include the careful design of annotation specifications, th
e selection and preprocessing of statements, and the creation of high-quality an
notation subsets for effective model evaluations. To demonstrate our dataset's u
tility, we propose a machine learning methodology for the detection of sentences
 relevant to mandatory reporting requirements set by the Australian Modern Slave
ry Act. We then follow this methodology to benchmark modern language models unde
r zero-shot and supervised learning settings.
****

Rongfeng Lu,Hangyu Chen,Zunjie Zhu,Yuhang Qin,Ming Lu,Le zhang,Chenggang Yan,ank
e xue

ThermalGaussian: Thermal 3D Gaussian Splatting

Thermography is especially valuable for the military and other users of surveill
ance cameras. Some recent methods based on Neural Radiance Fields (NeRF) are pro
posed to reconstruct the thermal scenes in 3D from a set of thermal and RGB imag
es. However, unlike NeRF, 3D Gaussian splatting (3DGS) prevails due to its rapid
 training and real-time rendering. In this work, we propose ThermalGaussian, the
 first thermal 3DGS approach capable of rendering high-quality images in RGB and
 thermal modalities. We first calibrate the RGB camera and the thermal camera to
 ensure that both modalities are accurately aligned. Subsequently, we use the re
gistered images to learn the multimodal 3D Gaussians. To prevent the overfitting
 of any single modality, we introduce several multimodal regularization constrai
nts. We also develop smoothing constraints tailored to the physical characterist
ics of the thermal modality.

Besides, we contribute a real-world dataset named RGBT-Scenes, captured by a han
d-hold thermal-infrared camera, facilitating future research on thermal scene re
construction. We conduct comprehensive experiments to show that ThermalGaussian
achieves photorealistic rendering of thermal images and improves the rendering q
uality of RGB images. With the proposed multimodal regularization constraints, w
e also reduced the model's storage cost by 90\%. Our project page is at https://
thermalgaussian.github.io/.
****

Girish Narayanswamy,Xin Liu,Kumar Ayush,Yuzhe Yang,Xuhai Xu,shun liao,Jake Garri
son,Shyam A. Tailor,Jacob Sunshine,Yun Liu,Tim Althoff,Shrikanth Narayanan,Pushm
eet Kohli,Jiening Zhan,Mark Malhotra,Shwetak Patel,Samy Abdel-Ghaffar,Daniel McD
uff

Scaling Wearable Foundation Models

Wearable sensors have become ubiquitous thanks to a variety of health tracking f
eatures. The resulting continuous and longitudinal measurements from everyday li
fe generate large volumes of data. However, making sense of these observations f
or scientific and actionable insights is non-trivial. Inspired by the empirical
success of generative modeling, where large neural networks learn powerful repre
sentations from vast amounts of text, image, video, or audio data, we investigat
e the scaling properties of wearable sensor foundation models across compute, da
ta, and model size. Using a dataset of up to 40 million hours of in-situ heart r
ate, heart rate variability, accelerometer, electrodermal activity, skin tempera
ture, and altimeter per-minute data from over 165,000 people, we create LSM, a m
ultimodal foundation model built on the largest wearable-signals dataset with th
e most extensive range of sensor modalities to date. Our results establish the s
caling laws of LSM for tasks such as imputation, interpolation and extrapolation
across both time and sensor modalities. Moreover, we highlight how LSM enables
sample-efficient downstream learning for tasks including exercise and activity r
ecognition.
****
Bofei Gao,Feifan Song,Zhe Yang,Zefan Cai,Yibo Miao,Qingxiu Dong,Lei Li,Chenghao
Ma,Liang Chen,Runxin Xu,Zhengyang Tang,Benyou Wang,Daoguang Zan,Shanghaoran Quan
,Ge Zhang,Lei Sha,Yichang Zhang,Xuancheng Ren,Tianyu Liu,Baobao Chang
Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Mo
dels
Recent advancements in large language models (LLMs) have led to significant brea
kthroughs in mathematical reasoning capabilities.
However, existing benchmarks like GSM8K or MATH are now being solved with high a
ccuracy (e.g., OpenAI o1 achieves 94.8% on MATH dataset), indicating their inade
quacy for truly challenging these models. To bridge this gap, we propose a compr
ehensive and challenging benchmark specifically designed to assess LLMs' mathema
tical reasoning at the Olympiad level. Unlike existing Olympiad-related benchmar
ks, our dataset focuses exclusively on mathematics and comprises a vast collecti
on of 4428 competition-level problems with rigorous human annotation. These prob
lems are meticulously categorized into over 33 sub-domains and span more than 10
distinct difficulty levels, enabling a holistic assessment of model performance
in Olympiad-mathematical reasoning. Furthermore, we conducted an in-depth analy
sis based on this benchmark. Our experimental results show that even the most ad
vanced models, OpenAI o1-mini and OpenAI o1-preview, struggle with highly challe
nging Olympiad-level problems, with 60.54% and 52.55% accuracy, highlighting sig
nificant challenges in Olympiad-level mathematical reasoning.
****
Jang Hyun Cho,Boris Ivanovic,Yulong Cao,Edward Schmerling,Yue Wang,Xinshuo Weng,
Boyi Li,Yurong You,Philipp Kraehenbuehl,Yan Wang,Marco Pavone
Language-Image Models with 3D Understanding
Multi-modal large language models (MLLMs) have shown incredible capabilities in
a variety of 2D vision and language tasks. We extend MLLMs’ perceptual capabilit
ies to ground and reason about images in 3-dimensional space. To that end, we fi
rst develop a large-scale pretraining dataset for 2D and 3D called LV3D by combi
ning multiple existing 2D and 3D recognition datasets under a common task formul
ation: as multi-turn question-answering. Next, we introduce a new MLLM named CUB
E-LLM and pre-train it on LV3D. We show that pure data scaling makes a strong 3D
perception capability without 3D specific architectural design or training obje
ctive. CUBE-LLM exhibits intriguing properties similar to LLMs: (1) CUBE-LLM can
apply chain-of-thought prompting to improve 3D understanding from 2D context in
formation. (2) CUBE-LLM can follow complex and diverse instructions and adapt to
versatile input and output formats. (3) CUBE-LLM can be visually prompted such
as 2D box or a set of candidate 3D boxes from specialists. Our experiments on ou
tdoor benchmarks demonstrate that CUBE-LLM significantly outperforms existing ba
selines by 21.3 points of AP-BEV on the Talk2Car dataset for 3D grounded reasoni
ng and 17.7 points on the DriveLM dataset for complex reasoning about driving sc
enarios, respectively. CUBE-LLM also shows competitive results in general MLLM b
enchmarks such as refCOCO for 2D grounding with (87.0) average score, as well as

visual question answering benchmarks such as VQAv2, GQA, SQA, POPE, etc. for co
mplex reasoning.
****
Zheng Yi Ho,Siyuan Liang,Sen Zhang,Yibing Zhan,Dacheng Tao
NoVo: Norm Voting off Hallucinations with Attention Heads in Large Language Mode
ls
Hallucinations in Large Language Models (LLMs) remain a major obstacle, particul
arly in high-stakes applications where factual accuracy is critical. While repre
sentation editing and reading methods have made strides in reducing hallucinatio
ns, their heavy reliance on specialised tools and training on in-domain samples,
makes them difficult to scale and prone to overfitting. This limits their accur
acy gains and generalizability to diverse datasets. This paper presents a lightw
eight method, Norm Voting (NoVo), which harnesses the untapped potential of atte
ntion head norms to dramatically enhance factual accuracy in zero-shot multiple-
choice questions (MCQs). NoVo begins by automatically selecting truth-correlated
head norms with an efficient, inference-only algorithm using only 30 random sam
ples, allowing NoVo to effortlessly scale to diverse datasets. Afterwards, selec
ted head norms are employed in a simple voting algorithm, which yields significa
nt gains in prediction accuracy. On TruthfulQA MC1, NoVo surpasses the current s
tate-of-the-art and all previous methods by an astounding margin---at least 19 a
ccuracy points. NoVo demonstrates exceptional generalization to 20 diverse datas
ets, with significant gains in over 90\% of them, far exceeding all current repr
esentation editing and reading methods. NoVo also reveals promising gains to fin
etuning strategies and building textual adversarial defence. NoVo's effectivenes
s with head norms opens new frontiers in LLM interpretability, robustness and re
liability. Our code is available at: https://github.com/hozhengyi/novo
****
Fan Wu,Emily Black,Varun Chandrasekaran
Generative Monoculture in Large Language Models
We introduce {\em generative monoculture}, a behavior observed in large language
models (LLMs) characterized by a significant narrowing of model output diversit
y relative to available training data for a given task: for example, generating
only positive book reviews for books with a mixed reception. While in some cases
, generative monoculture enhances performance (e.g., LLMs more often produce eff
icient code), the dangers are exacerbated in others (e.g., LLMs refuse to share
diverse opinions). As LLMs are increasingly used in high-impact settings such as
education and web search, careful maintenance of LLM output diversity is essent
ial to ensure a variety of facts and perspectives are preserved over time. We ex
perimentally demonstrate the prevalence of generative monoculture through analys
is of book review and code generation tasks, and find that simple countermeasure
s such as altering sampling or prompting strategies are insufficient to mitigate
the behavior. Moreover, our results suggest that the root causes of generative
monoculture are likely embedded within the LLM's alignment processes, suggesting
a need for developing fine-tuning paradigms that preserve or promote diversity.
****
Chang Zou,Xuyang Liu,Ting Liu,Siteng Huang,Linfeng Zhang
Accelerating Diffusion Transformers with Token-wise Feature Caching
Diffusion transformers have shown significant effectiveness in both image and vi
deo synthesis at the expense of huge computation costs. To address this problem,
feature caching methods have been introduced to accelerate diffusion transforme
rs by caching the features in previous timesteps and reusing them in the followi
ng timesteps. However, previous caching methods ignore that different tokens exh
ibit different sensitivities to feature caching, and feature caching on some tok
ens may lead to 10$\times$ more destruction to the overall generation quality co
mpared with other tokens. In this paper, we introduce token-wise feature caching
, allowing us to adaptively select the most suitable tokens for caching, and fur
ther enable us to apply different caching ratios to neural layers in different t
ypes and depths. Extensive experiments on PixArt-alpha, OpenSora, and DiT demons
trate our effectiveness in both image and video generation with no requirements
for training. For instance, 2.36$\times$ and 1.93$\times$ acceleration are achie

ved on OpenSora and PixArt-α with almost no drop in generation quality. C
odes have been released in the supplementary material and Github.
****

Yuchen Zhou,Jiayuan Gu,Tung Yen Chiang,Fanbo Xiang,Hao Su
Point-SAM: Promptable 3D Segmentation Model for Point Clouds
The development of 2D foundation models for image segmentation has been signific
antly advanced by the Segment Anything Model (SAM). However, achieving similar s
uccess in 3D models remains a challenge due to issues such as non-unified data f
ormats, poor model scalability, and the scarcity of labeled data with diverse ma
sks. To this end, we propose a 3D promptable segmentation model Point-SAM, focus
ing on point clouds. We employ an efficient transformer-based architecture tailo
red for point clouds, extending SAM to the 3D domain. We then distill the rich k
nowledge from 2D SAM for Point-SAM training by introducing a data engine to gene
rate part-level and object-level pseudo-labels at scale from 2D SAM. Our model o
utperforms state-of-the-art 3D segmentation models on several indoor and outdoor
 benchmarks and demonstrates a variety of applications, such as interactive 3D a
nnotation and zero-shot 3D instance proposal.
****

Michael Eli Sander,Gabriel Peyré
Towards Understanding the Universality of Transformers for Next-Token Prediction
Causal Transformers are trained to predict the next token for a given context. W
hile it is widely accepted that self-attention is crucial for encoding the causa
l structure of sequences, the precise underlying mechanism behind this in-contex
t autoregressive learning ability remains unclear. In this paper, we take a step
 towards understanding this phenomenon by studying the approximation ability of
Transformers for next-token prediction. Specifically, we explore the capacity of
 causal Transformers to predict the next token x_{t+1} given an autoregressive
 sequence $(x_1, \dots, x_t)$ as a prompt, where $ x_{t+1} = f(x_t) $, and $ f $
 is a context-dependent function that varies with each sequence.
On the theoretical side, we focus on specific instances, namely when $ f $ is li
near or when $ (x_t)$ is periodic. We explicitly construct a Transformer (with l
inear, exponential, or softmax attention) that learns the mapping f in-context
 through a causal kernel descent method. The causal kernel descent method we pro
pose provably estimates $x_{t+1} $ based solely on past and current observations
 $ (x_1, \dots, x_t) $, with connections to the Kaczmarz algorithm in Hilbert sp
aces. We present experimental results that validate our theoretical findings and
 suggest their applicability to more general mappings f.
****

Ao Li,Wei Fang,Hongbo Zhao,Le Lu,Ge Yang,Minfeng Xu
MaRS: A Fast Sampler for Mean Reverting Diffusion based on ODE and SDE Solvers
In applications of diffusion models, controllable generation is of practical sig
nificance, but is also challenging. Current methods for controllable generation
primarily focus on modifying the score function of diffusion models, while Mean
Reverting (MR) Diffusion directly modifies the structure of the stochastic diffe
rential equation (SDE), making the incorporation of image conditions simpler and
 more natural. However, current training-free fast samplers are not directly app
licable to MR Diffusion. And thus MR Diffusion requires hundreds of NFEs (number
 of function evaluations) to obtain high-quality samples. In this paper, we prop
ose a new algorithm named MaRS (MR Sampler) to reduce the sampling NFEs of MR Di
ffusion. We solve the reverse-time SDE and the probability flow ordinary differe
ntial equation (PF-ODE) associated with MR Diffusion, and derive semi-analytical
 solutions. The solutions consist of an analytical function and an integral para
meterized by a neural network. Based on this solution, we can generate high-qual
ity samples in fewer steps. Our approach does not require training and supports
all mainstream parameterizations, including noise prediction, data prediction an
d velocity prediction. Extensive experiments demonstrate that MR Sampler maintai
ns high sampling quality with a speedup of 10 to 20 times across ten different i
mage restoration tasks. Our algorithm accelerates the sampling procedure of MR D
iffusion, making it more practical in controllable generation.
****

Qiqiang Lin,Muning Wen,Qiuying Peng,Guanyu Nie,Junwei Liao,Jun Wang,Xiaoyun Mo,Jiamu Zhou,Cheng Cheng,Yin Zhao,Jun Wang,Weinan Zhang
Robust Function-Calling for On-Device Language Model via Function Masking
Large language models have demonstrated impressive value in performing as autonomous agents when equipped with external tools and API calls. Nonetheless, effectively harnessing their potential for executing complex tasks crucially relies on enhancements in their function-calling capabilities. This paper identifies a critical gap in existing function-calling models, where performance varies significantly across benchmarks, often due to over-fitting to specific naming conventions. To address such an issue, we introduce Hammer, a novel family of foundation models specifically engineered for on-device function calling. Hammer employs an augmented dataset that enhances models' sensitivity to irrelevant functions and incorporates function masking techniques to minimize over-fitting. Our empirical evaluations reveal that Hammer not only outperforms larger models but also demonstrates robust generalization across diverse benchmarks, achieving state-of-the-art results. Our open-source contributions include a specialized dataset for irrelevance detection, a tuning framework for enhanced generalization, and the Hammer models, establishing a new standard for function-calling performance.
****
Pantelis Vafidis,Aman Bhargava,Antonio Rangel
Disentangling Representations through Multi-task Learning
Intelligent perception and interaction with the world hinges on internal representations that capture its underlying structure ("disentangled" or "abstract" representations). Disentangled representations serve as world models, isolating latent factors of variation in the world along approximately orthogonal directions, thus facilitating feature-based generalization. We provide experimental and theoretical results guaranteeing the emergence of disentangled representations in agents that optimally solve multi-task evidence accumulation classification tasks, canonical in the neuroscience literature. The key conceptual finding is that, by producing accurate multi-task classification estimates, a system implicitly represents a set of coordinates specifying a disentangled representation of the underlying latent state of the data it receives. The theory provides conditions for the emergence of these representations in terms of noise, number of tasks, and evidence accumulation time, when the classification boundaries are affine in the latent space. Surprisingly, the theory also produces closed-form expressions for extracting the disentangled representation from the model's latent state $\mathbf Z(t)$. We experimentally validate these predictions in RNNs trained on multi-task classification, which learn disentangled representations in the form of continuous attractors, leading to zero-shot out-of-distribution (OOD) generalization in predicting latent factors. We demonstrate the robustness of our framework across autoregressive architectures, decision boundary geometries and in tasks requiring classification confidence estimation. We find that transformers are particularly suited for disentangling representations, which might explain their unique world understanding abilities. Overall, our framework establishes a formal link between competence at multiple tasks and the formation of disentangled, interpretable world models in both biological and artificial systems, and helps explain why ANNs often arrive at human-interpretable concepts, and how they both may acquire exceptional zero-shot generalization capabilities.
****
Takashi Furuya,Koichi Taniguchi,Satoshi Okuda
Quantitative Approximation for Neural Operators in Nonlinear Parabolic Equations
Neural operators serve as universal approximators for general continuous operators. In this paper, we derive the approximation rate of solution operators for the nonlinear parabolic partial differential equations (PDEs), contributing to the quantitative approximation theorem for solution operators of nonlinear PDEs. Our results show that neural operators can efficiently approximate these solution operators without the exponential growth in model complexity, thus strengthening the theoretical foundation of neural operators. A key insight in our proof is to transfer PDEs into the corresponding integral equations via Duahamel's principle, and to leverage the similarity between neural operators and Picard's iterati

on—a classical algorithm for solving PDEs. This approach is potentially generali
zable beyond parabolic PDEs to a class of PDEs which can be solved by Picard's i
teration.
****
Xinyu Yang,Tianqi Chen,Beidi Chen
APE: Faster and Longer Context-Augmented Generation via Adaptive Parallel Encodi
ng
Context-augmented generation (CAG) techniques, including RAG and ICL, require th
e efficient combination of multiple contexts to generate responses to user queri
es. Directly inputting these contexts as a sequence introduces a considerable co
mputational burden by re-encoding the combined selection of contexts for every r
equest. To address this, we explore the promising potential of parallel encoding
 to independently pre-compute and cache each context's KV states. This approach
enables the direct loading of cached states during inference while accommodating
 more contexts through position reuse across contexts. However, due to misalignm
ents in attention distribution, directly applying parallel encoding results in a
 significant performance drop. To enable effective and efficient CAG, we propose
 Adaptive Parallel Encoding (**APE**), which brings shared prefix, attention tem
perature, and scaling factor to align the distribution of parallel encoding with
 sequential encoding. Results on RAG and ICL tasks demonstrate that APE can pres
erve 98\% and 93\% sequential encoding performance using the same inputs while o
utperforming parallel encoding by 3.6\% and 7.9\%, respectively. It also scales
to many-shot CAG, effectively encoding hundreds of contexts in parallel. Efficie
ncy evaluation shows that APE can achieve an end-to-end 4.5$\times$ speedup by r
educing 28$\times$ prefilling time for a 128K-length context. The code is availa
ble at
https://github.com/Infini-AI-Lab/APE.
****
Guangqi Jiang,Yifei Sun,Tao Huang,Huanyu Li,Yongyuan Liang,Huazhe Xu
Robots Pre-train Robots: Manipulation-Centric Robotic Representation from Large-
Scale Robot Datasets
The pre-training of visual representations has enhanced the efficiency of robot
learning. Due to the lack of large-scale in-domain robotic datasets, prior works
 utilize in-the-wild human videos to pre-train robotic visual representation. De
spite their promising results, representations from human videos are inevitably
subject to distribution shifts and lack the dynamics information crucial for tas
k completion. We first evaluate various pre-trained representations in terms of
their correlation to the downstream robotic manipulation tasks (i.e., manipulati
on centricity). Interestingly, we find that the ''manipulation centricity'' is a
 strong indicator of success rates when applied to downstream tasks. Drawing fro
m these findings, we propose **M**anipulation **C**entric **R**epresentation (**
MCR**), a foundation representation learning framework capturing both visual fea
tures and the dynamics information such as actions and proprioceptions of manipu
lation tasks to improve manipulation centricity. Specifically, we pre-train a vi
sual encoder on the DROID robotic dataset and leverage motion-relevant data such
 as robot proprioceptive states and actions. We introduce a novel contrastive lo
ss that aligns visual observations with the robot's proprioceptive state-action
dynamics, combined with an action prediction loss and a time contrastive loss du
ring pre-training. Empirical results across four simulation domains with 20 robo
tic manipulation tasks demonstrate that **MCR** outperforms the strongest baseli
ne by 14.8\%. Additionally, **MCR** significantly boosts the success rate in thr
ee real-world manipulation tasks by 76.9\%. Project website: robots-pretrain-rob
ots.github.io
****
Haoxi Li,Xueyang Tang,Jie ZHANG,Song Guo,Sikai Bai,Peiran Dong,Yue Yu
Causally Motivated Sycophancy Mitigation for Large Language Models
Incorporating user preferences into large language models (LLMs) can enhance the
 personalization and reliability of model outputs and facilitate the application
 of LLMs to real-world scenarios. However, leveraging user preferences can be a
double-edged sword. Recent studies have found that improper utilization can incu

r sycophancy, where LLMs prioritize alignment with user preferences over the cor
rectness of their outputs. To address sycophancy in LLMs, we analyze and model t
he problem through the lens of structured causal models (SCMs). We attribute syc
ophancy to LLMs' reliance on spurious correlations between user preferences and
model outputs in this paper. Based on the proposed SCMs, we develop a novel fram
ework, termed **CAUSM**, to mitigate sycophancy in LLMs by exploiting a signific
ant causal signature. Specifically, we eliminate the spurious correlations embed
ded in the intermediate layers of LLMs through causally motivated head reweighti
ng, and then calibrate the intra-head knowledge along the causal representation
direction. Extensive experiments are conducted across diverse language tasks to
demonstrate the superiority of our method over state-of-the-art competitors in m
itigating sycophancy in LLMs.
****
Yiran Zhao,Wenxuan Zhang,Yuxi Xie,Anirudh Goyal,Kenji Kawaguchi,Michael Shieh
Understanding and Enhancing Safety Mechanisms of LLMs via Safety-Specific Neuron
Safety alignment for large language models (LLMs) has become a critical issue du
e to their rapid progress. However, our understanding of effective safety mechan
isms in LLMs remains limited, leading to safety alignment training that mainly f
ocuses on improving optimization, data-level enhancement, or adding extra struct
ures to intentionally block harmful outputs. To address this gap, we develop a n
euron detection method to identify safety neurons—those consistently crucial for
 handling and defending against harmful queries. Our findings reveal that these
safety neurons constitute less than 1% of all parameters, are language-specif
ic and are predominantly located in self-attention layers. Moreover, safety is c
ollectively managed by these neurons in the first several layers. Based on these
 observations, we introduce a $\underline{S}$afety $\underline{N}$euron $\underl
ine{Tun}$ing method, named $\texttt{SN-Tune}$, that exclusively tune safety neur
ons without compromising models' general capabilities. $\texttt{SN-Tune}$ signif
icantly enhances the safety of instruction-tuned models, notably reducing the ha
rmful scores of Llama3-8B-Instruction from 65.5 to 2.0, Mistral-7B-Instruct-
v0.2 from 70.8 to 4.5, and Vicuna-13B-1.5 from 93.5 to 3.0. Moreover, $\
texttt{SN-Tune}$ can be applied to base models on efficiently establishing LLMs'
 safety mechanism. In addition, we propose $\underline{R}$obust $\underline{S}$a
fety $\underline{N}$euron $\underline{Tun}$ing method ($\texttt{RSN-Tune}$), whi
ch preserves the integrity of LLMs' safety mechanisms during downstream task fin
e-tuning by separating the safety neurons from models' foundation neurons.
****
Mingxun Zhou,Elaine Shi,Giulia Fanti
Pacmann: Efficient Private Approximate Nearest Neighbor Search
We propose a new private Approximate Nearest Neighbor (ANN) search scheme
named Pacmann
that allows a client to perform ANN search
in a vector database
without revealing the query vector to the server.
Unlike prior constructions that run encrypted search on the server side,
Pacmann carefully offloads limited computation and storage to the client,
no longer requiring computationally-intensive cryptographic techniques.
Specifically, clients run a graph-based ANN search, where in each hop on the gra
ph, the client privately retrieves local graph information from the server.
To make this efficient, we combine two ideas:
(1) we adapt a leading graph-based ANN search algorithm to be compatible with pr
ivate information retrieval (PIR) for subgraph retrieval;
(2) we use a recent class of PIR schemes that trade offline preprocessing for on
line computational efficiency.
Pacmann achieves significantly better search quality than
the state-of-the-art private ANN search schemes,
showing up to 2.5$\times$ better search accuracy on
real-world datasets than prior work and
reaching 90\% quality of a state-of-the-art
non-private ANN algorithm.

Moreover on large datasets with up to 100 million vectors,
Pacmann shows better scalability
than prior private ANN schemes
with up to 62\% reduction in computation time
and 22\% reduction in overall latency.
****
Xingxuan Zhang,Haoran Wang,Jiansheng Li,Yuan Xue,Shikai Guan,Renzhe Xu,Hao Zou,Han Yu,Peng Cui
Understanding the Generalization of In-Context Learning in Transformers: An Empirical Study
Large language models (LLMs) like GPT-4 and LLaMA-3 utilize the powerful in-context learning (ICL) capability of Transformer architecture to learn on the fly from limited examples. While ICL underpins many LLM applications, its full potential remains hindered by a limited understanding of its generalization boundaries and vulnerabilities. We present a systematic investigation of transformers' generalization capability with ICL relative to training data coverage by defining a task-centric framework along three dimensions: inter-problem, intra-problem, and intra-task generalization. Through extensive simulation and real-world experiments, encompassing tasks such as function fitting, API calling, and translation, we find that transformers lack inter-problem generalization with ICL, but excel in intra-task and intra-problem generalization. When the training data includes a greater variety of mixed tasks, it significantly enhances the generalization ability of ICL on unseen tasks and even on known simple tasks. This guides us in designing training data to maximize the diversity of tasks covered and to combine different tasks whenever possible, rather than solely focusing on the target task for testing.
****
Andrew Liu,Henry Prior,Gargi Balasubramaniam,Rivka Moroshko,Amir Zait,Ilia Labzovsky,Danny Karmon,Ishita Dasgupta,Kim Stachenfeld,Kenneth Marino
ReCogLab: a framework testing relational reasoning & cognitive hypotheses on LLMs
A fundamental part of human cognition is the ability to not only recall previous memories, but also reason across them to draw conclusions. In cognitive science and psychology, this is termed relational reasoning and a number of effects and biases have been observed in human cognition. Designing experiments to measure these reasoning effects is effortful and does not transfer easily to analyzing language model reasoning patterns. To make exploring language models on relational reasoning easier, we introduce ReCogLab – a generative framework for constructing reasoning examples. Unlike static datasets, our framework has a number of benefits that help us in our goal of flexible evaluation of LLMs. First, our framework allows us to control the difficulty and context-length of the problem, allowing us to scale with model capability and evaluate LLMs at a variety of scales. Second, the ability to change the configuration of a dataset dynamically allows us to probe models on different aspects and capabilities. Finally, the flexibility of our approach enables the recreation of classic cognitive science experiments and the systematic study of relational reasoning biases in language models. We demonstrate several such experiments and present our findings on a wide variety of open and closed-source language models. We release all data and code at https://github.com/google-deepmind/recoglab.
****
Jingwei Xu,Junyu Lai,Yunpeng Huang
MeteoRA: Multiple-tasks Embedded LoRA for Large Language Models
The pretrain+fine-tune paradigm is foundational for deploying large language models (LLMs) across various downstream applications. Within this framework, Low-Rank Adaptation (LoRA) stands out for its parameter-efficient fine-tuning (PEFT), producing numerous reusable task-specific LoRA adapters. However, this approach requires explicit task intention selection, posing challenges for autonomous task sensing and switching during inference with multiple existing LoRA adapters embedded in a single LLM. In this work, we introduce MeteoRA (Multiple-Tasks embedded LoRA), a scalable and efficient framework that reuses multiple task-specific

LoRA adapters into the base LLM via a full-mode Mixture-of-Experts (MoE) architecture. This framework also includes novel MoE forward acceleration strategies to address the efficiency challenges of traditional MoE implementations. Our evaluation, using the LlaMA2-13B and LlaMA3-8B base models equipped with 28 existing LoRA adapters through MeteoRA, demonstrates equivalent performance with the traditional PEFT method. Moreover, the LLM equipped with MeteoRA achieves superior performance in handling composite tasks, effectively solving ten sequential problems in a single inference pass, thereby demonstrating the framework's enhanced capability for timely adapter switching.
****
Ayano Hiranaka,Shang-Fu Chen,Chieh-Hsin Lai,Dongjun Kim,Naoki Murata,Takashi Shibuya,Wei-Hsiang Liao,Shao-Hua Sun,Yuki Mitsufuji
HERO: Human-Feedback Efficient Reinforcement Learning for Online Diffusion Model Finetuning
Controllable generation through Stable Diffusion (SD) fine-tuning aims to improve fidelity, safety, and alignment with human guidance. Existing reinforcement learning from human feedback methods usually rely on predefined heuristic reward functions or pretrained reward models built on large-scale datasets, limiting their applicability to scenarios where collecting such data is costly or difficult. To effectively and efficiently utilize human feedback, we develop a framework, HERO, which leverages online human feedback collected on the fly during model learning. Specifically, HERO features two key mechanisms: (1) Feedback-Aligned Representation Learning, an online training method that captures human feedback and provides informative learning signals for fine-tuning, and (2) Feedback-Guided Image Generation, which involves generating images from SD's refined initialization samples, enabling faster convergence towards the evaluator's intent. We demonstrate that HERO is 4x more efficient in online feedback for body part anomaly correction compared to the best existing method. Additionally, experiments show that HERO can effectively handle tasks like reasoning, counting, personalization, and reducing NSFW content with only 0.5K online feedback. The code and project page are available at https://hero-dm.github.io/.
****
Chengyuan Deng,Jie Gao,Jalaj Upadhyay,Chen Wang,Samson Zhou
On the Price of Differential Privacy for Hierarchical Clustering
Hierarchical clustering is a fundamental unsupervised machine learning task with the aim of organizing data into a hierarchy of clusters. Many applications of hierarchical clustering involve sensitive user information, therefore motivating recent studies on differentially private hierarchical clustering under the rigorous framework of Dasgupta's objective. However, it has been shown that any privacy-preserving algorithm under edge-level differential privacy necessarily suffers a large error. To capture practical applications of this problem, we focus on the weight privacy model, where each edge of the input graph is at least unit weight. We present a novel algorithm in the weight privacy model that shows significantly better approximation than known impossibility results in the edge-level DP setting. In particular, our algorithm achieves $O(\log^{1.5}n/\varepsilon)$ multiplicative error for ε-DP and runs in polynomial time, where n is the size of the input graph, and the cost is never worse than the optimal additive error in existing work. We complement our algorithm by showing if the unit-weight constraint does not apply, the lower bound for weight-level DP hierarchical clustering is essentially the same as the edge-level DP, i.e. $\Omega(n^2/\varepsilon)$ additive error. As a result, we also obtain a new lower bound of $\tilde{\Omega}(1/\varepsilon)$ additive error for balanced sparsest cuts in the weight-level DP model, which may be of independent interest. Finally, we evaluate our algorithm on synthetic and real-world datasets. Our experimental results show that our algorithm performs well in terms of extra cost and has good scalability to large graphs.
****
Andrew Luo,Jacob Yeung,Rushikesh Zawar,Shaurya Rajat Dewan,Margaret Marie Henderson,Leila Wehbe,Michael J. Tarr
Brain Mapping with Dense Features: Grounding Cortical Semantic Selectivity in Na

tural Images With Vision Transformers
We introduce BrainSAIL (Semantic Attribution and Image Localization), a method f
or linking neural selectivity with spatially distributed semantic visual concept
s in natural scenes. BrainSAIL leverages recent advances in large-scale artifici
al neural networks, using them to provide insights into the functional topology
of the brain. To overcome the challenge presented by the co-occurrence of multip
le categories in natural images, BrainSAIL exploits semantically consistent, den
se spatial features from pre-trained vision models, building upon their demonstr
ated ability to robustly predict neural activity. This method derives clean, spa
tially dense embeddings without requiring any additional training, and employs a
 novel denoising process that leverages the semantic consistency of images under
 random augmentations. By unifying the space of whole-image embeddings and dense
 visual features and then applying voxel-wise encoding models to these features,
 we enable the identification of specific subregions of each image which drive s
electivity patterns in different areas of the higher visual cortex. This provide
s a powerful tool for dissecting the neural mechanisms that underlie semantic vi
sual processing for natural images. We validate BrainSAIL on cortical regions wi
th known category selectivity, demonstrating its ability to accurately localize
and disentangle selectivity to diverse visual concepts. Next, we demonstrate Bra
inSAIL's ability to characterize high-level visual selectivity to scene properti
es and low-level visual features such as depth, luminance, and saturation, provi
ding insights into the encoding of complex visual information. Finally, we use B
rainSAIL to directly compare the feature selectivity of different brain encoding
 models across different regions of interest in visual cortex. Our innovative me
thod paves the way for significant advances in mapping and decomposing high-leve
l visual representations in the human brain.
****
Naveen Gupta,Medha Sawhney,Arka Daw,Youzuo Lin,Anuj Karpatne
A Unified Framework for Forward and Inverse Problems in Subsurface Imaging using
 Latent Space Translations
In subsurface imaging, learning the mapping from velocity maps to seismic wavefo
rms (forward problem) and waveforms to velocity (inverse problem) is important f
or several applications. While traditional techniques for solving forward and in
verse problems are computationally prohibitive, there is a growing interest to l
everage recent advances in deep learning to learn the mapping between velocity m
aps and seismic waveform images directly from data. Despite the variety of archi
tectures explored in previous works, several open questions still remain unanswe
red such as the effect of latent space sizes, the importance of manifold learnin
g, the complexity of translation models, and the value of jointly solving forwar
d and inverse problems. We propose a unified framework to systematically charact
erize prior research in this area termed the Generalized Forward-Inverse (GFI) f
ramework, building on the assumption of manifolds and latent space translations.
 We show that GFI encompasses previous works in deep learning for subsurface ima
ging, which can be viewed as specific instantiations of GFI. We also propose two
 new model architectures within the framework of GFI: Latent U-Net and Invertibl
e X-Net, leveraging the power of U-Nets for domain translation and the ability o
f IU-Nets to simultaneously learn forward and inverse translations, respectively
. We show that our proposed models achieve state-of-the-art (SOTA) performance f
or forward and inverse problems on a wide range of synthetic datasets, and also
investigate their zero-shot effectiveness on two real-world-like datasets. The c
ode is available at https://github.com/KGML-lab/Generalized-Forward-Inverse-Fram
ework-for-DL4SI
****
Akash Kumar,Zsolt Kira,Yogesh S Rawat
Contextual Self-paced Learning for Weakly Supervised Spatio-Temporal Video Groun
ding
In this work, we focus on Weakly Supervised Spatio-Temporal Video Grounding (WST
VG). It is a multimodal task aimed at localizing specific subjects spatio-tempo
rally based on textual queries without bounding box supervision. Motivated by re
cent advancements in multi-modal foundation models for grounding tasks, we first

explore the potential of state-of-the-art object detection models for WSTVG. Despite their robust zero-shot capabilities, our adaptation reveals significant limitations, including inconsistent temporal predictions, inadequate understanding of complex queries, and challenges in adapting to difficult scenarios. We propose CoSPaL (Contextual Self-Paced Learning), a novel approach which is designed to overcome these limitations. CoSPaL integrates three core components: (1) Tubelet Phrase Grounding (TPG), which introduces spatio-temporal prediction by linking textual queries to tubelets; (2) Contextual Referral Grounding (CRG), which improves comprehension of complex queries by extracting contextual information to refine object identification over time; and (3) Self-Paced Scene Understanding (SPS), a training paradigm that progressively increases task difficulty, enabling the model to adapt to complex scenarios by transitioning from coarse to fine-grained understanding.
****
Yue Zhao,Yuanjun Xiong,Philipp Kraehenbuehl
Image and Video Tokenization with Binary Spherical Quantization
We propose a new transformer-based image and video tokenizer with Binary Spherical Quantization (BSQ). BSQ projects the high-dimensional visual embedding to a lower-dimensional hypersphere and then applies binary quantization. BSQ is (1) parameter-efficient without an explicit codebook, (2) scalable to arbitrary token dimensions, and (3) compact: compressing visual data by up to 100×
with minimal distortion. Our tokenizer uses a transformer encoder and decoder with simple block-wise causal masking to support variable-length videos as input. The resulting BSQ-ViT achieves state-of-the-art visual reconstruction quality on image and video reconstruction benchmarks with 2.4× throughput compared to the best prior methods. Furthermore, by learning an autoregressive prior for adaptive arithmetic coding, BSQ-ViT achieves comparable visual compression results with commonly used compression standards, e.g. JPEG2000/WebP for images and H.264/H.265 for videos. BSQ-ViT also enables masked language models to achieve competitive image synthesis quality to GAN and diffusion approaches.
****
Pritam Sarkar,Sayna Ebrahimi,Ali Etemad,Ahmad Beirami,Sercan O Arik,Tomas Pfister
Mitigating Object Hallucination in MLLMs via Data-augmented Phrase-level Alignment
Despite their significant advancements, Multimodal Large Language Models
(MLLMs) often generate factually inaccurate information, referred to as hallucination.
In this work, we address object hallucinations in MLLMs, where information
is generated about an object not present in the input image. We introduce Data-augmented
Phrase-level Alignment (DPA), a novel loss which can be applied to
instruction-tuned off-the-shelf MLLMs to mitigate hallucinations, while preserving
their general vision-language capabilities. To fine-tune MLLMs with DPA, we first
generate a set of 'hallucinated' and 'correct' response pairs through generative data
augmentation by selectively altering the ground-truth information of the correct
responses at a phrase level. The DPA loss is then used to train MLLMs to reduce
the likelihood of hallucinated phrases compared to the correct ones. Our thorough
evaluation on various benchmarks confirms the effectiveness of DPA in mitigating
hallucination while retaining the out-of-the-box performance of the MLLMs on
general tasks. For instance, MLLMs finetuned with DPA, which we refer to as Hallucination
Attenuated Language and Vision Assistant (HALVA), improve F1 by up
to 13.4% on hallucination visual question-answering and reduce the hallucination
rate by up to 4.2% on image description tasks.
****

Jan Robine,Marc Höftmann,Stefan Harmeling
Simple, Good, Fast: Self-Supervised World Models Free of Baggage
What are the essential components of world models? How far do we get with world models that are not employing RNNs, transformers, discrete representations, and image reconstructions? This paper introduces SGF, a Simple, Good, and Fast world model that uses self-supervised representation learning, captures short-time dependencies through frame and action stacking, and enhances robustness against model errors through data augmentation. We extensively discuss SGF’s connections to established world models, evaluate the building blocks in ablation studies, and demonstrate good performance through quantitative comparisons on the Atari 100k benchmark. The code is available at https://github.com/jrobine/sgf.
****
Trung X. Pham,Tri Ton,Chang D. Yoo
MDSGen: Fast and Efficient Masked Diffusion Temporal-Aware Transformers for Open-Domain Sound Generation
We introduce MDSGen, a novel framework for vision-guided open-domain sound generation optimized for model parameter size, memory consumption, and inference speed. This framework incorporates two key innovations: (1) a redundant video feature removal module that filters out unnecessary visual information, and (2) a temporal-aware masking strategy that leverages temporal context for enhanced audio generation accuracy. In contrast to existing resource-heavy Unet-based models, MDSGen employs denoising masked diffusion transformers, facilitating efficient generation without reliance on pre-trained diffusion models. Evaluated on the benchmark VGGSound dataset, our smallest model (5M parameters) achieves 97.9% alignment accuracy, using 172x fewer parameters, 371% less memory, and offering 36x faster inference than the current 860M-parameter state-of-the-art model (93.9% accuracy). The larger model (131M parameters) reaches nearly 99% accuracy while requiring 6.5x fewer parameters. These results highlight the scalability and effectiveness of our approach. The code is available at https://bit.ly/mdsgen.
****
Wenzheng Pan,Hao Xiong,Jiale Ma,Wentao Zhao,Yang Li,Junchi Yan
UniCO: On Unified Combinatorial Optimization via Problem Reduction to Matrix-Encoded General TSP
Various neural solvers have been devised for combinatorial optimization (CO), which are often tailored for specific problem types, e.g., TSP, CVRP and SAT, etc. Yet, it remains an open question how to achieve universality regarding problem representing and learning with a general framework. This paper first proposes **UniCO**, to unify a set of CO problems by reducing them into the *general* TSP form featured by distance matrices. The applicability of this strategy depends on the efficiency of the problem reduction and solution transition procedures, which we show that at least ATSP, HCP, and SAT are readily feasible. The hope is to allow for the effective and even simultaneous use of as many types of CO instances as possible to train a neural TSP solver, and optionally finetune it for specific problem types. In particular, unlike the prevalent TSP benchmarks based on Euclidean instances with 2-D coordinates, our studied domain of TSP could involve non-metric, asymmetric or discrete distances without explicit node coordinates, which is much less explored in TSP literature while poses new intellectual challenges. Along this direction, we devise two neural TSP solvers with and without supervision to conquer such matrix-formulated input, respectively: 1) **MatPOENet** and 2) **MatDIFFNet**. The former is a reinforcement learning-based sequential model with pseudo one-hot embedding (POE) scheme; and the latter is a Diffusion-based generative model with the mix-noised reference mapping scheme. Experiments on ATSP, 2DTSP, HCP- and SAT-distributed general TSPs show the strong ability towards arbitrary matrix-encoded TSP with structure and size variation.
****
Jason Piquenot,Maxime Berar,Romain Raveaux,Pierre Héroux,Jean-Yves RAMEL,Sébastien Adam
Grammar Reinforcement Learning: path and cycle counting in graphs with a Context-Free Grammar and Transformer approach
This paper presents Grammar Reinforcement Learning (GRL), a reinforcement learni

ng algorithm that uses Monte Carlo Tree Search (MCTS) and a transformer architec
ture that models a Pushdown Automaton (PDA) within a context-free grammar (CFG)
framework. Taking as use case the problem of efficiently counting paths and cycl
es in graphs, a key challenge in network analysis, computer science, biology, an
d social sciences, GRL discovers new matrix-based formulas for path/cycle counti
ng that improve computational efficiency by factors of two to six w.r.t state-of
-the-art approaches. Our contributions include: (i) a framework for generating t
ransformers that operate within a CFG, (ii) the development of GRL for optimizin
g formulas within grammatical structures, and (iii) the discovery of novel formu
las for graph substructure counting, leading to significant computational improv
ements.
****
Shuhan Song,Ping Li,Ming Dun,Maolei Huang,Huawei Cao,Xiaochun Ye
GPromptShield: Elevating Resilience in Graph Prompt Tuning Against Adversarial A
ttacks
The paradigm of ``pre-training and prompt-tuning", with its effectiveness and li
ghtweight characteristics, has rapidly spread from the language field to the gra
ph field. Several pioneering studies have designed specialized prompt functions
for diverse downstream graph tasks based on various graph pre-training strategie
s. These prompts concentrate on the compatibility between the pre-training prete
xt and downstream graph tasks, aiming to bridge the gap between them. However, d
esigning prompts to blindly adapt to downstream tasks based on this concept negl
ects crucial security issues. By conducting covert attacks on downstream graph d
ata, we find that even when the downstream task data closely matches that of the
 pre-training tasks, it is still feasible to generate highly misleading prompts
using simple deceptive techniques. In this paper, we shift the primary focus of
graph prompts from compatibility to vulnerability issues in adversarial attack s
cenarios. We design a highly extensible shield defense system for the prompts, w
hich enhances their robustness from two perspectives: \textbf{\textit{Direct Han
dling}} and \textbf{\textit{Indirect Amplification}}. When downstream graph data
 exhibits unreliable biases, the former directly combats invalid information by
adding hybrid multi-defense prompts to the input graph's feature space, while th
e latter employs a training strategy that circumvents invalid part and amplifies
 valid part. We provide a theoretical derivation that proves their feasibility,
indicating that unbiased prompts exist under certain conditions on unreliable da
ta. Extensive experiments across various scenarios of adversarial attack (includ
ing adaptive and non-adaptive attacks) indicate that the prompts within our shie
ld defense system exhibit enhanced resilience and superiority. Our work explores
 new perspectives in the field of graph prompts, offering a novel option for dow
nstream robust prompt tuning.
****
Shengpeng Ji,Ziyue Jiang,Wen Wang,Yifu Chen,Minghui Fang,Jialong Zuo,Qian Yang,X
ize Cheng,Zehan Wang,Ruiqi Li,Ziang Zhang,Xiaoda Yang,Rongjie Huang,Yidi Jiang,Q
ian Chen,Siqi Zheng,Zhou Zhao
WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language
Modeling
Language models have been effectively applied to modeling natural signals, such
as images, video, speech, and audio. A crucial component of these models is the
codec tokenizer, which compresses high-dimensional natural signals into lower-di
mensional discrete tokens. In this paper, we introduce WavTokenizer, which offer
s several advantages over previous SOTA acoustic codec models in the audio domai
n: 1) extreme compression. By compressing the layers of quantizers and the tempo
ral dimension of the discrete codec, one-second audio of 24kHz sampling rate req
uires only a single quantizer with 40 or 75 tokens. 2) improved subjective quali
ty. Despite the reduced number of tokens, WavTokenizer achieves state-of-the-art
 reconstruction quality with outstanding UTMOS scores and inherently contains ri
cher semantic information. Specifically, we achieve these results by designing a
 broader VQ space, extended contextual windows, and improved attention networks,
 as well as introducing a powerful multi-scale discriminator and an inverse Four
ier transform structure. We conducted extensive reconstruction experiments in th

e domains of speech, audio, and music. WavTokenizer exhibited strong performance across various objective and subjective metrics compared to state-of-the-art models. We also tested semantic information, VQ utilization, and adaptability to generative models. Comprehensive ablation studies confirm the necessity of each module in WavTokenizer. The code is available at https://github.com/jishengpeng/WavTokenizer.
****
Songming Liu,Lingxuan Wu,Bangguo Li,Hengkai Tan,Huayu Chen,Zhengyi Wang,Ke Xu,Hang Su,Jun Zhu
RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation
Bimanual manipulation is essential in robotics, yet developing foundation models is extremely challenging due to the inherent complexity of coordinating two robot arms (leading to multi-modal action distributions) and the scarcity of training data. In this paper, we present the Robotics Diffusion Transformer (RDT), a pioneering diffusion foundation model for bimanual manipulation. RDT builds on diffusion models to effectively represent multi-modality, with innovative designs of a scalable Transformer to deal with the heterogeneity of multi-modal inputs and to capture the nonlinearity and high frequency of robotic data. To address data scarcity, we further introduce a Physically Interpretable Unified Action Space, which can unify the action representations of various robots while preserving the physical meanings of original actions, facilitating learning transferrable physical knowledge. With these designs, we managed to pre-train RDT on the largest collection of multi-robot datasets to date and scaled it up to 1.2B parameters, which is the largest diffusion-based foundation model for robotic manipulation. We finally fine-tuned RDT on a self-created multi-task bimanual dataset with over 6K+ episodes to refine its manipulation capabilities. Experiments on real robots demonstrate that RDT significantly outperforms existing methods. It exhibits zero-shot generalization to unseen objects and scenes, understands and follows language instructions, learns new skills with just 1$\sim$5 demonstrations, and effectively handles complex, dexterous tasks. We refer to https://rdt-robotics.github.io/rdt-robotics/ for the code and videos.
****
Noah Amsel,Gilad Yehudai,Joan Bruna
Quality over Quantity in Attention Layers: When Adding More Heads Hurts
Attention-based mechanisms are widely used in machine learning, most prominently in transformers. However, hyperparameters such as the number of attention heads and the attention rank (i.e., the query/key dimension) are set nearly the same way in all realizations of this architecture, without theoretical justification. In this paper, we prove that the rank can have a dramatic effect on the representational capacity of attention. This effect persists even when the number of heads and the parameter count are very large. Specifically, we present a simple and natural target function based on nearest neighbor search that can be represented using a single full-rank attention head for any sequence length, but that cannot be approximated by a low-rank attention layer even on short sequences unless the number of heads is exponential in the embedding dimension. Thus, for this target function, rank is what determines an attention layer's power. We show that, for short sequences, using multiple layers allows the target to be approximated by low-rank attention; for long sequences, we conjecture that full-rank attention is necessary regardless of depth. Finally, we present experiments with standard multilayer transformers that validate our theoretical findings. They demonstrate that, because of how standard transformer implementations set the rank, increasing the number of attention heads can severely decrease accuracy on certain tasks.
****
John Gkountouras,Matthias Lindemann,Phillip Lippe,Efstratios Gavves,Ivan Titov
Language Agents Meet Causality -- Bridging LLMs and Causal World Models
Large Language Models (LLMs) have recently shown great promise in planning and reasoning applications. These tasks demand robust systems, which arguably require a causal understanding of the environment. While LLMs can acquire and reflect common sense causal knowledge from their pretraining data, this information is of

ten incomplete, incorrect, or inapplicable to a specific environment. In contras
t, causal representation learning (CRL) focuses on identifying the underlying ca
usal structure within a given environment. We propose a framework that integrate
s CRLs with LLMs to enable causally-aware reasoning and planning. This framework
 learns a causal world model, with causal variables linked to natural language e
xpressions. This mapping provides LLMs with a flexible interface to process and
generate descriptions of actions and states in text form. Effectively, the causa
l world model acts as a simulator that the LLM can query and interact with. We e
valuate the framework on causal inference and planning tasks across temporal sca
les and environmental complexities. Our experiments demonstrate the effectivenes
s of the approach, with the causally-aware method outperforming LLM-based reason
ers, especially for longer planning horizons.
****

Qiqi Hou,Randall Rauwendaal,Zifeng Li,Hoang Le,Farzad Farhadzadeh,Fatih Porikli,
Alexei Bourd,Amir Said
Sort-free Gaussian Splatting via Weighted Sum Rendering
Recently, 3D Gaussian Splatting (3DGS) has emerged as a significant advancement
in 3D scene reconstruction, attracting considerable attention due to its ability
 to recover high-fidelity details while maintaining low complexity. Despite the
promising results achieved by 3DGS, its rendering performance is constrained by
its dependence on costly non-commutative alpha-blending operations. These operat
ions mandate complex view dependent sorting operations that introduce computatio
nal overhead, especially on the resource-constrained platforms such as mobile ph
ones. In this paper, we propose Weighted Sum Rendering, which approximates alpha
 blending with weighted sums, thereby removing the need for sorting. This simpli
fies implementation, delivers superior performance, and eliminates the ``popping
'' artifacts caused by sorting. Experimental results show that optimizing a gene
ralized Gaussian splatting formulation to the new differentiable rendering yield
s competitive image quality. The method was implemented and tested in a mobile d
evice GPU, achieving on average $1.23\times$ faster rendering.
****

Gengshan Yang,Andrea Bajcsy,Shunsuke Saito,Angjoo Kanazawa
Agent-to-Sim: Learning Interactive Behavior Models from Casual Longitudinal Vide
os
We present Agent-to-Sim (ATS), a framework for learning interactive behavior mod
els of 3D agents from casual longitudinal video collections. Different from prio
r works that rely on marker-based tracking and multiview cameras, ATS learns nat
ural behaviors of animal agents non-invasively through video observations record
ed over a long time-span (e.g. a month) in a single environment.
Modeling 3D behavior of an agent requires persistent 3D tracking (e.g., knowing
which point corresponds to which) over a long time period. To obtain such data,
we develop a coarse-to-fine registration method that tracks the agent and the ca
mera over time through a canonical 3D space, resulting in a complete and persist
ent spacetime 4D representation. We then train a generative model of agent behav
iors using paired data of perception and motion of an agent queried from the 4D
reconstruction. ATS enables real-to-sim transfer from video recordings of an age
nt to an interactive behavior simulator. We demonstrate results on animals given
 monocular RGBD videos captured by a smartphone. Project page: gengshan-y.github
.io/agent2sim-www.
****

Danny Wang,Ruihong Qiu,Guangdong Bai,Zi Huang
GOLD: Graph Out-of-Distribution Detection via Implicit Adversarial Latent Genera
tion
Despite graph neural networks' (GNNs) great success in modelling graph-structure
d data, out-of-distribution (OOD) test instances still pose a great challenge fo
r current GNNs. One of the most effective techniques to detect OOD nodes is to e
xpose the detector model with an additional OOD node-set, yet the extra OOD inst
ances are often difficult to obtain in practice. Recent methods for image data a
ddress this problem using OOD data synthesis, typically relying on pre-trained g
enerative models like Stable Diffusion. However, these approaches require vast a

mounts of additional data, as well as one-for-all pre-trained generative models, which are not available for graph data. Therefore, we propose the GOLD framework for graph OOD detection, an implicit adversarial learning pipeline with synthetic OOD exposure without pre-trained models. The implicit adversarial training process employs a novel alternating optimisation framework by training: (1) a latent generative model to regularly imitate the in-distribution (ID) embeddings from an evolving GNN, and (2) a GNN encoder and an OOD detector to accurately classify ID data while increasing the energy divergence between the ID embeddings and the generative model's synthetic embeddings. This novel approach implicitly transforms the synthetic embeddings into pseudo-OOD instances relative to the ID data, effectively simulating exposure to OOD scenarios without auxiliary data. Extensive OOD detection experiments are conducted on five benchmark graph datasets, verifying the superior performance of GOLD without using real OOD data compared with the state-of-the-art OOD exposure and non-exposure baselines.
****

Namgyu Kang,Jaemin Oh,Youngjoon Hong,Eunbyung Park
PIG: Physics-Informed Gaussians as Adaptive Parametric Mesh Representations
The numerical approximation of partial differential equations (PDEs) using neural networks has seen significant advancements through Physics-Informed Neural Networks (PINNs). Despite their straightforward optimization framework and flexibility in implementing various PDEs, PINNs often suffer from limited accuracy due to the spectral bias of Multi-Layer Perceptrons (MLPs), which struggle to effectively learn high-frequency and nonlinear components. Recently, parametric mesh representations in combination with neural networks have been investigated as a promising approach to eliminate the inductive bias of MLPs. However, they usually require high-resolution grids and a large number of collocation points to achieve high accuracy while avoiding overfitting. In addition, the fixed positions of the mesh parameters restrict their flexibility, making accurate approximation of complex PDEs challenging. To overcome these limitations, we propose Physics-Informed Gaussians (PIGs), which combine feature embeddings using Gaussian functions with a lightweight neural network. Our approach uses trainable parameters for the mean and variance of each Gaussian, allowing for dynamic adjustment of their positions and shapes during training. This adaptability enables our model to optimally approximate PDE solutions, unlike models with fixed parameter positions. Furthermore, the proposed approach maintains the same optimization framework used in PINNs, allowing us to benefit from their excellent properties. Experimental results show the competitive performance of our model across various PDEs, demonstrating its potential as a robust tool for solving complex PDEs. Our project page is available at https://namgyukang.github.io/Physics-Informed-Gaussians/
****

Tariq Berrada,Pietro Astolfi,Melissa Hall,Marton Havasi,Yohann Benchetrit,Adriana Romero-Soriano,Karteek Alahari,Michal Drozdzal,Jakob Verbeek
Boosting Latent Diffusion with Perceptual Objectives
Latent diffusion models (LDMs) power state-of-the-art high-resolution generative image models. LDMs learn the data distribution in the latent space of an autoencoder (AE) and produce images by mapping the generated latents into RGB image space using the AE decoder. While this approach allows for efficient model training and sampling, it induces a disconnect between the training of the diffusion model and the decoder, resulting in a loss of detail in the generated images. To remediate this disconnect, we propose to leverage the internal features of the decoder to define a latent perceptual loss (LPL). This loss encourages the models to create sharper and more realistic images. Our loss can be seamlessly integrated with common autoencoders used in latent diffusion models, and can be applied to different generative modeling paradigms such as DDPM with epsilon and velocity prediction, as well as flow matching. Extensive experiments with models trained on three datasets at 256 and 512 resolution show improved quantitative -- with boosts between 6% and 20% in FID -- and qualitative results when using our perceptual loss.
****

Seanie Lee,Haebin Seong,Dong Bok Lee,Minki Kang,Xiaoyin Chen,Dominik Wagner,Yosh

ua Bengio,Juho Lee,Sung Ju Hwang
HarmAug: Effective Data Augmentation for Knowledge Distillation of Safety Guard Models
Safety guard models that detect malicious queries aimed at large language models (LLMs) are essential for ensuring the secure and responsible deployment of LLMs in real-world applications.
However, deploying existing safety guard models with billions of parameters alongside LLMs on mobile devices is impractical due to substantial memory requirements and latency.
To reduce this cost, we distill a large teacher safety guard model into a smaller one using a labeled dataset of instruction-response pairs with binary harmfulness labels. Due to the limited diversity of harmful instructions in the existing labeled dataset, naively distilled models tend to underperform compared to larger models. To bridge the gap between small and large models, we propose **HarmAug**, a simple yet effective data augmentation method that involves jailbreaking an LLM and prompting it to generate harmful instructions. Given a prompt such as, "Make a single harmful instruction prompt that would elicit offensive content", we add an affirmative prefix (e.g., "I have an idea for a prompt:") to the LLM's response. This encourages the LLM to continue generating the rest of the response, leading to sampling harmful instructions. Another LLM generates a response to the harmful instruction, and the teacher model labels the instruction-response pair. We empirically show that our HarmAug outperforms other relevant baselines. Moreover, a 435-million-parameter safety guard model trained with HarmAug achieves an F1 score comparable to larger models with over 7 billion parameters, and even outperforms them in AUPRC, while operating at less than 25\% of their computational cost. Our [code](https://anonymous.4open.science/r/HarmAug/), [safety guard model](https://huggingface.co/AnonHB/HarmAug_Guard_Model_deberta_v3_large_finetuned), and [synthetic dataset](https://huggingface.co/datasets/AnonHB/HarmAug_generated_dataset) are publicly available.
****
Yilong Li,Jingyu Liu,Hao Zhang,M Badri Narayanan,Utkarsh Sharma,Shuai Zhang,Yijing Zeng,Jayaram Raghuram,Suman Banerjee
PALMBENCH: A COMPREHENSIVE BENCHMARK OF COMPRESSED LARGE LANGUAGE MODELS ON MOBILE PLATFORMS
Deploying large language models (LLMs) locally on mobile devices is advantageous in scenarios where transmitting data to remote cloud servers is either undesirable due to privacy concerns or impractical due to network connection. Recent advancements have facilitated the local deployment of LLMs. However, local deployment also presents challenges, particularly in balancing quality (generative performance), latency, and throughput within the hardware constraints of mobile devices. In this paper, we introduce our lightweight, all-in-one automated benchmarking framework that allows users to evaluate LLMs on mobile devices. We provide a comprehensive benchmark of various popular LLMs with different quantization configurations (both weights and activations) across multiple mobile platforms with varying hardware capabilities. Unlike traditional benchmarks that assess full-scale models on high-end GPU clusters, we focus on evaluating resource efficiency (memory and power consumption) and harmful output for compressed models on mobile devices. Our key observations include: i) differences in energy efficiency and throughput across mobile platforms; ii) the impact of quantization on memory usage, GPU execution time, and power consumption; and iii) accuracy and performance degradation of quantized models compared to their non-quantized counterparts; and iv) the frequency of hallucinations and toxic content generated by compressed LLMs on
mobile devices.
****
Tudor Ioan Cebere,Aurélien Bellet,Nicolas Papernot
Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model
Machine learning models can be trained with formal privacy guarantees via differentially private optimizers such as DP-SGD. In this work, we focus on a threat model where the adversary has access only to the final model, with no visibility

into intermediate updates. In the literature, this ``hidden state'' threat model exhibits a significant gap between the lower bound from empirical privacy auditing and the theoretical upper bound provided by privacy accounting. To challenge this gap, we propose to audit this threat model with adversaries that craft a gradient sequence designed to maximize the privacy loss of the final model without relying on intermediate updates. Our experiments show that this approach consistently outperforms previous attempts at auditing the hidden state model. Furthermore, our results advance the understanding of achievable privacy guarantees within this threat model. Specifically, when the crafted gradient is inserted at every optimization step, we show that concealing the intermediate model updates in DP-SGD does not enhance the privacy guarantees. The situation is more complex when the crafted gradient is not inserted at every step: our auditing lower bound matches the privacy upper bound only for an adversarially-chosen loss landscape and a sufficiently large batch size. This suggests that existing privacy upper bounds can be improved in certain regimes.
****

Huaisheng Zhu,Teng Xiao,Vasant G Honavar
DSPO: Direct Score Preference Optimization for Diffusion Model Alignment
Diffusion-based Text-to-Image (T2I) models have achieved impressive success in generating high-quality images from textual prompts. While large language models (LLMs) effectively leverage Direct Preference Optimization (DPO) for fine-tuning on human preference data without the need for reward models, diffusion models have not been extensively explored in this area. Current preference learning methods applied to T2I diffusion models immediately adapt existing techniques from LLMs. However, this direct adaptation introduces an estimated loss specific to T2I diffusion models. This estimation can potentially lead to suboptimal performance through our empirical results. In this work, we propose Direct Score Preference Optimization (DSPO), a novel algorithm that aligns the pretraining and fine-tuning objectives of diffusion models by leveraging score matching, the same objective used during pretraining. It introduces a new perspective on preference learning for diffusion models. Specifically, DSPO distills the score function of human-preferred image distributions into pretrained diffusion models, fine-tuning the model to generate outputs that align with human preferences. We theoretically show that DSPO shares the same optimization direction as reinforcement learning algorithms in diffusion models under certain conditions. Our experimental results demonstrate that DSPO outperforms preference learning baselines for T2I diffusion models in human preference evaluation tasks and enhances both visual appeal and prompt alignment of generated images.
****

Chenxin Li,Hengyu Liu,Zhiwen Fan,Wuyang Li,Yifan Liu,Panwang Pan,Yixuan Yuan
InstantSplamp: Fast and Generalizable Stenography Framework for Generative Gaussian Splatting
With the rapid development of large generative models for 3D, especially the evolution from NeRF representations to more efficient Gaussian Splatting, the synthesis of 3D assets has become increasingly fast and efficient, enabling the large-scale publication and sharing of generated 3D objects. However, while existing methods can add watermarks or steganographic information to individual 3D assets, they often require time-consuming per-scene training and optimization, leading to watermarking overheads that can far exceed the time required for asset generation itself, making deployment impractical for generating large collections of 3D objects. To address this, we propose InstantSplamp a framework that seamlessly integrates the 3D steganography pipeline into large 3D generative models without introducing explicit additional time costs. Guided by visual foundation models,InstantSplamp subtly injects hidden information like copyright tags during asset generation, enabling effective embedding and recovery of watermarks within generated 3D assets while preserving original visual quality. Experiments across various potential deployment scenarios demonstrate that \model~strikes an optimal balance between rendering quality and hiding fidelity, as well as between hiding performance and speed. Compared to existing per-scene optimization techniques for 3D assets, InstantSplamp reduces their watermarking training overheads that

are multiples of generation time to nearly zero, paving the way for real-world d
eployment at scale. Project page: https://gaussian-stego.github.io/.
****
Weifeng Lin,Xinyu Wei,Renrui Zhang,Le Zhuo,Shitian Zhao,Siyuan Huang,Junlin Xie,
Peng Gao,Hongsheng Li
PixWizard: Versatile Image-to-Image Visual Assistant with Open-Language Instruct
ions
This paper presents a versatile image-to-image visual assistant, PixWizard, desi
gned for image generation, manipulation, and translation based on free-from lang
uage instructions. To this end, we tackle a variety of vision tasks into a unifi
ed image-text-to-image generation framework and curate an Omni Pixel-to-Pixel In
struction-Tuning Dataset. By constructing detailed instruction templates in natu
ral language, we comprehensively include a large set of diverse vision tasks suc
h as text-to-image generation, image restoration, image grounding, dense image p
rediction, image editing, controllable generation, inpainting/outpainting, and m
ore. Furthermore, we adopt Diffusion Transformers (DiT) as our foundation model
and extend its capabilities with a flexible any resolution mechanism, enabling t
he model to dynamically process images based on the aspect ratio of the input, c
losely aligning with human perceptual processes. The model also incorporates str
ucture-aware and semantic-aware guidance to facilitate effective fusion of infor
mation from the input image. Our experiments demonstrate that PixWizard not only
 shows impressive generative and understanding abilities for images with divers
e resolutions but also exhibits generalization capabilities with unseen tasks an
d human instructions.
****
Jialiang Cheng,Ning Gao,Yun Yue,Zhiling Ye,Jiadi Jiang,Jian Sha
EDiT: A Local-SGD-Based Efficient Distributed Training Method for Large Language
 Models
Distributed training methods are crucial for large language models (LLMs). Howev
er, existing distributed training methods often suffer from communication bottle
necks, stragglers, and limited elasticity, particularly in heterogeneous or larg
e-scale environments. Local SGD methods have been proposed to address these issu
es, but their effectiveness remains limited to small-scale training due to addit
ional memory overhead and lack of concerns on efficiency and stability. To tackl
e these issues, we propose EDiT, an innovative Efficient Distributed Training me
thod that combines a tailored Local SGD approach with model sharding techniques
to enhance large-scale training efficiency. EDiT performs layer-wise parameter s
ynchronization during forward pass, reducing communication and memory overhead a
nd enabling overlap. Besides, EDiT employs a pseudo gradient penalty strategy to
 suppress loss spikes, which ensures training stability and improves performance
. Additionally, we introduce A-EDiT, a fully asynchronous variant of EDiT that a
ccommodates heterogeneous clusters. Building on EDiT/A-EDiT, we conduct a series
 of experiments to validate large-scale asynchronous training for LLMs, accompan
ied by comprehensive analyses. Experimental results demonstrate the superior per
formance of EDiT/A-EDiT, establishing them as robust solutions for distributed L
LM training in diverse computational ecosystems. The code is available at Atorch
 codebase: https://github.com/intelligent-machine-learning/atorch/tree/main/ator
ch/local_sgd.
****
Shiping Gao,Fanqi Wan,Jiajian Guo,Xiaojun Quan,Qifan Wang
Advantage-Guided Distillation for Preference Alignment in Small Language Models
Alignment techniques enable Large Language Models (LLMs) to generate outputs tha
t align with human preferences and play a crucial role in their effectiveness. H
owever, their impact often diminishes when applied to Small Language Models (SLM
s), likely due to the limited capacity of these models. Instead of directly appl
ying existing alignment techniques to SLMs, we propose to utilize a well-aligned
 teacher LLM to guide the alignment process for these models, thereby facilitati
ng the transfer of the teacher's knowledge of human preferences to the student m
odel. To achieve this, we first explore a straightforward approach, Dual-Constra
ined Knowledge Distillation (DCKD), that employs knowledge distillation with two

KL-divergence constraints from the aligned teacher to the unaligned student. To further enhance the student's ability to distinguish between preferred and disp
referred responses, we then propose Advantage-Guided Distillation for Preference Alignment (ADPA), which leverages an advantage function from the aligned teache
r to deliver more nuanced, distribution-level reward signals for the student's a
lignment. Our experimental results show that these two approaches appreciably im
prove the alignment of SLMs and narrow the performance gap with larger counterpa
rts. Among them, ADPA demonstrates superior performance and achieves even greate
r effectiveness when integrated with DCKD. Our code is available at https://gith
ub.com/SLIT-AI/ADPA .
****
Lianghui Zhu,Xinggang Wang,Xinlong Wang
JudgeLM: Fine-tuned Large Language Models are Scalable Judges
Evaluating Large Language Models (LLMs) in open-ended scenarios is challenging b
ecause existing benchmarks and metrics can not measure them comprehensively. To
address this problem, we propose to fine-tune LLMs as scalable judges (JudgeLM)
to evaluate LLMs efficiently and effectively in open-ended benchmarks. We first
propose a comprehensive, large-scale, high-quality dataset containing task seeds
, LLMs-generated answers, and GPT-4-generated judgments for fine-tuning high-per
formance judges, as well as a new benchmark for evaluating the judges. We train
JudgeLM at different scales from 7B, 13B, to 33B parameters, and conduct a syste
matic analysis of its capabilities and behaviors. We then analyze the key biases
in fine-tuning LLM as a judge and consider them as position bias, knowledge bia
s, and format bias. To address these issues, JudgeLM introduces a bag of techniq
ues including swap augmentation, reference support, and reference drop, which cl
early enhance the judge's performance. JudgeLM obtains the state-of-the-art judg
e performance on both the existing PandaLM benchmark and our proposed new benchm
ark. Our JudgeLM is efficient and the JudgeLM-7B only needs 3 minutes to judge 5
K samples with 8 A100 GPUs. JudgeLM obtains high agreement with the teacher judg
e, achieving an agreement exceeding 90% that even surpasses human-to-human agree
ment. JudgeLM also demonstrates extended capabilities in being judges of the sin
gle answer, multimodal models, multiple answers, multi-turn chat, etc.
****
João Loula,Benjamin LeBrun,Li Du,Ben Lipkin,Clemente Pasti,Gabriel Grand,Tianyu
Liu,Yahya Emara,Marjorie Freedman,Jason Eisner,Ryan Cotterell,Vikash Mansinghka,
Alexander K. Lew,Tim Vieira,Timothy J. O'Donnell
Syntactic and Semantic Control of Large Language Models via Sequential Monte Car
lo
A wide range of LM applications require generating text that conforms to syntact
ic or semantic constraints. Imposing such constraints can be naturally framed as
probabilistic conditioning, but exact generation from the resulting distributio
n—which can differ substantially from the LM’s base distribution—is generally in
tractable. In this work, we develop an architecture for controlled LM generation
based on sequential Monte Carlo (SMC). This SMC framework allows us to flexibly
incorporate domain- and problem-specific constraints at inference
time, and efficiently reallocate computational resources in light of new informa
tion during the course of generation. By comparing to a number of alternatives a
nd ablations on four challenging domains—Python code generation for data science
, text-to-SQL, goal inference, and molecule synthesis—we demonstrate that, with
little overhead, our approach allows small open-source language models to outper
form models over 8× larger, as well as closed-source, fine-tuned ones. In suppor
t of the probabilistic perspective, we show that these performance improvements
are driven by better approximation to the posterior distribution. [Our system](h
ttps://github.com/probcomp/genparse) builds on the framework of Lew et al. (2023
) and integrates with its language model probabilistic programming language, giv
ing users a simple, programmable way to apply SMC to a broad variety of controll
ed generation problems.
****
Chongyi Zheng,Jens Tuyls,Joanne Peng,Benjamin Eysenbach
Can a MISL Fly? Analysis and Ingredients for Mutual Information Skill Learning

Self-supervised learning has the potential of lifting several of the key challen
ges in reinforcement learning today, such as exploration, representation learnin
g, and reward design. Recent work (METRA) has effectively argued that moving awa
y from mutual information and instead optimizing a certain Wasserstein distance
is important for good performance. In this paper, we argue that the benefits see
n in that paper can largely be explained within the existing framework of mutual
 information skill learning (MISL).
Our analysis suggests a new MISL method (contrastive successor features) that re
tains the excellent performance of METRA with fewer moving parts, and highlights
 connections between skill learning, contrastive representation learning, and su
ccessor features. Finally, through careful ablation studies, we provide further
insight into some of the key ingredients for both our method and METRA.
****
Daniel Levy,Siba Smarak Panigrahi,Sékou-Oumar Kaba,Qiang Zhu,Kin Long Kelvin Lee
,Mikhail Galkin,Santiago Miret,Siamak Ravanbakhsh
SymmCD: Symmetry-Preserving Crystal Generation with Diffusion Models
Generating novel crystalline materials has potential to lead to advancements in
fields such as electronics, energy storage, and catalysis. The defining characte
ristic of crystals is their symmetry, which plays a central role in determining
their physical properties. However, existing crystal generation methods either f
ail to generate materials that display the symmetries of real-world crystals, or
 simply replicate the symmetry information from examples in a database. To addr
ess this limitation, we propose SymmCD, a novel diffusion-based generative model
 that explicitly incorporates crystallographic symmetry into the generative proc
ess. We decompose crystals into two components and learn their joint distributio
n through diffusion: 1) the asymmetric unit, the smallest subset of the crystal
 which can generate the whole crystal through symmetry transformations, and; 2)
the symmetry transformations needed to be applied to each atom in the asymmetric
 unit. We also use a novel and interpretable representation for these transforma
tions, enabling generalization across different crystallographic symmetry groups
. We showcase the competitive performance of SymmCD on a subset of the Materials
 Project, obtaining diverse and valid crystals with realistic symmetries and pre
dicted properties.
****
Solenne Gaucher,Martino Bernasconi,Matteo Castiglioni,Andrea Celli,Vianney Perch
et
Feature-Based Online Bilateral Trade
Bilateral trade models the problem of facilitating trades between a seller and a
 buyer having private valuations for the item being sold. In the online version
of the problem, the learner faces a new seller and buyer at each time step, and
has to post a price for each of the two parties without any knowledge of their v
aluations. We consider a scenario where, at each time step, before posting price
s the learner observes a context vector containing information about the feature
s of the item for sale. The valuations of both the seller and the buyer follow a
n unknown linear function of the context. In this setting, the learner could lev
erage previous transactions in an attempt to estimate private valuations. We cha
racterize the regret regimes of different settings, taking as a baseline the bes
t context-dependent prices in hindsight. First, in the setting in which the lear
ner has two-bit feedback and strong budget balance constraints, we propose an al
gorithm with $O(\log T)$ regret. Then, we study the same set-up with noisy valua
tions, providing a tight $\widetilde O(T^{2/3})$ regret upper bound. Finally, we
 show that loosening budget balance constraints allows the learner to operate un
der more restrictive feedback. Specifically, we show how to address the one-bit,
 global budget balance setting through a reduction from the two-bit, strong budg
et balance setup. This established a fundamental trade-off between the quality o
f the feedback and the strictness of the budget constraints.
****
Carlo Abate,Filippo Maria Bianchi
MaxCutPool: differentiable feature-aware Maxcut for pooling in graph neural netw
orks

We propose a novel approach to compute the MAXCUT in attributed graphs, i.e., gr
aphs with features associated with nodes and edges. Our approach works well on a
ny kind of graph topology and can find solutions that jointly optimize the MAXCU
T along with other objectives. Based on the obtained MAXCUT partition, we implem
ent a hierarchical graph pooling layer for Graph Neural Networks, which is spars
e, trainable end-to-end, and particularly suitable for downstream tasks on heter
ophilic graphs.
****

SUBBA REDDY OOTA,Akshett Rai Jindal,Ishani Mondal,Khushbu Pahwa,Satya Sai Srinat
h Namburi GNVV,Manish Shrivastava,Maneesh Kumar Singh,Bapi Raju Surampudi,Manish
 Gupta

Correlating instruction-tuning (in multimodal models) with vision-language proce
ssing (in the brain)

Transformer-based language models, though not explicitly trained to mimic brain
recordings, have demonstrated surprising alignment with brain activity. Progress
 in these models—through increased size, instruction-tuning, and multimodality—h
as led to better representational alignment with neural data. Recently, a new cl
ass of instruction-tuned multimodal LLMs (MLLMs) have emerged, showing remarkabl
e zero-shot capabilities in open-ended multimodal vision tasks. However, it is u
nknown whether MLLMs, when prompted with natural instructions, lead to better br
ain alignment and effectively capture instruction-specific representations. To a
ddress this, we first investigate the brain alignment, i.e., measuring the degre
e of predictivity of neural visual activity using text output response embedding
s from MLLMs as participants engage in watching natural scenes. Experiments with
 10 different instructions (like image captioning, visual question answering, et
c.) show that MLLMs exhibit significantly better brain alignment than vision-on
ly models and perform comparably to non-instruction-tuned multimodal models like
 CLIP. We also find that while these MLLMs are effective at generating high-qual
ity responses suitable to the task-specific instructions, not all instructions a
re relevant for brain alignment. Further, by varying instructions, we make the M
LLMs encode instruction-specific visual concepts related to the input image. Thi
s analysis shows that MLLMs effectively capture count-related and recognition-re
lated concepts, demonstrating strong alignment with brain activity. Notably, the
 majority of the explained variance of the brain encoding models is shared betwe
en MLLM embeddings of image captioning and other instructions. These results ind
icate that enhancing MLLMs' ability to capture more task-specific information co
uld allow for better differentiation between various types of instructions, and
hence improve their precision in predicting brain responses.
****

Zhaorun Chen,Francesco Pinto,Minzhou Pan,Bo Li

SafeWatch: An Efficient Safety-Policy Following Video Guardrail Model with Trans
parent Explanations

With the rise of generative AI and rapid growth of high-quality video generation
, video guardrails have become more crucial than ever to ensure safety and secur
ity across platforms. Current video guardrails, however, are either overly simpl
istic, relying on pure classification models trained on simple policies with lim
ited unsafe categories, which lack detailed explanations, or prompting multimoda
l large language models (MLLMs) with long safety guidelines, which are inefficie
nt and impractical for guardrailing real-world content. To bridge this gap, we p
ropose SafeWatch, an efficient MLLM-based video guardrail model designed to foll
ow customized safety policies and provide multi-label video guardrail outputs wi
th content-specific explanations in a zero-shot manner. In particular, unlike tr
aditional MLLM-based guardrails that encode all safety policies autoregressively
, causing inefficiency and bias, SafeWatch uniquely encodes each policy chunk in
 parallel and eliminates their position bias such that all policies are attended
 simultaneously with equal importance. In addition, to improve efficiency and ac
curacy, SafeWatch incorporates a policy-aware visual token pruning algorithm tha
t adaptively selects the most relevant video tokens for each policy, discarding
noisy or irrelevant information. This allows for more focused, policy-compliant
guardrail with significantly reduced computational overhead. Considering the lim

itations of existing video guardrail benchmarks, we propose SafeWatch-Bench, a l
arge-scale video guardrail benchmark comprising over 2M videos spanning six safe
ty categories which covers over 30 tasks to ensure a comprehensive coverage of a
ll potential safety scenarios. We have conducted extensive experiments, showing
that SafeWatch outperforms all SOTA video guardrails on SafeWatch-Bench by 28.2%
, and achieves a 13.6% improvement on existing benchmarks, all while reducing in
ference costs by an average of 10%. SafeWatch also demonstrates strong policy-fo
llowing abilities and outperforms previous SOTAs by 5.6% and 15.6% in zero-shot
generalizability to new policies and new prompting tasks. Additionally, both LLM
-as-a-judge and human evaluators confirm the high quality of the explanations pr
ovided by SafeWatch. Our project is open-sourced at https://safewatch-aiguard.gi
thub.io.
****
Hakaze Cho,Mariko Kato,Yoshihiro Sakai,Naoya Inoue
Revisiting In-context Learning Inference Circuit in Large Language Models
In-context Learning (ICL) is an emerging few-shot learning paradigm on Language
Models (LMs) with inner mechanisms un-explored. There are already existing works
 describing the inner processing of ICL, while they struggle to capture all the
inference phenomena in large language models. Therefore, this paper proposes a c
omprehensive circuit to model the inference dynamics and try to explain the obse
rved phenomena of ICL. In detail, we divide ICL inference into 3 major operation
s: (1) Input Text Encode: LMs encode every input text (in the demonstrations and
 queries) into linear representation in the hidden states with sufficient inform
ation to solve ICL tasks. (2) Semantics Merge: LMs merge the encoded representat
ions of demonstrations with their corresponding label tokens to produce joint re
presentations of labels and demonstrations. (3) Feature Retrieval and Copy: LMs
search the joint representations of demonstrations similar to the query represen
tation on a task subspace, and copy the searched representations into the query.
 Then, language model heads capture these copied label representations to a cert
ain extent and decode them into predicted labels. Through careful measurements,
the proposed inference circuit successfully captures and unifies many fragmented
 phenomena observed during the ICL process, making it a comprehensive and practi
cal explanation of the ICL inference process. Moreover, ablation analysis by dis
abling the proposed steps seriously damages the ICL performance, suggesting the
proposed inference circuit is a dominating mechanism. Additionally, we confirm a
nd list some bypass mechanisms that solve ICL tasks in parallel with the propose
d circuit.
****
Melis Ilayda Bal,Pier Giuseppe Sessa,Mojmir Mutny,Andreas Krause
Optimistic Games for Combinatorial Bayesian Optimization with Application to Pro
tein Design
Bayesian optimization (BO) is a powerful framework to optimize black-box expensi
ve-to-evaluate functions via sequential interactions. In several important probl
ems (e.g. drug discovery, circuit design, neural architecture search, etc.), tho
ugh, such functions are defined over large $\textit{combinatorial and unstructur
ed}$ spaces. This makes existing BO algorithms not feasible due to the intractab
le maximization of the acquisition function over these domains. To address this
issue, we propose $\textbf{GameOpt}$, a novel game-theoretical approach to combi
natorial BO. $\textbf{GameOpt}$ establishes a cooperative game between the diffe
rent optimization variables, and selects points that are game $\textit{equilibri
a}$ of an upper confidence bound acquisition function. These are stable configur
ations from which no variable has an incentive to deviate$-$ analog to local opt
ima in continuous domains. Crucially, this allows us to efficiently break down t
he complexity of the combinatorial domain into individual decision sets, making
$\textbf{GameOpt}$ scalable to large combinatorial spaces. We demonstrate the ap
plication of $\textbf{GameOpt}$ to the challenging $\textit{protein design}$ pro
blem and validate its performance on four real-world protein datasets. Each prot
ein can take up to 20^{X} possible configurations, where X is the length of
a protein, making standard BO methods infeasible. Instead, our approach iterativ
ely selects informative protein configurations and very quickly discovers highly

active protein variants compared to other baselines.
****
Yufan Zhuang,Chandan Singh,Liyuan Liu,Jingbo Shang,Jianfeng Gao
Vector-ICL: In-context Learning with Continuous Vector Representations
Large language models (LLMs) have shown remarkable in-context learning (ICL) cap abilities on textual data. We explore whether these capabilities can be extended to continuous vectors from diverse domains, obtained from black-box pretrained encoders. By aligning input data with an LLM's embedding space through lightweig ht projectors, we observe that LLMs can effectively process and learn from these projected vectors, which we term Vector-ICL. In particular, we find that pretra ining projectors with general language modeling objectives enables Vector-ICL, w hile task-specific finetuning further enhances performance. In our experiments a cross various tasks and modalities, including text reconstruction, numerical fun ction regression, text classification, summarization, molecule captioning, time-series classification, graph classification, and fMRI decoding, Vector-ICL often surpasses both few-shot ICL and domain-specific model or tuning. We further con duct analyses and case studies, indicating the potential of LLMs to process vect or representations beyond traditional token-based paradigms.
****
Aditya T. Vadlamani,Anutam Srinivasan,Pranav Maneriker,Ali Payani,srinivasan par thasarathy
A Generic Framework for Conformal Fairness
Conformal Prediction (CP) is a popular method for uncertainty quantification wit h machine learning models. While conformal prediction provides probabilistic gua rantees regarding the coverage of the true label, these guarantees are agnostic to the presence of sensitive attributes within the dataset. In this work, we for malize \textit{Conformal Fairness}, a notion of fairness using conformal predict ors, and provide a theoretically well-founded algorithm and associated framework to control for the gaps in coverage between different sensitive groups. Our fra mework leverages the exchangeability assumption (implicit to CP) rather than the typical IID assumption, allowing us to apply the notion of Conformal Fairness t o data types and tasks that are not IID, such as graph data. Experiments were co nducted on graph and tabular datasets to demonstrate that the algorithm can cont rol fairness-related gaps in addition to coverage aligned with theoretical expec tations.
****
Jun Luo,Chen Chen,Shandong Wu
Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Langu age Models
Federated prompt learning benefits federated learning with CLIP-like Vision-Lang uage Model's (VLM's) robust representation learning ability through prompt learn ing. However, current federated prompt learning methods are habitually restricte d to the traditional FL paradigm, where the participating clients are generally only allowed to download a single globally aggregated model from the server. Whi le justifiable for training full-sized models under federated settings, in this work, we argue that this paradigm is ill-suited for lightweight prompts. By faci litating the clients to download multiple pre-aggregated prompts as fixed non-lo cal experts, we propose Personalized Federated Mixture of Adaptive Prompts (pFed MoAP), a novel FL framework that personalizes the prompt learning process throug h the lens of Mixture of Experts (MoE). pFedMoAP implements a local attention-ba sed gating network that learns to generate enhanced text features for better ali gnment with local image data, benefiting from both local and downloaded non-loca l adaptive prompt experts. Extensive experiments on 9 datasets under various fed erated settings demonstrate the efficacy of the proposed pFedMoAP algorithm. The code is available at https://github.com/ljaiverson/pFedMoAP.
****
Zehui Chen,Kuikun Liu,Qiuchen Wang,Jiangning Liu,Wenwei Zhang,Kai Chen,Feng Zhao
MindSearch: Mimicking Human Minds Elicits Deep AI Searcher
Information seeking and integration is a complex cognitive task that consumes en ormous time and effort. Inspired by the remarkable progress of Large Language Mo

dels, recent works attempt to solve this task by combining LLMs and search engin
es. However, these methods still obtain unsatisfying performance due to three ch
allenges: (1) complex requests often cannot be accurately and completely retriev
ed by the search engine once (2) corresponding information to be integrated is s
pread over multiple web pages along with massive noise, and (3) a large number o
f web pages with long contents may quickly exceed the maximum context length of
LLMs. Inspired by the cognitive process when humans solve these problems, we int
roduce MindSearch to mimic the human minds in web information seeking and integr
ation, which can be instantiated by a simple yet effective LLM-based multi-agent
 framework. The WebPlanner models the human mind of multi-step information seeki
ng as a dynamic graph construction process: it decomposes the user query into at
omic sub-questions as nodes in the graph and progressively extends the graph bas
ed on the search result from WebSearcher. Tasked with each sub-question, WebSear
cher performs hierarchical information retrieval with search engines and collect
s valuable information for WebPlanner. The multi-agent design of MindSearch enab
les the whole framework to seek and integrate information parallelly from larger
-scale (e.g., more than 300) web pages in 3 minutes, which is worth 3 hours of h
uman effort. MindSearch demonstrates significant improvement in the response qua
lity in terms of depth and breadth, on both close-set and open-set QA problems.
Besides, responses from MindSearch based on InternLM2.5-7B are preferable by hum
ans to ChatGPT-Web and Perplexity.ai applications, which implies that MindSearch
 can already deliver a competitive solution to the proprietary AI search engine.
****

Zilong Wang,Zifeng Wang,Long Le,Steven Zheng,Swaroop Mishra,Vincent Perot,Yuwei
Zhang,Anush Mattapalli,Ankur Taly,Jingbo Shang,Chen-Yu Lee,Tomas Pfister
Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting
Retrieval augmented generation (RAG) combines the generative abilities of large
language models (LLMs) with external knowledge sources to provide more accurate
and up-to-date responses. Recent RAG advancements focus on improving retrieval o
utcomes through iterative LLM refinement or self-critique capabilities acquired
through additional instruction tuning of LLMs. In this work, we introduce Specul
ative RAG - a framework that leverages a larger generalist LM to efficiently ver
ify multiple RAG drafts produced in parallel by a smaller, distilled specialist
LM. Each draft is generated from a distinct subset of retrieved documents, offer
ing diverse perspectives on the evidence while reducing input token counts per d
raft. This approach enhances comprehension of each subset and mitigates potentia
l position bias over long context. Our method accelerates RAG by delegating draf
ting to the smaller specialist LM, with the larger generalist LM performing a si
ngle verification pass over the drafts. Extensive experiments demonstrate that S
peculative RAG achieves state-of-the-art performance with reduced latency on Tri
viaQA, MuSiQue, PopQA, PubHealth, and ARC-Challenge benchmarks. It notably enhan
ces accuracy by up to 12.97% while reducing latency by 50.83% compared to conven
tional RAG systems on PubHealth.
****

Tiexin Qin,Mengxu ZHU,Chunyang Li,Terry Lyons,Hong Yan,Haoliang Li
Deep Signature: Characterization of Large-Scale Molecular Dynamics
Understanding protein dynamics are essential for deciphering protein functional
mechanisms and developing molecular therapies. However, the complex high-dimensi
onal dynamics and interatomic interactions of biological processes pose signific
ant challenge for existing computational techniques. In this paper, we approach
this problem for the first time by introducing Deep Signature, a novel computati
onally tractable framework that characterizes complex dynamics and interatomic i
nteractions based on their evolving trajectories. Specifically, our approach inc
orporates soft spectral clustering that locally aggregates cooperative dynamics
to reduce the size of the system, as well as signature transform that collects i
terated integrals to provide a global characterization of the non-smooth interac
tive dynamics. Theoretical analysis demonstrates that Deep Signature exhibits se
veral desirable properties, including invariance to translation, near invariance
 to rotation, equivariance to permutation of atomic coordinates, and invariance
under time reparameterization. Furthermore, experimental results on three benchm

arks of biological processes verify that our approach can achieve superior perfo
rmance compared to baseline methods.
****
Arnab Bhattacharyya,Sutanu Gayen,Kuldeep S. Meel,Dimitrios Myrisiotis,A. Pavan,N
. V. Vinodchandran
Computational Explorations of Total Variation Distance
We investigate some previously unexplored (or underexplored) computational aspec
ts of total variation (TV) distance.
First, we give a simple deterministic polynomial-time algorithm for checking equ
ivalence between mixtures of product distributions, over arbitrary alphabets.
This corresponds to a special case, whereby the TV distance between the two dist
ributions is zero.
Second, we prove that unless $\mathsf{NP} \subseteq \mathsf{RP}$ it is impossibl
e to efficiently estimate the TV distance between arbitrary Ising models, even i
n a bounded-error randomized setting.
****
Kaizhe Hu,Zihang Rui,Yao He,Yuyao Liu,Pu Hua,Huazhe Xu
Stem-OB: Generalizable Visual Imitation Learning with Stem-Like Convergent Obser
vation through Diffusion Inversion
Visual imitation learning methods demonstrate strong performance, yet they lack
generalization when faced with visual input perturbations like variations in lig
hting and textures. This limitation hampers their practical application in real-
world settings. To address this, we propose ***Stem-OB*** that leverages the inv
ersion process of pretrained image diffusion models to suppress low-level visual
 differences while maintaining high-level scene structures. This image inversion
 process is akin to transforming the observation into a shared representation, f
rom which other observations also stem. *Stem-OB* offers a simple yet effective
plug-and-play solution that stands in contrast to data augmentation approaches.
It demonstrates robustness to various unspecified appearance changes without the
 need for additional training. We provide theoretical insights and empirical res
ults that validate the efficacy of our approach in simulated and real settings.
Stem-OB shows an exceptionally significant improvement in real-world robotic t
asks, where challenging light and appearance changes are present, with an averag
e increase of **22.2%** in success rates compared to the best baseline. Please r
efer to [this link](https://stem-ob.github.io/) for more videos and details.
****
Niklas Muennighoff,Luca Soldaini,Dirk Groeneveld,Kyle Lo,Jacob Morrison,Sewon Mi
n,Weijia Shi,Evan Pete Walsh,Oyvind Tafjord,Nathan Lambert,Yuling Gu,Shane Arora
,Akshita Bhagia,Dustin Schwenk,David Wadden,Alexander Wettig,Binyuan Hui,Tim Det
tmers,Douwe Kiela,Ali Farhadi,Noah A. Smith,Pang Wei Koh,Amanpreet Singh,Hannane
h Hajishirzi
OLMoE: Open Mixture-of-Experts Language Models
We introduce OLMoE, a fully open, state-of-the-art language model leveraging spa
rse Mixture-of-Experts (MoE). OLMoE-1B-7B has 7 billion (B) parameters but uses
only 1B per input token. We pretrain it on 5 trillion tokens and further adapt i
t to create OLMoE-1B-7B-Instruct. Our models outperform all available models wit
h similar active parameters, even surpassing larger ones like Llama2-13B-Chat an
d DeepSeekMoE-16B. We present novel findings on MoE training, define and analyze
 new routing properties showing high specialization in our model, and open-sourc
e all our work: model weights, training data, code, and logs.
****
Yanbo Zhang,Benedikt Hartl,Hananel Hazan,Michael Levin
Diffusion Models are Evolutionary Algorithms
In a convergence of machine learning and biology, we reveal that diffusion model
s are evolutionary algorithms. By considering evolution as a denoising process a
nd reversed evolution as diffusion, we mathematically demonstrate that diffusion
 models inherently perform evolutionary algorithms, naturally encompassing selec
tion, mutation, and reproductive isolation. Building on this equivalence, we pro
pose the Diffusion Evolution method: an evolutionary algorithm utilizing iterati
ve denoising -- as originally introduced in the context of diffusion models -- t

o heuristically refine solutions in parameter spaces. Unlike traditional approac
hes, Diffusion Evolution efficiently identifies multiple optimal solutions and o
utperforms prominent mainstream evolutionary algorithms. Furthermore, leveraging
 advanced concepts from diffusion models, namely latent space diffusion and acce
lerated sampling, we introduce Latent Space Diffusion Evolution, which finds sol
utions for evolutionary tasks in high-dimensional complex parameter space while
significantly reducing computational steps. This parallel between diffusion and
evolution not only bridges two different fields but also opens new avenues for m
utual enhancement, raising questions about open-ended evolution and potentially
utilizing non-Gaussian or discrete diffusion models in the context of Diffusion
Evolution.
****
Tianchi Xie,Jiangning Zhu,Guozu Ma,Minzhi Lin,Wei Chen,Weikai Yang,Shixia Liu
Structural-Entropy-Based Sample Selection for Efficient and Effective Learning
Sample selection improves the efficiency and effectiveness of machine learning m
odels by providing informative and representative samples. Typically, samples ca
n be modeled as a sample graph, where nodes are samples and edges represent thei
r similarities. Most existing methods are based on local information, such as th
e training difficulty of samples, thereby overlooking global information, such a
s connectivity patterns. This oversight can result in suboptimal selection becau
se global information is crucial for ensuring that the selected samples well rep
resent the structural properties of the graph. To address this issue, we employ
structural entropy to quantify global information and losslessly decompose it fr
om the whole graph to individual nodes using the Shapley value. Based on the dec
omposition, we present $\textbf{S}$tructural-$\textbf{E}$ntropy-based sample $\t
extbf{S}$election ($\textbf{SES}$), a method that integrates both global and loc
al information to select informative and representative samples. SES begins by c
onstructing a kNN-graph among samples based on their similarities. It then mea
sures sample importance by combining structural entropy (global metric) with tra
ining difficulty (local metric). Finally, SES applies importance-biased blue noi
se sampling to select a set of diverse and representative samples. Comprehensive
 experiments on three learning scenarios --- supervised learning, active learnin
g, and continual learning --- clearly demonstrate the effectiveness of our metho
d.
****
Zhengyang Geng,Ashwini Pokle,Weijian Luo,Justin Lin,J Zico Kolter
Consistency Models Made Easy
Consistency models (CMs) offer faster sampling than traditional diffusion models
, but their training is resource-intensive. For example, as of 2024, training a
state-of-the-art CM on CIFAR-10 takes one week on 8 GPUs. In this work, we propo
se an effective scheme for training CMs that largely improves the efficiency of
building such models. Specifically, by expressing CM trajectories via a particul
ar differential equation, we argue that diffusion models can be viewed as a spec
ial case of CMs. We can thus fine-tune a consistency model starting from a pretr
ained diffusion model and progressively approximate the full consistency conditi
on to stronger degrees over the training process. Our resulting method, which we
 term Easy Consistency Tuning (ECT), achieves vastly reduced training times whil
e improving upon the quality of previous methods: for example, ECT achieves a 2-
step FID of 2.73 on CIFAR10 within 1 hour on a single A100 GPU, matching Consist
ency Distillation trained for hundreds of GPU hours. Owing to this computational
 efficiency, we investigate the scaling laws of CMs under ECT, showing that they
 obey the classic power law scaling, hinting at their ability to improve efficie
ncy and performance at larger scales. Our [code](https://github.com/locuslab/ect
) is available.
****
Jerry Zhi-Yang He,Sashrika Pandey,Mariah L Schrum,Anca Dragan
Context Steering: Controllable Personalization at Inference Time
To deliver high-quality, personalized responses, large language models (LLMs) mu
st effectively incorporate context — personal, demographic, and cultural informa
tion specific to an end-user. For example, asking the model to explain Newton's

second law with the context "I am a toddler'' should produce a response differen
t from when the context is "I am a physics professor''. However, leveraging the
context in practice is a nuanced and challenging task, and is often dependent on
 the specific situation or user base. The model must strike a balance between pr
oviding specific, personalized responses and maintaining general applicability.
Current solutions, such as prompt-engineering and fine-tuning, require collectio
n of contextually appropriate responses as examples, making them time-consuming
and less flexible to use across different contexts. In this work, we introduce C
ontext Steering (CoS) —a simple, training-free decoding approach that amplifies
the influence of the context in next token predictions. CoS computes contextual
influence by comparing the output probabilities from two LLM forward passes: one
 that includes the context and one that does not. By linearly scaling the contex
tual influence, CoS allows practitioners to flexibly control the degree of perso
nalization for different use cases. We show that CoS can be applied to autoregre
ssive LLMs, and demonstrates strong performance in personalized recommendations.
 Additionally, we show that CoS can function as a Bayesian Generative model to i
nfer and quantify correlations between open-ended texts, broadening its potentia
l applications.
****
Carles Domingo-Enrich,Michal Drozdzal,Brian Karrer,Ricky T. Q. Chen
Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryle
ss Stochastic Optimal Control
Dynamical generative models that produce samples through an iterative process, s
uch as Flow Matching and denoising diffusion models, have seen widespread use, b
ut there have not been many theoretically-sound methods for improving these mode
ls with reward fine-tuning. In this work, we cast reward fine-tuning as stochast
ic optimal control (SOC). Critically, we prove that a very specific *memoryless*
 noise schedule must be enforced during fine-tuning, in order to account for the
 dependency between the noise variable and the generated samples. We also propos
e a new algorithm named *Adjoint Matching* which outperforms existing SOC algori
thms, by casting SOC problems as a regression problem. We find that our approach
 significantly improves over existing methods for reward fine-tuning, achieving
better consistency, realism, and generalization to unseen human preference rewar
d models, while retaining sample diversity.
****
Yuxuan Yao,Zixuan Zeng,Chun Gu,Xiatian Zhu,Li Zhang
Reflective Gaussian Splatting
Novel view synthesis has experienced significant advancements owing to increasin
gly capable NeRF- and 3DGS-based methods. However, reflective object reconstruct
ion remains challenging, lacking a proper solution to achieve real-time, high-qu
ality rendering while accommodating inter-reflection. To fill this gap, we intro
duce a Reflective Gaussian splatting (Ref-Gaussian) framework characterized with
 two components: (I) Physically based deferred rendering that empowers the rende
ring equation with pixel-level material properties via formulating split-sum app
roximation; (II) Gaussian-grounded inter-reflection that realizes the desired in
ter-reflection function within a Gaussian splatting paradigm for the first time.
 To enhance geometry modeling, we further introduce material-aware normal propag
ation and an initial per-Gaussian shading stage, along with 2D Gaussian primitiv
es. Extensive experiments on standard datasets demonstrate that Ref-Gaussian sur
passes existing approaches in terms of quantitative metrics, visual quality, and
 compute efficiency. Further, we show that our method serves as a unified soluti
on for both reflective and non-reflective scenes, going beyond the previous alte
rnatives focusing on only reflective scenes. Also, we illustrate that Ref-Gaussi
an supports more applications such as relighting and editing.
****
Hao Wang,zhengnan li,Haoxuan Li,Xu Chen,Mingming Gong,BinChen,Zhichao Chen
Optimal Transport for Time Series Imputation
Missing data imputation through distribution alignment has demonstrated advantag
es for non-temporal datasets but exhibits suboptimal performance in time-series
applications. The primary obstacle is crafting a discrepancy measure that simult

aneously (1) captures temporal patterns—accounting for periodicity and temporal dependencies inherent in time-series—and (2) accommodates non-stationarity, ensuring robustness amidst multiple coexisting temporal patterns. In response to these challenges, we introduce the Proximal Spectrum Wasserstein (PSW) discrepancy, a novel discrepancy tailored for comparing two \textit{sets} of time-series based on optimal transport. It incorporates a pairwise spectral distance to encapsulate temporal patterns, and a selective matching regularization to accommodate non-stationarity. Subsequently, we develop the PSW for Imputation (PSW-I) framework, which iteratively refines imputation results by minimizing the PSW discrepancy. Extensive experiments demonstrate that PSW-I effectively accommodates temporal patterns and non-stationarity, outperforming prevailing time-series imputation methods. Code is available at https://github.com/FMLYD/PSW-I.
****

Yuang Wang,Yujian Zhang,Sida Peng,Xingyi He,Haoyu Guo,Yujun Shen,Hujun Bao,Xiaowei Zhou
UniRestore3D: A Scalable Framework For General Shape Restoration
Shape restoration aims to recover intact 3D shapes from defective ones, such as those that are incomplete, noisy, and low-resolution. Previous works have achieved impressive results in shape restoration subtasks thanks to advanced generative models. While effective for specific shape defects, they are less applicable in real-world scenarios involving multiple defect types simultaneously. Additionally, training on limited subsets of defective shapes hinders knowledge transfer across restoration types and thus affects generalization. In this paper, we address the task of general shape restoration, which restores shapes with various types of defects through a unified model, thereby naturally improving the applicability and scalability. Our approach first standardizes the data representation across different restoration subtasks using high-resolution TSDF grids and constructs a large-scale dataset with diverse types of shape defects. Next, we design an efficient hierarchical shape generation model and a noise-robust defective shape encoder that enables effective impaired shape understanding and intact shape generation. Moreover, we propose a scalable training strategy for efficient model training. The capabilities of our proposed method are demonstrated across multiple shape restoration subtasks and validated on various datasets, including Objaverse, ShapeNet, GSO, and ABO.
****

Brian R.Y. Huang,Maximilian Li,Leonard Tang
Endless Jailbreaks with Bijection Learning
Despite extensive safety measures, LLMs are vulnerable to adversarial inputs, or jailbreaks, which can elicit unsafe behaviors. In this work, we introduce bijection learning, a powerful attack algorithm which automatically fuzzes LLMs for safety vulnerabilities using randomly-generated encodings whose complexity can be tightly controlled. We leverage in-context learning to teach models bijective encodings, pass encoded queries to the model to bypass built-in safety mechanisms, and finally decode responses back into English. Our attack is extremely effective on a wide range of frontier language models. By controlling complexity parameters such as number of key-value mappings in the encodings, we find a close relationship between the capability level of the attacked LLM and the average complexity of the most effective bijection attacks. Our work highlights that new vulnerabilities in frontier models can emerge with scale: more capable models are more severely jailbroken by bijection attacks.
****

Neel Dey,Benjamin Billot,Hallee E. Wong,Clinton Wang,Mengwei Ren,Ellen Grant,Adrian V Dalca,Polina Golland
Learning General-purpose Biomedical Volume Representations using Randomized Synthesis
Current volumetric biomedical foundation models struggle to generalize as public 3D datasets are small and do not cover the broad diversity of medical procedures, conditions, anatomical regions, and imaging protocols. We address this by creating a representation learning method that instead anticipates strong domain shifts at training time itself. We first propose a data engine that synthesizes hi

ghly variable training samples that would enable generalization to new biomedica
l contexts. To then train a single 3D network for any voxel-level task, we devel
op a contrastive learning method that pretrains the network to be stable against
 nuisance imaging variation simulated by the data engine, a key inductive bias f
or generalization. This network's features can be used as robust representations
 of input images for downstream tasks and its weights provide a strong, _dataset
-agnostic_ initialization for finetuning on new datasets. As a result, we set ne
w standards across _both_ multimodality registration and few-shot segmentation,
a first for any 3D biomedical vision model, all without (pre-)training on any ex
isting dataset of real images.
****

Lei Yang
Discrete Distribution Networks
We introduce a novel generative model, the Discrete Distribution Networks (DDN),
 that approximates data distribution using hierarchical discrete distributions.
We posit that since the features within a network inherently capture distributio
nal information, enabling the network to generate multiple samples simultaneousl
y, rather than a single output, may offer an effective way to represent distribu
tions. Therefore, DDN fits the target distribution, including continuous ones, b
y generating multiple discrete sample points. To capture finer details of the ta
rget data, DDN selects the output that is closest to the Ground Truth (GT) from
the coarse results generated in the first layer. This selected output is then fe
d back into the network as a condition for the second layer, thereby generating
new outputs more similar to the GT. As the number of DDN layers increases, the r
epresentational space of the outputs expands exponentially, and the generated sa
mples become increasingly similar to the GT. This hierarchical output pattern of
 discrete distributions endows DDN with unique properties: more general zero-sho
t conditional generation and 1D latent representation. We demonstrate the effica
cy of DDN and its intriguing properties through experiments on CIFAR-10 and FFHQ
. The code is available at https://discrete-distribution-networks.github.io/
****

Zeyi Liao,Lingbo Mo,Chejian Xu,Mintong Kang,Jiawei Zhang,Chaowei Xiao,Yuan Tian,
Bo Li,Huan Sun
EIA: ENVIRONMENTAL INJECTION ATTACK ON GENERALIST WEB AGENTS FOR PRIVACY LEAKAGE
Recently, generalist web agents have demonstrated remarkable potential in autono
mously completing a wide range of tasks on real websites, significantly boosting
 human productivity. However, web tasks, such as booking flights, usually involv
e users' personally identifiable information (PII), which may be exposed to pote
ntial privacy risks if web agents accidentally interact with compromised website
s—a scenario that remains largely unexplored in the literature.
In this work, we narrow this gap by conducting the first study on the privacy ri
sks of generalist web agents in adversarial environments. First, we present a re
alistic threat model for attacks on the website, where we consider two adversari
al targets: stealing users' specific PII or the entire user request.
Then, we propose a novel attack method, termed Environmental Injection Attack (E
IA). EIA injects malicious content designed to adapt well to environments where
the agents operate and our work instantiates EIA specifically for privacy scenar
ios in web environments.
We collect 177 action steps that involve diverse PII categories on realistic web
sites from the Mind2Web dataset, and conduct experiments using one of the most c
apable generalist web agent frameworks to date. The results demonstrate that EIA
 achieves up to 70\% attack success rate (ASR) in stealing users' specific PII a
nd 16\% ASR in stealing a full user request at an action step. Additionally, by
evaluating the detectability and testing defensive system prompts, we indicate t
hat EIA is challenging to detect and mitigate.
Notably, attacks that are not well adapted for a webpage can be detected through
 careful human inspection, leading to our discussion about the trade-off between
 security and autonomy. However, extra attackers' efforts can make EIA seamlessl
y adapted, rendering such human supervision ineffective. Thus, we further discus
s the implications on defenses at the pre- and post-deployment stages of the web

sites without relying on human supervision and call for more advanced defense st
rategies.
****
Zhuoshi Pan,Qianhui Wu,Huiqiang Jiang,Xufang Luo,Hao Cheng,Dongsheng Li,Yuqing Y
ang,Chin-Yew Lin,H. Vicky Zhao,Lili Qiu,Jianfeng Gao
SeCom: On Memory Construction and Retrieval for Personalized Conversational Agen
ts
To deliver coherent and personalized experiences in long-term conversations, exi
sting approaches typically perform retrieval augmented response generation by co
nstructing memory banks from conversation history at either the turn-level, sess
ion-level, or through summarization techniques.
In this paper, we explore the impact of different memory granularities and prese
nt two key findings: (1) Both turn-level and session-level memory units are subo
ptimal, affecting not only the quality of final responses, but also the accuracy
 of the retrieval process.
(2) The redundancy in natural language introduces noise, hindering precise retri
eval. We demonstrate that *LLMLingua-2*, originally designed for prompt compress
ion to accelerate LLM inference, can serve as an effective denoising method to e
nhance memory retrieval accuracy.

Building on these insights, we propose **SeCom**, a method that constructs a mem
ory bank with topical segments by introducing a conversation **Se**gmentation mo
del, while performing memory retrieval based on **Com**pressed memory units.
Experimental results show that **SeCom** outperforms turn-level, session-level,
and several summarization-based methods on long-term conversation benchmarks suc
h as *LOCOMO* and *Long-MT-Bench+*. Additionally, the proposed conversation segm
entation method demonstrates superior performance on dialogue segmentation datas
ets such as *DialSeg711*, *TIAGE*, and *SuperDialSeg*.
****
Jiaxin Wen,Ruiqi Zhong,Akbir Khan,Ethan Perez,Jacob Steinhardt,Minlie Huang,Samu
el R. Bowman,He He,Shi Feng
Language Models Learn to Mislead Humans via RLHF
Language models (LMs) can produce errors that are hard to detect for humans, esp
ecially when the task is complex.
RLHF, the most popular post-training method, may exacerbate this problem: to ach
ieve higher rewards, LMs might get better at convincing humans that they are rig
ht even when they are wrong. We study this phenomenon under a standard RLHF pipe
line, calling it ``U-Sophistry'' since it is \textbf{U}nintended by model develo
pers. Specifically, we ask time-constrained (e.g., 3-10 minutes) human subjects
to evaluate the correctness of model outputs and calculate humans' accuracy agai
nst gold labels. On a question-answering task (QuALITY) and programming task (AP
PS), RLHF makes LMs better at convincing our subjects but not at completing the
task correctly. RLHF also makes the model harder to evaluate: our subjects' fals
e positive rate increases by 24.1% on QuALITY and 18.3% on APPS.
Finally, we show that probing, a state-of-the-art approach for detecting \textbf
{I}ntended Sophistry (e.g.~backdoored LMs), does not generalize to U-Sophistry.
Our results highlight an important failure mode of RLHF and call for more resear
ch in assisting humans to align them.
****
Xinyue Hu,Zhibin Duan,Bo Chen,Mingyuan Zhou
Enhancing Uncertainty Estimation and Interpretability with Bayesian Non-negative
 Decision Layer
Although deep neural networks have demonstrated significant success due to their
powerful expressiveness, most models struggle to meet practical requirements for
uncertainty estimation. Concurrently, the entangled nature of deep neural net-
works leads to a multifaceted problem, where various localized explanation tech-
niques reveal that multiple unrelated features influence the decisions, thereby
un-
dermining interpretability. To address these challenges, we develop a Bayesian
Nonnegative Decision Layer (BNDL), which reformulates deep neural networks

as a conditional Bayesian non-negative factor analysis. By leveraging stochastic latent variables, the BNDL can model complex dependencies and provide robust uncertainty estimation. Moreover, the sparsity and non-negativity of the latent variables encourage the model to learn disentangled representations and decision layers, thereby improving interpretability. We also offer theoretical guarantees that BNDL can achieve effective disentangled learning. In addition, we developed a corresponding variational inference method utilizing a Weibull variational inference network to approximate the posterior distribution of the latent variables.
Our experimental results demonstrate that with enhanced disentanglement capabilities, BNDL not only improves the model’s accuracy but also provides reliable uncertainty estimation and improved interpretability.
****
Liang Wang,Shaozhen Liu,Yu Rong,Deli Zhao,Qiang Liu,Shu Wu,Liang Wang
MolSpectra: Pre-training 3D Molecular Representation with Multi-modal Energy Spectra
Establishing the relationship between 3D structures and the energy states of molecular systems has proven to be a promising approach for learning 3D molecular representations. However, existing methods are limited to modeling the molecular energy states from classical mechanics. This limitation results in a significant oversight of quantum mechanical effects, such as quantized (discrete) energy level structures, which offer a more accurate estimation of molecular energy and can be experimentally measured through energy spectra. In this paper, we propose to utilize the energy spectra to enhance the pre-training of 3D molecular representations (MolSpectra), thereby infusing the knowledge of quantum mechanics into the molecular representations. Specifically, we propose SpecFormer, a multi-spectrum encoder for encoding molecular spectra via masked patch reconstruction. By further aligning outputs from the 3D encoder and spectrum encoder using a contrastive objective, we enhance the 3D encoder's understanding of molecules. Evaluations on public benchmarks reveal that our pre-trained representations surpass existing methods in predicting molecular properties and modeling dynamics.
****
Shuchen Wu,Mirko Thalmann,Peter Dayan,Zeynep Akata,Eric Schulz
Building, Reusing, and Generalizing Abstract Representations from Concrete Sequences
Humans excel at learning abstract patterns across different sequences, filtering out
irrelevant details, and transferring these generalized concepts to new sequences.
In contrast, many sequence learning models lack the ability to abstract, which leads to memory inefficiency and poor transfer. We introduce a non-parametric hierarchical variable learning model (HVM) that learns chunks from sequences and abstracts contextually similar chunks as variables. HVM efficiently organizes memory while uncovering abstractions, leading to compact sequence representations.
When learning on language datasets such as babyLM, HVM learns a more efficient dictionary than standard compression algorithms such as Lempel-Ziv. In a sequence recall task requiring the acquisition and transfer of variables embedded in sequences,
we demonstrate HVM’s sequence likelihood correlates with human recall times. In contrast, large language models (LLMs) struggle to transfer abstract variables as effectively as humans. From HVM’s adjustable layer of abstraction, we demonstrate that the model realizes a precise trade-off between compression and generalization.
Our work offers a cognitive model that captures the learning and transfer of abstract

representations in human cognition and differentiates itself from LLMs.
****
Margaret Li,Sneha Kudugunta,Luke Zettlemoyer
(Mis)Fitting Scaling Laws: A Survey of Scaling Law Fitting Techniques in Deep Le
arning
Modern foundation models rely heavily on using scaling laws to guide crucial tra
ining decisions. Researchers often extrapolate the optimal architecture and hype
r parameters settings from smaller training runs by describing the relationship
between, loss, or task performance, and scale. All components of this process va
ry, from the specific equation being fit, to the training setup, to the optimiza
tion method. Each of these factors may affect the fitted law, and therefore, the
 conclusions of a given study. We discuss discrepancies in the conclusions that
several prior works reach, on questions such as the optimal token to parameter r
atio. We augment this discussion with our own analysis of the critical impact th
at changes in specific details may effect in a scaling study, and the resulting
altered conclusions. Additionally, we survey over 50 papers that study scaling t
rends: while 45 of these papers quantify these trends using a power law, most un
der-report crucial details needed to reproduce their findings. To mitigate this,
 we we propose a checklist for authors to consider while contributing to scaling
 law research.
****
Michal Lukasik,Zhao Meng,Harikrishna Narasimhan,Yin-Wen Chang,Aditya Krishna Men
on,Felix Yu,Sanjiv Kumar
Better autoregressive regression with LLMs via regression-aware fine-tuning
Decoder-based large language models (LLMs) have proven highly versatile, with re
markable successes even on problems ostensibly removed from traditional language
 generation. One such example is solving regression problems, where the targets
 are real numbers rather than textual tokens. A common approach to use LLMs on
such problems is to perform fine-tuning based on the cross-entropy loss, and use
 autoregressive sampling at inference time. Another approach relies on fine-tuni
ng a separate predictive head with a suitable loss such as squared error. While
each approach has had success, there has been limited study on principled ways o
f using decoder LLMs for regression. In this work, we compare different prior wo
rks under a unified view, and introduce regression-aware fine-tuning(RAFT), a no
vel approach based on the Bayes-optimal decision rule. We demonstrate how RAFT i
mproves over established baselines on several benchmarks and model families.
****
Vinh Tong,Dung Trung Hoang,Anji Liu,Guy Van den Broeck,Mathias Niepert
Learning to Discretize Denoising Diffusion ODEs
Diffusion Probabilistic Models (DPMs) are generative models showing competitive
performance in various domains, including image synthesis and 3D point cloud gen
eration. Sampling from pre-trained DPMs involves multiple neural function evalua
tions (NFEs) to transform Gaussian noise samples into images, resulting in highe
r computational costs compared to single-step generative models such as GANs or
VAEs. Therefore, reducing the number of NFEs while preserving generation quality
 is crucial. To address this, we propose LD3, a lightweight framework designed t
o learn the optimal time discretization for sampling. LD3 can be combined with v
arious samplers and consistently improves generation quality without having to r
etrain resource-intensive neural networks. We demonstrate analytically and empir
ically that LD3 improves sampling efficiency with much less computational overhe
ad. We evaluate our method with extensive experiments on 7 pre-trained models, c
overing unconditional and conditional sampling in both pixel-space and latent-sp
ace DPMs. We achieve FIDs of 2.38 (10 NFE), and 2.27 (10 NFE) on unconditional C
IFAR10 and AFHQv2 in 5-10 minutes of training. LD3 offers an efficient approach
to sampling from pre-trained diffusion models. Code is available at https://gith
ub.com/vinhsuhi/LD3.
****
Grace Liu,Michael Tang,Benjamin Eysenbach
A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL
 without Rewards, Demonstrations, or Subgoals

In this paper, we present empirical evidence of skills and directed exploration emerging from a simple RL algorithm long before any successful trials are observed. For example, in a manipulation task, the agent is given a single observation of the goal state (see Fig. 1) and learns skills, first for moving its end-effector, then for pushing the block, and finally for picking up and placing the block. These skills emerge before the agent has ever successfully placed the block at the goal location and without the aid of any reward functions, demonstrations, or manually-specified distance metrics. Once the agent has learned to reach the goal state reliably, exploration is reduced. Implementing our method involves a simple modification of prior work and does not require density estimates, ensembles, or any additional hyperparameters. Intuitively, the proposed method seems like it should be terrible at exploration, and we lack a clear theoretical understanding of why it works so effectively, though our experiments provide some hints.
****
Haoyue Dai,Ignavier Ng,Jianle Sun,Zeyu Tang,Gongxu Luo,Xinshuai Dong,Peter Spirtes,Kun Zhang
When Selection Meets Intervention: Additional Complexities in Causal Discovery
We address the common yet often-overlooked selection bias in interventional studies, where subjects are selectively enrolled into experiments. For instance, participants in a drug trial are usually patients of the relevant disease; A/B tests on mobile applications target existing users only, and gene perturbation studies typically focus on specific cell types, such as cancer cells. Ignoring this bias leads to incorrect causal discovery results. Even when recognized, the existing paradigm for interventional causal discovery still fails to address it. This is because subtle differences in _when_ and _where_ interventions happen can lead to significantly different statistical patterns. We capture this dynamic by introducing a graphical model that explicitly accounts for both the observed world (where interventions are applied) and the counterfactual world (where selection occurs while interventions have not been applied). We characterize the Markov property of the model, and propose a provably sound algorithm to identify causal relations as well as selection mechanisms up to the equivalence class, from data with soft interventions and unknown targets. Through synthetic and real-world experiments, we demonstrate that our algorithm effectively identifies true causal relations despite the presence of selection bias.
****
Hyeongjun Heo,Seonghun Oh,Jae Yong Lee,Young Min Kim,Yonghyeon Lee
Isometric Regularization for Manifolds of Functional Data
While conventional data are represented as discrete vectors, Implicit Neural Representations (INRs) utilize neural networks to represent data points as continuous functions. By incorporating a shared network that maps latent vectors to individual functions, one can model the distribution of functional data, which has proven effective in many applications, such as learning 3D shapes, surface reflectance, and operators.
However, the infinite-dimensional nature of these representations makes them prone to overfitting, necessitating sufficient regularization. Naïve regularization methods -- those commonly used with discrete vector representations -- may enforce smoothness to increase robustness but result in a loss of data fidelity due to improper handling of function coordinates.
To overcome these challenges, we start by interpreting the mapping from latent variables to INRs as a parametrization of a Riemannian manifold. We then recognize that preserving geometric quantities -- such as distances and angles -- between the latent space and the data manifold is crucial. As a result, we obtain a manifold with minimal intrinsic curvature, leading to robust representations while maintaining high-quality data fitting. Our experiments on various data modalities demonstrate that our method effectively discovers a well-structured latent space, leading to robust data representations even for challenging datasets, such as those that are small or noisy.
****
Yixin Ji,Yang Xiang,Juntao Li,Qingrong Xia,Ping Li,Xinyu Duan,Zhefeng Wang,Min Z

hang
Beware of Calibration Data for Pruning Large Language Models
As large language models (LLMs) are widely applied across various fields, model compression has become increasingly crucial for reducing costs and improving inference efficiency. Post-training pruning is a promising method that does not require resource-intensive iterative training and only needs a small amount of calibration data to assess the importance of parameters. Recent research has enhanced post-training pruning from different aspects but few of them systematically explore the effects of calibration data, and it is unclear if there exist better calibration data construction strategies. We fill this blank and surprisingly observe that calibration data is also crucial to post-training pruning, especially for high sparsity. Through controlled experiments on important influence factors of calibration data, including the pruning settings, the amount of data, and its similarity with pre-training data, we observe that a small size of data is adequate, and more similar data to its pre-training stage can yield better performance. As pre-training data is usually inaccessible for advanced LLMs, we further provide a self-generating calibration data synthesis strategy to construct feasible calibration data. Experimental results on recent strong open-source LLMs (e.g., DCLM, and LLaMA-3) show that the proposed strategy can enhance the performance of strong pruning methods (e.g., Wanda, DSnoT, OWL) by a large margin (up to 2.68%).
****
Yusuke Hirota,Min-Hung Chen,Chien-Yi Wang,Yuta Nakashima,Yu-Chiang Frank Wang,Ryo Hachiuma
SANER: Annotation-free Societal Attribute Neutralizer for Debiasing CLIP
Large-scale vision-language models, such as CLIP, are known to contain societal bias regarding protected attributes (e.g., gender, age). This paper aims to address the problems of societal bias in CLIP. Although previous studies have proposed to debias societal bias through adversarial learning or test-time projecting, our comprehensive study of these works identifies two critical limitations: 1) loss of attribute information when it is explicitly disclosed in the input and 2) use of the attribute annotations during debiasing process. To mitigate societal bias in CLIP and overcome these limitations simultaneously, we introduce a simple-yet-effective debiasing method called SANER (societal attribute neutralizer) that eliminates attribute information from CLIP text features only of attribute-neutral descriptions. Experimental results show that SANER, which does not require attribute annotations and preserves original information for attribute-specific descriptions, demonstrates superior debiasing ability than the existing methods.
****
Daniel Musekamp,Marimuthu Kalimuthu,David Holzmüller,Makoto Takamoto,Mathias Niepert
Active Learning for Neural PDE Solvers
Solving partial differential equations (PDEs) is a fundamental problem in engineering and science. While neural PDE solvers can be more efficient than established numerical solvers, they often require large amounts of training data that is costly to obtain. Active learning (AL) could help surrogate models reach the same accuracy with smaller training sets by querying classical solvers with more informative initial conditions and PDE parameters. While AL is more common in other domains, it has yet to be studied extensively for neural PDE solvers. To bridge this gap, we introduce AL4PDE, a modular and extensible active learning benchmark. It provides multiple parametric PDEs and state-of-the-art surrogate models for the solver-in-the-loop setting, enabling the evaluation of existing and the development of new AL methods for PDE solving. We use the benchmark to evaluate batch active learning algorithms such as uncertainty- and feature-based methods. We show that AL reduces the average error by up to 71\% compared to random sam

pling and significantly reduces worst-case errors. Moreover, AL generates simila r datasets across repeated runs, with consistent distributions over the PDE para meters and initial conditions. The acquired datasets are reusable, providing ben efits for surrogate models not involved in the data generation.
****

Sha Guo,Jing Chen,Zixuan Hu,Zhuo Chen,Wenhan Yang,Yu Lin,Xing Jiang,LINGYU DUAN
Which Tasks Should Be Compressed Together? A Causal Discovery Approach for Effic ient Multi-Task Representation Compression
Conventional image compression methods are inadequate for intelligent analysis, as they overemphasize pixel-level precision while neglecting semantic significan ce and the interaction among multiple tasks. This paper introduces a Taskonomy-Aware Multi-Task Compression framework comprising (1) inter-coherent task groupi ng, which organizes synergistic tasks into shared representations to improve mul ti-task accuracy and reduce encoding volume, and (2) a conditional entropy-based directed acyclic graph (DAG) that captures causal dependencies among grouped re presentations. By leveraging parent representations as contextual priors for chi ld representations, the framework effectively utilizes cross-task information to improve entropy model accuracy. Experiments on diverse vision tasks, including Keypoint 2D, Depth Z-buffer, Semantic Segmentation, Surface Normal, Edge Texture , and Autoencoder, demonstrate significant bitrate-performance gains, validating the method’s capability to reduce system entropy uncertainty. These findings un derscore the potential of leveraging representation disentanglement, synergy, an d causal modeling to learn compact representations, which enable efficient multi -task compression in intelligent systems.
****

Haotian Xia,Zhengbang Yang,Junbo Zou,Rhys Tracy,Yuqing Wang,Chi Lu,Christopher L ai,Yanjun He,Xun Shao,Zhuoqing Xie,Yuan-fang Wang,Weining Shen,Hanjie Chen
SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Lang uage Models
Multimodal Large Language Models (MLLMs) are advancing the ability to reason abo ut complex sports scenarios by integrating textual and visual information. To co mprehensively evaluate their capabilities, we introduce SPORTU, a benchmark desi gned to assess MLLMs across multi-level sports reasoning tasks. SPORTU comprises two key components: SPORTU-text, featuring 900 multiple-choice questions with h uman-annotated explanations for rule comprehension and strategy understanding. T his component focuses on testing models' ability to reason about sports solely t hrough question-answering (QA), without requiring visual inputs; SPORTU-video, c onsisting of 1,701 slow-motion video clips across 7 different sports and 12,048 QA pairs, designed to assess multi-level reasoning, from simple sports recogniti on to complex tasks like foul detection and rule application. We evaluated four prevalent LLMs mainly utilizing few-shot learning paradigms supplemented by chai n-of-thought (CoT) prompting on the SPORTU-text part. GPT-4o achieves the highes t accuracy of 71\%, but still falls short of human-level performance, highlighti ng room for improvement in rule comprehension and reasoning. The evaluation for the SPORTU-video part includes 6 proprietary and 8 open-source MLLMs. Experiment s show that models fall short on hard tasks that require deep reasoning and rule -based understanding. GPT-4o performs the best with only 57.8\% accuracy on the hard task, showing large room for improvement. We hope that SPORTU will serve as a critical step toward evaluating models' capabilities in sports understanding and reasoning. The dataset is available at [https://github.com/chili-lab/SPORTU] (https://github.com/chili-lab/SPORTU).
****

Yougang Lyu,Lingyong Yan,Zihan Wang,Dawei Yin,Pengjie Ren,Maarten de Rijke,Zhaoc hun Ren
MACPO: Weak-to-Strong Alignment via Multi-Agent Contrastive Preference Optimizat ion
As large language models (LLMs) are rapidly advancing and achieving near-human c apabilities on specific tasks, aligning them with human values is becoming more urgent. In scenarios where LLMs outperform humans, we face a weak-to-strong alig nment problem where we need to effectively align strong student LLMs through wea

k supervision generated by weak teachers. Existing alignment methods mainly focu
s on strong-to-weak alignment and self-alignment settings, and it is impractical
 to adapt them to the much harder weak-to-strong alignment setting. To fill this
 gap, we propose a multi-agent contrastive preference optimization (MACPO) frame
work. MACPO facilitates weak teachers and strong students to learn from each oth
er by iteratively reinforcing unfamiliar positive behaviors while penalizing fam
iliar negative ones. To get this, we devise a mutual positive behavior augmentat
ion strategy to encourage weak teachers and strong students to learn from each o
ther's positive behavior and further provide higher quality positive behavior fo
r the next iteration. Additionally, we propose a hard negative behavior construc
tion strategy to induce weak teachers and strong students to generate familiar n
egative behavior by fine-tuning on negative behavioral data. Experimental result
s on the HH-RLHF and PKU-SafeRLHF datasets, evaluated using both automatic metri
cs and human judgments, demonstrate that MACPO simultaneously improves the align
ment performance of strong students and weak teachers. Moreover, as the number o
f weak teachers increases, MACPO achieves better weak-to-strong alignment perfor
mance through more iteration optimization rounds.
****
Xinyang Liu,Yilin He,Bo Chen,Mingyuan Zhou
Advancing Graph Generation through Beta Diffusion
Diffusion models have excelled in generating natural images and are now being ad
apted to a variety of data types, including graphs. However, conventional models
 often rely on Gaussian or categorical diffusion processes, which can struggle t
o accommodate the mixed discrete and continuous components characteristic of gra
ph data. Graphs typically feature discrete structures and continuous node attrib
utes that often exhibit rich statistical patterns, including sparsity,
bounded ranges, skewed distributions, and long-tailed behavior. To address these
 challenges, we introduce Graph Beta Diffusion (GBD), a generative model specifi
cally designed to handle the diverse nature of graph data. GBD leverages a beta
diffusion process, effectively modeling both continuous and discrete elements. A
dditionally, we propose a modulation technique that enhances the realism of gene
rated graphs by stabilizing critical graph topology while maintaining flexibilit
y for other components. GBD competes strongly with existing models across multip
le general and biochemical graph benchmarks, showcasing its ability to capture t
he intricate balance between discrete and continuous features inherent in real-w
orld graph data. Our PyTorch code is available at https://github.com/xinyangATK/
GraphBetaDiffusion.
****
Paidamoyo Chapfuwa,Ilker Demirel,Lorenzo Pisani,Javier Zazo,Elon Portugaly,H. Ja
bran Zahid,Julia Greissl
Scalable Universal T-Cell Receptor Embeddings from Adaptive Immune Repertoires
T cells are a key component of the adaptive immune system, targeting infections,
 cancers, and allergens with specificity encoded by their T cell receptors (TCRs
), and retaining a memory of their targets. High-throughput TCR repertoire seque
ncing captures a cross-section of TCRs that encode the immune history of any sub
ject, though the data are heterogeneous, high dimensional, sparse, and mostly un
labeled.
Sets of TCRs responding to the same antigen, *i.e.*, a protein fragment, co-occu
r in subjects sharing immune genetics and exposure history. Here, we leverage TC
R co-occurrence across a large set of TCR repertoires and employ the GloVe (Penn
ington et al., 2014) algorithm to derive low-dimensional, dense vector represen
tations (embeddings) of TCRs. We then aggregate these TCR embeddings to generate
 subject-level embeddings based on observed *subject-specific* TCR subsets. Furt
her, we leverage random projection theory to improve GloVe's computational effic
iency in terms of memory usage and training time. Extensive experimental results
 show that TCR embeddings targeting the same pathogen have high cosine similarit
y, and subject-level embeddings encode both immune genetics and pathogenic expos
ure history.
****
Zihan Ye,Shreyank N Gowda,Shiming Chen,Xiaowei Huang,Haotian Xu,Fahad Shahbaz Kh

an,Yaochu Jin,Kaizhu Huang,Xiaobo Jin
ZeroDiff: Solidified Visual-semantic Correlation in Zero-Shot Learning
Zero-shot Learning (ZSL) aims to enable classifiers to identify unseen classes. This is typically achieved by generating visual features for unseen classes base d on learned visual-semantic correlations from seen classes. However, most curre nt generative approaches heavily rely on having a sufficient number of samples f rom seen classes. Our study reveals that a scarcity of seen class samples result s in a marked decrease in performance across many generative ZSL techniques. We argue, quantify, and empirically demonstrate that this decline is largely attri butable to spurious visual-semantic correlations. To address this issue, we intr oduce ZeroDiff, an innovative generative framework for ZSL that incorporates dif fusion mechanisms and contrastive representations to enhance visual-semantic cor relations. ZeroDiff comprises three key components: (1) Diffusion augmentation, which naturally transforms limited data into an expanded set of noised data to m itigate generative model overfitting; (2) Supervised-contrastive (SC)-based repr esentations that dynamically characterize each limited sample to support visual feature generation; and (3) Multiple feature discriminators employing a Wasserst ein-distance-based mutual learning approach, evaluating generated features from various perspectives, including pre-defined semantics, SC-based representations, and the diffusion process. Extensive experiments on three popular ZSL benchmark s demonstrate that ZeroDiff not only achieves significant improvements over exis ting ZSL methods but also maintains robust performance even with scarce training data. Our codes are available at https://github.com/FouriYe/ZeroDiff_ICLR25.
****
Chenxu Wu,Qingpeng Kong,Zihang Jiang,S Kevin Zhou
Self-Supervised Diffusion MRI Denoising via Iterative and Stable Refinement
Magnetic Resonance Imaging (MRI), including diffusion MRI (dMRI), serves as a ``microscope'' for anatomical structures and routinely mitigates the influence of low signal-to-noise ratio scans by compromising temporal or spatial resolution. However, these compromises fail to meet clinical demands for both efficiency and precision. Consequently, denoising is a vital preprocessing step, particularly for dMRI, where clean data is unavailable. In this paper, we introduce Di-Fusion , a fully self-supervised denoising method that leverages the latter diffusion s teps and an adaptive sampling process. Unlike previous approaches, our single-st age framework achieves efficient and stable training without extra noise model t raining and offers adaptive and controllable results in the sampling process. Ou r thorough experiments on real and simulated data demonstrate that Di-Fusion ach ieves state-of-the-art performance in microstructure modeling, tractography trac king, and other downstream tasks. Code is available at https://github.com/Fouier L/Di-Fusion.
****
Yunhan Zhao,Xiang Zheng,Lin Luo,Yige Li,Xingjun Ma,Yu-Gang Jiang
BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks
In this paper, we focus on black-box defense for VLMs against jailbreak attacks. Existing black-box defense methods are either unimodal or bimodal. Unimodal meth ods enhance either the vision or language module of the VLM, while bimodal metho ds robustify the model through text-image representation realignment.
However, these methods suffer from two limitations: 1) they fail to fully exploi t the cross-modal information, or 2) they degrade the model performance on benig n inputs.
To address these limitations, we propose a novel blue-team method BlueSuffix tha t defends target VLMs against jailbreak attacks without compromising its perform ance under black-box setting. BlueSuffix includes three key components: 1) a vis ual purifier against jailbreak images, 2) a textual purifier against jailbreak t exts, and 3) a blue-team suffix generator using reinforcement fine-tuning for en hancing cross-modal robustness. We empirically show on four VLMs (LLaVA, MiniGPT -4, InstructionBLIP, and Gemini) and four safety benchmarks (Harmful Instruction , AdvBench, MM-SafetyBench, and RedTeam-2K) that BlueSuffix outperforms the base line defenses by a significant margin.

Our BlueSuffix opens up a promising direction for defending VLMs against jailbre
ak attacks. Code is available at https://github.com/Vinsonzyh/BlueSuffix.
****

Enea Monzio Compagnoni,Tianlin Liu,Rustem Islamov,Frank Norbert Proske,Antonio O
rvieto,Aurelien Lucchi

Adaptive Methods through the Lens of SDEs: Theoretical Insights on the Role of N
oise

Despite the vast empirical evidence supporting the efficacy of adaptive optimiza
tion methods in deep learning, their theoretical understanding is far from compl
ete. This work introduces novel SDEs for commonly used adaptive optimizers: Sign
SGD, RMSprop(W), and Adam(W). These SDEs offer a quantitatively accurate descrip
tion of these optimizers and help illuminate an intricate relationship between a
daptivity, gradient noise, and curvature. Our novel analysis of SignSGD highligh
ts a noteworthy and precise contrast to SGD in terms of convergence speed, stati
onary distribution, and robustness to heavy-tail noise. We extend this analysis
to AdamW and RMSpropW, for which we observe that the role of noise is much more
complex. Crucially, we support our theoretical analysis with experimental eviden
ce by verifying our insights: this includes numerically integrating our SDEs usi
ng Euler-Maruyama discretization on various neural network architectures such as
 MLPs, CNNs, ResNets, and Transformers. Our SDEs accurately track the behavior o
f the respective optimizers, especially when compared to previous SDEs derived f
or Adam and RMSprop. We believe our approach can provide valuable insights into
best training practices and novel scaling rules.
****

Rylan Schaeffer,Dan Valentine,Luke Bailey,James Chua,Cristobal Eyzaguirre,Zane D
urante,Joe Benton,Brando Miranda,Henry Sleight,Tony Tong Wang,John Hughes,Rajash
ree Agrawal,Mrinank Sharma,Scott Emmons,Sanmi Koyejo,Ethan Perez

Failures to Find Transferable Image Jailbreaks Between Vision-Language Models

The integration of new modalities into frontier AI systems offers exciting capab
ilities, but also increases the possibility such systems can be adversarially ma
nipulated in undesirable ways.

In this work, we focus on a popular class of vision-language models (VLMs) that
generate text outputs conditioned on visual and textual inputs.

We conducted a large-scale empirical study to assess the transferability of grad
ient-based universal image "jailbreaks" using a diverse set of over 40 open-para
meter VLMs, including 18 new VLMs that we publicly release.

Overall, we find that transferable gradient-based image jailbreaks are extremely
 difficult to obtain.

When an image jailbreak is optimized against a single VLM or against an ensemble
 of VLMs, the jailbreak successfully jailbreaks the attacked VLM(s), but exhibi
ts little-to-no transfer to any other VLMs; transfer is not affected by whether
the attacked and target VLMs possess matching vision backbones or language model
s, whether the language model underwent instruction-following and/or safety-alig
nment training, or many other factors.

Only two settings display partially successful transfer: between identically-pre
trained and identically-initialized VLMs with slightly different VLM training da
ta, and between different training checkpoints of a single VLM.

Leveraging these results, we then demonstrate that transfer can be significantly
 improved against a specific target VLM by attacking larger ensembles of "highly
-similar" VLMs.

These results stand in stark contrast to existing evidence of universal and tran
sferable text jailbreaks against language models and transferable adversarial at
tacks against image classifiers, suggesting that VLMs may be more robust to grad
ient-based transfer attacks.
****

Guoxin Chen,Zhong Zhang,Xin Cong,Fangda Guo,Yesai Wu,Yankai Lin,Wenzheng Feng,Ya
sheng Wang

Learning Evolving Tools for Large Language Models

Tool learning enables large language models (LLMs) to interact with external too
ls and APIs, greatly expanding the application scope of LLMs. However, due to th

e dynamic nature of external environments, these tools and APIs may become outda
ted over time, preventing LLMs from correctly invoking tools. Existing research
primarily focuses on static environments and overlooks this issue, limiting the
adaptability of LLMs in real-world applications. In this paper, we propose ToolE
VO, a novel framework designed to enhance the adaptive and reflective capabiliti
es of LLMs against tool variability. By leveraging Monte Carlo Tree Search, Tool
EVO facilitates active exploration and interaction of LLMs within dynamic enviro
nments, allowing for autonomous self-reflection and self-updating of tool usage
based on environmental feedback. Additionally, we introduce ToolQA-D, a benchmar
k specifically designed to evaluate the impact of tool variability. Extensive ex
periments demonstrate the effectiveness and stability of our approach, highlight
ing the importance of adaptability to tool variability for effective tool learni
ng.
****
Seung Hyun Cheon,Anneke Wernerfelt,Sorelle Friedler,Berk Ustun
Feature Responsiveness Scores: Model-Agnostic Explanations for Recourse
Machine learning models routinely automate decisions in applications like lendin
g and hiring. In such settings, consumer protection rules require companies that
 deploy models to explain predictions to decision subjects. These rules are moti
vated, in part, by the belief that explanations can promote *recourse* by reveal
ing information that individuals can use to contest or improve their outcomes. I
n practice, many companies comply with these rules by providing individuals with
 a list of the most important features for their prediction, which they identify
 based on feature importance scores from feature attribution methods such as SHA
P or LIME. In this work, we show how these practices can undermine consumers by
highlighting features that would not lead to an improved outcome and by explaini
ng predictions that cannot be changed. We propose to address these issues by hig
hlighting features based on their *responsiveness score*—i.e., the probability t
hat an individual can attain a target prediction by changing a specific feature.
 We develop efficient methods to compute responsiveness scores for any model and
 any dataset. We conduct an extensive empirical study on the responsiveness of e
xplanations in lending. Our results show that standard practices in consumer fin
ance can backfire by presenting consumers with *reasons without recourse*, and d
emonstrate how our approach improves consumer protection by highlighting respons
ive features and identifying fixed predictions.
****
Stefan Sylvius Wagner,Maike Behrendt,Marc Ziegele,Stefan Harmeling
The Power of LLM-Generated Synthetic Data for Stance Detection in Online Politic
al Discussions
Stance detection holds great potential to improve online political discussions t
hrough its deployment in discussion platforms for purposes such as content moder
ation, topic summarisation or to facilitate more balanced discussions. Typically
, transformer-based models are employed directly for stance detection, requiring
 vast amounts of data. However, the wide variety of debate topics in online poli
tical discussions makes data collection particularly challenging. LLMs have revi
ved stance detection, but their online deployment in online political discussion
s faces challenges like inconsistent outputs, biases, and vulnerability to adver
sarial attacks. We show how LLM-generated synthetic data can improve stance dete
ction for online political discussions by using reliable traditional stance dete
ction models for online deployment, while leveraging the text generation capabil
ities of LLMs for synthetic data generation in a secure offline environment. To
achieve this, (i) we generate synthetic data for specific debate questions by pr
ompting a Mistral-7B model and show that fine-tuning with the generated syntheti
c data can substantially improve the performance of stance detection, while rema
ining interpretable and aligned with real world data. (ii) Using the synthetic d
ata as a reference, we can improve performance even further by identifying the m
ost informative samples in an unlabelled dataset, i.e., those samples which the
stance detection model is most uncertain about and can benefit from the most. By
 fine-tuning with both synthetic data and the most informative samples, we surpa
ss the performance of the baseline model that is fine-tuned on all true labels,

while labelling considerably less data.
****
Liang Chen,Sinan Tan,Zefan Cai,Weichu Xie,Haozhe Zhao,Yichi Zhang,Junyang Lin,Jinze Bai,Tianyu Liu,Baobao Chang
A Spark of Vision-Language Intelligence: 2-Dimensional Autoregressive Transformer for Efficient Finegrained Image Generation
This work tackles the information loss bottleneck of vector-quantization (VQ) autoregressive image generation by introducing a novel model architecture called the 2-Dimensional Autoregression (DnD) Transformer. The DnD-Transformer predicts more codes for an image by introducing a new direction, **model depth**, along with the sequence length. Compared to 1D autoregression and previous work using similar 2D image decomposition such as RQ-Transformer, the DnD-Transformer is an end-to-end model that can generate higher quality images with the same backbone model size and sequence length, opening a new optimization perspective for autoregressive image generation. Furthermore, our experiments reveal that the DnD-Transformer's potential extends beyond generating natural images. It can even generate images with rich text and graphical elements in a self-supervised manner, demonstrating an understanding of these combined modalities. This has not been previously demonstrated for popular vision generative models such as diffusion models, showing a spark of vision-language intelligence when trained solely on images. Code, datasets and models are open at https://github.com/chenllliang/DnD-Transformer.
****
Shuyi Jiang,Qihao Zhao,Hossein Rahmani,De Wen Soh,Jun Liu,Na Zhao
GaussianBlock: Building Part-Aware Compositional and Editable 3D Scene by Primitives and Gaussians
Recently, with the development of Neural Radiance Fields and Gaussian Splatting, 3D reconstruction techniques have achieved remarkably high fidelity. However, the latent representations learnt by these methods are highly entangled and lack interpretability. In this paper, we propose a novel part-aware compositional reconstruction method, called GaussianBlock, that enables semantically coherent and disentangled representations, allowing for precise and physical editing akin to building blocks, while simultaneously maintaining high fidelity.
Our GaussianBlock introduces a hybrid representation that leverages the advantages of both primitives, known for their flexible actionability and editability, and 3D Gaussians, which excel in reconstruction quality. Specifically, we achieve semantically coherent primitives through a novel attention-guided centering loss derived from 2D semantic priors, complemented by a dynamic splitting and fusion strategy.
Furthermore, we utilize 3D Gaussians that hybridize with primitives to refine structural details and enhance fidelity.
Additionally, a binding inheritance strategy is employed to strengthen and maintain the connection between the two.
Our reconstructed scenes are evidenced to be disentangled, compositional, and compact across diverse benchmarks, enabling seamless, direct and precise editing while maintaining high quality.
****
Alessandro Stolfo,Vidhisha Balachandran,Safoora Yousefi,Eric Horvitz,Besmira Nushi
Improving Instruction-Following in Language Models through Activation Steering
The ability to follow instructions is crucial for numerous real-world applications of language models. In pursuit of deeper insights and more powerful capabilities, we derive instruction-specific vector representations from language models and use them to steer models accordingly. These vectors are computed as the difference in activations between inputs with and without instructions, enabling a modular approach to activation steering. We demonstrate how this method can enhance model adherence to constraints such as output format, length, and word inclusion, providing inference-time control over instruction following. Our experiments across four models demonstrate how we can use the activation vectors to guide models to follow constraints even without explicit instructions and to enhance p

erformance when instructions are present. Additionally, we explore the compositi
onality of activation steering, successfully applying multiple instructions simu
ltaneously. Finally, we demonstrate that steering vectors computed on instructio
n-tuned models can transfer to improve base models. Our findings demonstrate tha
t activation steering offers a practical and scalable approach for fine-grained
control in language generation. Our code and data are available at https://githu
b.com/microsoft/llm-steer-instruct.
****
Zhenfang Chen,Delin Chen,Rui Sun,Wenjun Liu,Chuang Gan
Scaling Autonomous Agents via Automatic Reward Modeling And Planning
Large language models (LLMs) have demonstrated remarkable capabilities across a
range of text-generation tasks. However, LLMs still struggle with problems requi
ring multi-step decision-making and environmental feedback, such as online shopp
ing, scientific reasoning, and mathematical problem-solving. Unlike pure text da
ta, collecting large-scale decision-making data is challenging. Moreover, many p
owerful LLMs are only accessible through APIs, which hinders their fine-tuning f
or agent tasks due to cost and complexity. To address LLM agents' limitations, w
e propose a framework that can automatically learn a reward model from the envir
onment without human annotations. This model can be used to evaluate the action
trajectories of LLM agents and provide heuristics for task planning. Specificall
y, our approach involves employing one LLM-based agent to navigate an environmen
t randomly, generating diverse action trajectories. Subsequently, a separate LLM
 is leveraged to assign a task intent and synthesize a negative response alongsi
de the correct response for each trajectory. These triplets (task intent, positi
ve response, and negative response) are then utilized as training data to optimi
ze a reward model capable of scoring action trajectories. This reward model can
be integrated with LLM-based agents and various planning algorithms to enhance t
ask-solving performance. The effectiveness and generalizability of our framework
 are demonstrated through evaluations conducted on different agent benchmarks. I
n conclusion, our proposed framework represents a significant advancement in enh
ancing LLM agents' decision-making capabilities. By automating the learning of r
eward models, we overcome the challenges of data scarcity and API limitations, p
otentially revolutionizing the application of LLMs in complex and interactive en
vironments. This research paves the way for more sophisticated AI agents capable
 of tackling a wide range of real-world problems requiring multi-step decision-m
aking.
****
Itamar Zimerman,Ameen Ali Ali,Lior Wolf
Explaining Modern Gated-Linear RNNs via a Unified Implicit Attention Formulation
Recent advances in efficient sequence modeling have led to attention-free layers
, such as Mamba, RWKV, and various gated RNNs, all featuring sub-quadratic compl
exity in sequence length and excellent scaling properties, enabling the construc
tion of a new type of foundation models. In this paper, we present a unified vie
w of these models, formulating such layers as implicit causal self-attention lay
ers. The formulation includes most of their sub-components and is not limited to
 a specific part of the architecture. The framework compares the underlying mech
anisms on similar grounds for different layers and provides a direct means for a
pplying explainability methods. Our experiments show that our attention matrices
 and attribution method outperform an alternative and a more limited formulation
 that was recently proposed for Mamba. For the other architectures for which our
 method is the first to provide such a view, our method is effective and competi
tive in the relevant metrics compared to the results obtained by state-of-the-ar
t Transformer explainability methods. Our code is publicly available.
****
Guangyi Wang,Yuren Cai,lijiang Li,Wei Peng,Song-Zhi Su
PFDiff: Training-Free Acceleration of Diffusion Models Combining Past and Future
 Scores
Diffusion Probabilistic Models (DPMs) have shown remarkable potential in image g
eneration, but their sampling efficiency is hindered by the need for numerous de
noising steps. Most existing solutions accelerate the sampling process by propos

ing fast ODE solvers. However, the inevitable discretization errors of the ODE s
olvers are significantly magnified when the number of function evaluations (NFE)
 is fewer. In this work, we propose PFDiff, a novel training-free and orthogonal
 timestep-skipping strategy, which enables existing fast ODE solvers to operate
with fewer NFE. Specifically, PFDiff initially utilizes score replacement from p
ast time steps to predict a springboard. Subsequently, it employs this ``springb
oard" along with foresight updates inspired by Nesterov momentum to rapidly upda
te current intermediate states. This approach effectively reduces unnecessary NF
E while correcting for discretization errors inherent in first-order ODE solvers
. Experimental results demonstrate that PFDiff exhibits flexible applicability a
cross various pre-trained DPMs, particularly excelling in conditional DPMs and s
urpassing previous state-of-the-art training-free methods. For instance, using D
DIM as a baseline, we achieved 16.46 FID (4 NFE) compared to 138.81 FID with DDI
M on ImageNet 64x64 with classifier guidance, and 13.06 FID (10 NFE) on Stable D
iffusion with 7.5 guidance scale. Code is available at https://github.com/onefly
123/PFDiff.
****
Gideon Stein,Maha Shadaydeh,Jan Blunk,Niklas Penzel,Joachim Denzler
CausalRivers - Scaling up benchmarking of causal discovery for real-world time-s
eries
Causal discovery, or identifying causal relationships from observational data, i
s a notoriously challenging task, with numerous methods proposed to tackle it.
Despite this, in-the-wild evaluation of these methods is still lacking, as works
 frequently rely on synthetic data evaluation and sparse real-world examples und
er critical theoretical assumptions.
Real-world causal structures, however, are often complex, evolving over time, no
n-linear, and influenced by unobserved factors, making
it hard to decide on a proper causal discovery strategy.
To bridge this gap, we introduce CausalRivers, the largest in-the-wild causal di
scovery benchmarking kit for time-series data to date.
CausalRivers features an extensive dataset on river discharge that covers the ea
stern German territory (666 measurement stations) and the state of Bavaria (494
 measurement stations).
It spans the years 2019 to 2023 with a 15-minute temporal resolution.
Further, we provide additional data from a flood around the Elbe River, as an ev
ent with a pronounced distributional shift.
Leveraging multiple sources of information and time-series meta-data, we constru
cted two distinct causal ground truth graphs (Bavaria and eastern Germany).
These graphs can be sampled to generate thousands of subgraphs to benchmark caus
al discovery across diverse and challenging settings.
To demonstrate the utility of CausalRivers, we evaluate several causal discovery
 approaches through a set of experiments to identify areas for improvement.
CausalRivers has the potential to facilitate robust evaluations and comparisons
of causal discovery methods.
Besides this primary purpose, we also expect that this dataset will be relevant
for connected areas of research, such as time-series forecasting and anomaly det
ection.
Based on this, we hope to push benchmark-driven method development that fosters
advanced techniques for causal discovery, as is the case for many other areas of
 machine learning.
****
Han Wang,Binbin Chen,Tieying Zhang,Baoxiang Wang
Learning to Communicate Through Implicit Communication Channels
Effective communication is an essential component in collaborative multi-agent s
ystems. Situations where explicit messaging is not feasible have been common in
human society throughout history, which motivate the study of implicit communica
tion. Previous works on learning implicit communication mostly rely on theory of
 mind (ToM), where agents infer the mental states and intentions of others by in
terpreting their actions. However, ToM-based methods become less effective in ma
king accurate inferences in complex tasks. In this work, we propose the Implicit

Channel Protocol (ICP) framework, which allows agents to communicate through implicit communication channels similar to the explicit ones. ICP leverages a subset of actions, denoted as the scouting actions, and a mapping between information and these scouting actions that encodes and decodes the messages. We propose training algorithms for agents to message and act, including learning with a randomly initialized information map and with a delayed information map. The efficacy of ICP has been tested on the tasks of Guessing Numbers, Revealing Goals, and Hanabi, where ICP significantly outperforms baseline methods through more efficient information transmission.
****

Fadi Khatib,Yoni Kasten,Dror Moran,Meirav Galun,Ronen Basri
RESfM: Robust Deep Equivariant Structure from Motion
Multiview Structure from Motion is a fundamental and challenging computer vision problem. A recent deep-based approach utilized matrix equivariant architectures for simultaneous recovery of camera pose and 3D scene structure from large image collections. That work, however, made the unrealistic assumption that the point tracks given as input are almost clean of outliers. Here, we propose an architecture suited to dealing with outliers by adding a multiview inlier/outlier classification module that respects the model equivariance and by utilizing a robust bundle adjustment step. Experiments demonstrate that our method can be applied successfully in realistic settings that include large image collections and point tracks extracted with common heuristics that include many outliers, achieving state-of-the-art accuracies in almost all runs, superior to existing deep-based methods and on-par with leading classical (non-deep) sequential and global methods.
****

Wentao Zhang,Junliang Guo,Tianyu He,Li Zhao,Linli Xu,Jiang Bian
Video In-context Learning: Autoregressive Transformers are Zero-Shot Video Imitators
People interact with the real-world largely dependent on visual signal, which are ubiquitous and illustrate detailed demonstrations. In this paper, we explore utilizing visual signals as a new interface for models to interact with the environment. Specifically, we choose videos as a representative visual signal. And by training autoregressive Transformers on video datasets in a self-supervised objective, we find that the model emerges a zero-shot capability to infer the semantics from a demonstration video, and imitate the semantics to an unseen scenario. This allows the models to perform unseen tasks by watching the demonstration video in an in-context manner, without further fine-tuning. To validate the imitation capacity, we design various evaluation metrics including both objective and subjective measures. The results show that our models can generate high-quality video clips that accurately align with the semantic guidance provided by the demonstration videos, and we also show that the imitation capacity follows the scaling law.
****

Annan Yu,Dongwei Lyu,Soon Hoe Lim,Michael W. Mahoney,N. Benjamin Erichson
Tuning Frequency Bias of State Space Models
State space models (SSMs) leverage linear, time-invariant (LTI) systems to effectively learn sequences with long-range dependencies. By analyzing the transfer functions of LTI systems, we find that SSMs exhibit an implicit bias toward capturing low-frequency components more effectively than high-frequency ones. This behavior aligns with the broader notion of frequency bias in deep learning model training. We show that the initialization of an SSM assigns it an innate frequency bias and that training the model in a conventional way does not alter this bias. Based on our theory, we propose two mechanisms to tune frequency bias: either by scaling the initialization to tune the inborn frequency bias; or by applying a Sobolev-norm-based filter to adjust the sensitivity of the gradients to high-frequency inputs, which allows us to change the frequency bias via training. Using an image-denoising task, we empirically show that we can strengthen, weaken, or even reverse the frequency bias using both mechanisms. By tuning the frequency bias, we can also improve SSMs' performance on learning long-range sequences,

averaging an $88.26\\%$ accuracy on the Long-Range Arena (LRA) benchmark tasks.
****
Yein Park,Chanwoong Yoon,Jungwoo Park,Donghyeon Lee,Minbyul Jeong,Jaewoo Kang
ChroKnowledge: Unveiling Chronological Knowledge of Language Models in Multiple Domains
Large language models (LLMs) have brought significant changes to many aspects of our lives.
However, assessing and ensuring their chronological knowledge remains challenging.
Existing approaches fall short in addressing the temporal adaptability of knowledge, often relying on a fixed time-point view.
To overcome this, we introduce ChroKnowBench, a benchmark dataset designed to evaluate chronologically accumulated knowledge across three key aspects: multiple domains, time dependency, temporal state.
Our benchmark distinguishes between knowledge that evolves (e.g., personal history, scientific discoveries, amended laws) and knowledge that remain constant (e.g., mathematical truths, commonsense facts).
Building on this benchmark, we present ChroKnowledge (Chronological Categorization of Knowledge), a novel sampling-based framework for evaluating LLMs' non-parametric chronological knowledge.
Our evaluation led to the following observations:
(1) The ability of eliciting temporal knowledge varies depending on the data format that model was trained on.
(2) LLMs partially recall knowledge or show a cut-off at temporal boundaries rather than recalling all aspects of knowledge correctly.
Thus, we apply our ChroKnowPrompt, an in-depth prompting to elicit chronological knowledge by traversing step-by-step through the surrounding time spans.
We observe that it successfully recalls objects across both open-source and proprietary LLMs, demonstrating versatility, though it faces challenges with dynamic datasets and unstructured formats.
****
Amirabbas Afzali,Borna khodabandeh,Ali Rasekh,Mahyar JafariNodeh,Sepehr Kazemi Ranjbar,Simon Gottschalk
Aligning Visual Contrastive learning models via Preference Optimization
Contrastive learning models have demonstrated impressive abilities to capture semantic similarities by aligning representations in the embedding space. However, their performance can be limited by the quality of the training data and its inherent biases. While Preference Optimization (PO) methods such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) have been applied to align generative models with human preferences, their use in contrastive learning has yet to be explored. This paper introduces a novel method for training contrastive learning models using different PO methods to break down complex concepts. Our method systematically aligns model behavior with desired preferences, enhancing performance on the targeted task. In particular, we focus on enhancing model robustness against typographic attacks and inductive biases, commonly seen in contrastive vision-language models like CLIP. Our experiments demonstrate that models trained using PO outperform standard contrastive learning techniques while retaining their ability to handle adversarial challenges and maintain accuracy on other downstream tasks. This makes our method well-suited for tasks requiring fairness, robustness, and alignment with specific preferences. We evaluate our method for tackling typographic attacks on images and explore its ability to disentangle gender concepts and mitigate gender bias, showcasing the versatility of our approach.
****
Qi Zhang,Peiyao Xiao,Shaofeng Zou,Kaiyi Ji
MGDA Converges under Generalized Smoothness, Provably
Multi-objective optimization (MOO) is receiving more attention in various fields such as multi-task learning. Recent works provide some effective algorithms with theoretical analysis but they are limited by the standard L-smooth or bounded-gradient assumptions, which typically do not hold for neural networks, such as

Long short-term memory (LSTM) models and Transformers. In this paper, we study a more general and realistic class of generalized ℓ-smooth loss functions, where ℓ is a general non-decreasing function of gradient norm. We revisit and analyze the fundamental multiple gradient descent algorithm (MGDA) and its stochastic version with double sampling for solving the generalized ℓ-smooth MOO problems, which approximate the conflict-avoidant (CA) direction that maximizes the minimum improvement among objectives. We provide a comprehensive convergence analysis of these algorithms and show that they converge to an ϵ-accurate Pareto stationary point with a guaranteed ϵ-level average CA distance (i.e., the gap between the updating direction and the CA direction) over all iterations, where totally $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$ samples are needed for deterministic and stochastic settings, respectively. We prove that they can also guarantee a tighter ϵ-level CA distance in each iteration using more samples. Moreover, we analyze an efficient variant of MGDA named MGDA-FA using only $\mathcal{O}(1)$ time and space, while achieving the same performance guarantee as MGDA.
****

Christopher Musco,R. Teal Witter
Provably Accurate Shapley Value Estimation via Leverage Score Sampling
Originally introduced in game theory, Shapley values have emerged as a central tool in explainable machine learning, where they are used to attribute model predictions to specific input features. However, computing Shapley values exactly is expensive: for a model with n features, $O(2^n)$ model evaluations are necessary. To address this issue, approximation algorithms are widely used. One of the most popular is the Kernel SHAP algorithm, which is model agnostic and remarkably effective in practice. However, to the best of our knowledge, Kernel SHAP has no strong non-asymptotic complexity guarantees. We address this issue by introducing *Leverage SHAP*, a light-weight modification of Kernel SHAP that provides provably accurate Shapley value estimates with just $O(n\log n)$ model evaluations. Our approach takes advantage of a connection between Shapley value estimation and agnostic active learning by employing *leverage score sampling*, a powerful regression tool. Beyond theoretical guarantees, we show that Leverage SHAP consistently outperforms even the highly optimized implementation of Kernel SHAP available in the ubiquitous SHAP library [Lundberg \& Lee, 2017].
****

Tanishq Kumar,Zachary Ankner,Benjamin Frederick Spector,Blake Bordelon,Niklas Muennighoff,Mansheej Paul,Cengiz Pehlevan,Christopher Re,Aditi Raghunathan
Scaling Laws for Precision
Low precision training and inference affect both the quality and cost of language models, but current scaling laws do not account for this. In this work, we devise "precision-aware" scaling laws for both training and inference. We propose that training in lower precision reduces the model's "effective parameter count," allowing us to predict the additional loss incurred from training in low precision and post-train quantization. For inference, we find that the degradation introduced by post-training quantization increases as models are trained on more data, eventually making additional pretraining data actively harmful. For training, our scaling laws allow us to predict the loss of a model with different parts in different precisions, and suggest that training larger models in lower precision can be compute optimal. We unify the scaling laws for post and pretraining quantization to arrive at a single functional form that predicts degradation from training and inference in varied precisions. We fit on over 465 pretraining runs and validate our predictions on model sizes up to 1.7B parameters trained on up to 26B tokens.
****

Chaofan Li,Minghao Qin,Shitao Xiao,Jianlyu Chen,Kun Luo,Defu Lian,Yingxia Shao,Zheng Liu
Making Text Embedders Few-Shot Learners
Large language models (LLMs) with decoder-only architectures have demonstrated exceptional text-generation capabilities across a variety of tasks. Some researchers have also adapted these models for text representation tasks. However, in te

xt representation tasks, these models often face performance degradation on unse
en tasks. In-context learning (ICL), which leverages examples provided in the in
put context, enables LLMs to handle unseen tasks effectively. Inspired by this,
we aim to fully utilize the inherent properties of LLMs to enhance text represen
tation performance across different tasks through the ICL approach.

In this paper, we introduce a simple yet effective training strategy, which sign
ificantly improves text representation capabilities. Unlike previous models that
 prepend task instructions to the text, our method randomly samples a varying nu
mber of examples during training, endowing the embedding model with in-context l
earning abilities while maintaining its zero-shot capabilities. This approach do
es not require additional data construction or modifications to the model archit
ecture. On the contrary, we find that some popular modifications to the model, s
uch as bidirectional attention, can degrade performance, undermining the inheren
t characteristics of LLMs. We have publicly released our method at this \href{ht
tps://github.com/FlagOpen/FlagEmbedding}{repo}.
****

Jianwen Jiang,Chao Liang,Jiaqi Yang,Gaojie Lin,Tianyun Zhong,Yanbo Zheng
Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency
With the introduction of video diffusion model, audio-conditioned human video ge
neration has recently achieved significant breakthroughs in both the naturalness
 of motion and the synthesis of portrait details. Due to the limited control of
audio signals in driving human motion, existing methods often add auxiliary spat
ial signals such as movement regions to stabilize movements, which compromise th
e naturalness and freedom of motion. To address this issue, we propose an end-to
-end audio-only conditioned video diffusion model named Loopy. Specifically, we
designed two key modules: an inter- and intra-clip temporal module and an audio-
to-latents module. These enable the model to better utilize long-term motion dep
endencies and establish a stronger audio-portrait movement correlation. Conseque
ntly, the model can generate more natural and stable portrait videos with subtle
 facial expressions, without the need for manually setting movement constraints.
 Extensive experiments show that Loopy outperforms recent audio-driven portrait
diffusion models, delivering more lifelike and high-quality results across vario
us scenarios. Video samples are available at https://loopyavataranony.github.io/
****

Bochen Lyu,Di Wang,Zhanxing Zhu
A Solvable Attention for Neural Scaling Laws
Transformers and many other deep learning models are empirically shown to predic
tably enhance their performance as a power law in training time, model size, or
the number of training data points, which is termed as the neural scaling law. T
his paper studies this intriguing phenomenon particularly for the transformer ar
chitecture in theoretical setups. Specifically, we propose a framework for linea
r self-attention, the underpinning block of transformer without softmax, to lear
n in an in-context manner, where the corresponding learning dynamics is modeled
as a non-linear ordinary differential equation (ODE) system. Furthermore, we est
ablish a procedure to derive a tractable approximate solution for this ODE syste
m by reformulating it as a *Riccati equation*, which allows us to precisely char
acterize neural scaling laws for linear self-attention with training time, model
 size, data size, and the optimal compute. In addition, we reveal that the linea
r self-attention shares similar neural scaling laws with several other architect
ures when the context sequence length of the in-context learning is fixed, other
wise it would exhibit a different scaling law of training time.
****

Beining Yang,Qingyun Sun,Cheng Ji,Xingcheng Fu,Jianxin Li
ST-GCond: Self-supervised and Transferable Graph Dataset Condensation
The increasing scale of graph datasets significantly enhances deep learning mode
ls but also presents substantial training challenges. Graph dataset condensation
 has emerged to condense large datasets into smaller yet informative ones that m
aintain similar test performance. However, these methods require downstream usag
e to match the original dataset and task, which is impractical in real-world sce

narios. Our empirical studies show that existing methods fail in "cross-task" an
d "cross-dataset" scenarios, often performing worse than training from scratch.
To address these challenges, we propose a novel method: Self-supervised and Tran
sferable Graph dataset Condensation (ST-GCond). For cross-task transferability,
we propose a task-disentangled meta optimization strategy to adaptively update t
he condensed graph according to the task relevance, encouraging information pres
ervation for various tasks. For cross-dataset transferability, we propose a mult
i-teacher self-supervised optimization strategy to incorporate auxiliary self-su
pervised tasks to inject universal knowledge into the condensed graph. Additiona
lly, we incorporate mutual information guided joint condensation mitigating the
potential conflicts and ensure the condensing stability. Experiments on both nod
e-level and graph-level datasets show that ST-GCond outperforms existing methods
 by 2.5% to 18.7% in all cross-task and cross-dataset scenarios, and also achiev
es state-of-the-art performance on 5 out of 6 datasets in the single dataset and
 task scenario.
****
Evgenii Aleksandrovich Dzhivelikian,Petr Kuderov,Aleksandr Panov
Learning Successor Features with Distributed Hebbian Temporal Memory
This paper presents a novel approach to address the challenge of online sequence
 learning for decision making under uncertainty in non-stationary, partially obs
ervable environments. The proposed algorithm, Distributed Hebbian Temporal Memor
y (DHTM), is based on the factor graph formalism and a multi-component neuron mo
del. DHTM aims to capture sequential data relationships and make cumulative pred
ictions about future observations, forming Successor Features (SFs). Inspired by
 neurophysiological models of the neocortex, the algorithm uses distributed repr
esentations, sparse transition matrices, and local Hebbian-like learning rules t
o overcome the instability and slow learning of traditional temporal memory algo
rithms such as RNN and HMM. Experimental results show that DHTM outperforms LSTM
, RWKV and a biologically inspired HMM-like algorithm, CSCG, on non-stationary d
ata sets. Our results suggest that DHTM is a promising approach to address the c
hallenges of online sequence learning and planning in dynamic environments.
****
Zihui Zhang,Yafei YANG,Hongtao Wen,Bo Yang
GrabS: Generative Embodied Agent for 3D Object Segmentation without Scene Superv
ision
We study the hard problem of 3D object segmentation in complex point clouds
without requiring human labels of 3D scenes for supervision. By relying on the
similarity of pretrained 2D features or external signals such as motion to group
 3D
points as objects, existing unsupervised methods are usually limited to identify
ing
simple objects like cars or their segmented objects are often inferior due to th
e
lack of objectness in pretrained features. In this paper, we propose a new two-
stage pipeline called GrabS. The core concept of our method is to learn generati
ve
and discriminative object-centric priors as a foundation from object datasets in
 the
first stage, and then design an embodied agent to learn to discover multiple ob-
jects by querying against the pretrained generative priors in the second stage.
We
extensively evaluate our method on two real-world datasets and a newly created
synthetic dataset, demonstrating remarkable segmentation performance, clearly
surpassing all existing unsupervised methods.
****
Christopher Ackerman,Nina Panickssery
Inspection and Control of Self-Generated-Text Recognition Ability in Llama3-8b-I
nstruct
It has been reported that LLMs can recognize their own writing. As this has pote
ntial implications for AI safety, yet is relatively understudied, we investigate

the phenomenon, seeking to establish: whether it robustly occurs at the behavio
ral level, how the observed behavior is achieved, and whether it can be controll
ed. First, we find that the Llama3-8b–Instruct chat model - but not the base Lla
ma3-8b model - can reliably distinguish its own outputs from those of humans, an
d present evidence that the chat model is likely using its experience with its o
wn outputs, acquired during post-training, to succeed at the writing recognition
 task. Second, we identify a vector in the residual stream of the model that is
differentially activated when the model makes a correct self-written-text recogn
ition judgment, show that the vector activates in response to information releva
nt to self-authorship, present evidence that the vector is related to the concep
t of ``self'' in the model, and demonstrate that the vector is causally related
to the model’s ability to perceive and assert self-authorship. Finally, we show
that the vector can be used to control both the model’s behavior and its percept
ion, steering the model to claim or disclaim authorship by applying the vector t
o the model’s output as it generates it, and steering the model to believe or di
sbelieve it wrote arbitrary texts by applying the vector to them as the model re
ads them.
****
Qian Chen,Lei Li,Qian Li,Jianghua Wu,Akang Wang,Ruoyu Sun,Xiaodong Luo,Tsung-Hui
 Chang,Qingjiang Shi
When GNNs meet symmetry in ILPs: an orbit-based feature augmentation approach
A common characteristic in integer linear programs (ILPs) is symmetry, allowing
variables to be permuted without altering the underlying problem structure. Rece
ntly, GNNs have emerged as a promising approach for solving ILPs.
However, a significant challenge arises when applying GNNs to ILPs with symmetry
: classic GNN architectures struggle to differentiate between symmetric variable
s, which limits their predictive accuracy. In this work, we investigate the prop
erties of permutation equivalence and invariance in GNNs, particularly in relati
on to the inherent symmetry of ILP formulations. We reveal that the interaction
between these two factors contributes to the difficulty of distinguishing betwee
n symmetric variables.
To address this challenge, we explore the potential of feature augmentation and
propose several guiding principles for constructing augmented features. Building
 on these principles, we develop an orbit-based augmentation scheme that first g
roups symmetric variables and then samples augmented features for each group fro
m a discrete uniform distribution. Empirical results demonstrate that our propos
ed approach significantly enhances both training efficiency and predictive perfo
rmance.
****
Mingquan Feng,Yixin Huang,Weixin Liao,Yuhong Liu,Yizhou Liu,Junchi Yan
SINGER: Stochastic Network Graph Evolving Operator for High Dimensional PDEs
We present a novel framework, StochastIc Network Graph Evolving operatoR (SINGER
), for learning the evolution operator of high-dimensional partial differential
equations (PDEs). The framework uses a sub-network to approximate the solution a
t the initial time step and stochastically evolves the sub-network parameters ov
er time by a graph neural network to approximate the solution at later time step
s. The framework is designed to inherit the desirable properties of the parametr
ic solution operator, including graph topology, semigroup, and stability, with a
 theoretical guarantee. Numerical experiments on 8 evolution PDEs of 5,10,15,20-
dimensions show that our method outperforms existing baselines in almost all cas
es (31 out of 32), and that our method generalizes well to unseen initial condit
ions, equation dimensions, sub-network width, and time steps.
****
Guoxia Wang,Jinle Zeng,Xiyuan Xiao,Siming Wu,Jiabin Yang,Lujing Zheng,Zeyu Chen,
Jiang Bian,Dianhai Yu,Haifeng Wang
FlashMask: Efficient and Rich Mask Extension of FlashAttention
The computational and memory demands of vanilla attention scale quadratically wi
th the sequence length N, posing significant challenges for processing long se
quences in Transformer models. FlashAttention alleviates these challenges by eli
minating the $\mathcal{O}(N^2)$ memory dependency and reducing attention latency

through IO-aware memory optimizations. However, its native support for certain attention mask types is limited, and it does not inherently accommodate more complex masking requirements. Previous approaches resort to using dense masks with $\mathcal{O}(N^2)$ memory complexity, leading to inefficiencies. In this paper, we propose \ours{}, an extension of FlashAttention that introduces a column-wise sparse representation of attention masks. This approach efficiently represents a wide range of mask types and facilitates the development of optimized kernel implementations. By adopting this novel representation, \ours{} achieves linear memory complexity $\mathcal{O}(N)$, making it suitable for modeling long-context sequences. Moreover, this representation enables kernel optimizations that eliminate unnecessary computations by leveraging sparsity in the attention mask, without sacrificing computational accuracy, resulting in higher computational efficiency. We evaluate \ours{}'s performance in fine-tuning and alignment training of LLMs such as SFT, LoRA, DPO, and RM. \ours{} achieves significant throughput improvements, with end-to-end speedups ranging from 1.65x to 3.22x compared to existing FlashAttention dense method. Additionally, our kernel-level comparisons demonstrate that \ours{} surpasses the latest counterpart, FlexAttention, by 12.1\% to 60.7\% in terms of kernel TFLOPs/s, achieving 37.8\% to 62.3\% of the theoretical maximum FLOPs/s on the A100 GPU. The code is open-sourced on PaddlePaddle \footnote{\url{https://github.com/PaddlePaddle/Paddle}} and integrated into PaddleNLP\footnote{\url{https://github.com/PaddlePaddle/PaddleNLP}}, supporting models with over 100 billion parameters for contexts extending up to 128K tokens.
****
Qizhou Wang,Bo Han,Puning Yang,Jianing Zhu,Tongliang Liu,Masashi Sugiyama
Towards Effective Evaluations and Comparisons for LLM Unlearning Methods
The imperative to eliminate undesirable data memorization underscores the significance of machine unlearning for large language models (LLMs). Recent research has introduced a series of promising unlearning methods, notably boosting the practical significance of the field. Nevertheless, adopting a proper evaluation framework to reflect the true unlearning efficacy is also essential yet has not received adequate attention. This paper seeks to improve the evaluation of LLM unlearning by addressing two key challenges---a) the robustness of evaluation metrics and b) the trade-offs between competing goals. The first challenge stems from findings that current metrics are susceptible to various red teaming scenarios. It indicates that they may not reflect the true extent of knowledge retained by LLMs but rather tend to mirror superficial model behaviors, thus prone to attacks. We address this issue by devising and assessing a series of candidate metrics, selecting the most robust ones under various types of attacks. The second challenge arises from the conflicting goals of eliminating unwanted knowledge while retaining those of others. This trade-off between unlearning and retention often fails to conform the Pareto frontier, rendering it subtle to compare the efficacy between methods that excel only in either unlearning or retention. We handle this issue by proposing a calibration method that can restore the original performance on non-targeted data after unlearning, thereby allowing us to focus exclusively on assessing the strength of unlearning. Our evaluation framework notably enhances the effectiveness when assessing and comparing various LLM unlearning methods, further allowing us to benchmark existing works, identify their proper hyper-parameters, and explore new tricks to enhance their practical efficacy.
****
Hongfu Liu,Hengguan Huang,Xiangming Gu,Hao Wang,Ye Wang
On Calibration of LLM-based Guard Models for Reliable Content Moderation
Large language models (LLMs) pose significant risks due to the potential for generating harmful content or users attempting to evade guardrails. Existing studies have developed LLM-based guard models designed to moderate the input and output of threat LLMs, ensuring adherence to safety policies by blocking content that violates these protocols upon deployment. However, limited attention has been given to the reliability and calibration of such guard models. In this work, we empirically conduct comprehensive investigations of confidence calibration for 9 existing LLM-based guard models on 12 benchmarks in both user input and model output classification. Our findings reveal that current LLM-based guard models ten

d to 1) produce overconfident predictions, 2) exhibit significant miscalibration when subjected to jailbreak attacks, and 3) demonstrate limited robustness to the outputs generated by different types of response models. Additionally, we assess the effectiveness of post-hoc calibration methods to mitigate miscalibration. We demonstrate the efficacy of temperature scaling and, for the first time, highlight the benefits of contextual calibration for confidence calibration of guard models, particularly in the absence of validation sets. Our analysis and experiments underscore the limitations of current LLM-based guard models and provide valuable insights for the future development of well-calibrated guard models toward more reliable content moderation. We also advocate for incorporating reliability evaluation of confidence calibration when releasing future LLM-based guard models.
****

Songtao Huang,Zhen Zhao,Can Li,LEI BAI

TimeKAN: KAN-based Frequency Decomposition Learning Architecture for Long-term Time Series Forecasting

Real-world time series often have multiple frequency components that are intertwined with each other, making accurate time series forecasting challenging. Decomposing the mixed frequency components into multiple single frequency components is a natural choice. However, the information density of patterns varies across different frequencies, and employing a uniform modeling approach for different frequency components can lead to inaccurate characterization. To address this challenges, inspired by the flexibility of the recent Kolmogorov-Arnold Network (KAN), we propose a KAN-based Frequency Decomposition Learning architecture (TimeKAN) to address the complex forecasting challenges caused by multiple frequency mixtures. Specifically, TimeKAN mainly consists of three components: Cascaded Frequency Decomposition (CFD) blocks, Multi-order KAN Representation Learning (M-KAN) blocks and Frequency Mixing blocks. CFD blocks adopt a bottom-up cascading approach to obtain series representations for each frequency band. Benefiting from the high flexibility of KAN, we design a novel M-KAN block to learn and represent specific temporal patterns within each frequency band. Finally, Frequency Mixing blocks is used to recombine the frequency bands into the original format. Extensive experimental results across multiple real-world time series datasets demonstrate that TimeKAN achieves state-of-the-art performance as an extremely lightweight architecture. Code is available at https://github.com/huangst21/TimeKAN.
****

Kunal Singh,Ankan Biswas,Sayandeep Bhowmick,Pradeep Moturi,Siva Kishore Gollapalli

SBSC: Step-by-Step Coding for Improving Mathematical Olympiad Performance

We propose Step-by-Step Coding (SBSC): a multi-turn math reasoning framework that enables Large Language Models (LLMs) to generate sequence of programs for solving Olympiad level math problems. After each turn/step, by leveraging the code execution outputs and programs of previous steps, the model generates the next sub-task and the corresponding program to complete it. This way, SBSC, sequentially navigates to reach the final answer. SBSC allows more granular, flexible and precise approach to problem-solving compared to existing methods. Extensive experiments highlight the effectiveness of SBSC in tackling competition and Olympiad-level math problems. For Claude-3.5-Sonnet, we observe SBSC (greedy decoding) surpasses existing state-of-the-art (SOTA) program generation based reasoning strategies by absolute 10.7% on AMC12, 8% on AIME and 12.6% on MathOdyssey. Given SBSC is multi-turn in nature, we also benchmark SBSC’s greedy decoding against self- consistency decoding results of existing SOTA math reasoning strategies and observe performance gain by absolute 6.2% on AMC, 6.7% on AIME and 7.4% on MathOdyssey.
****

João Bravo

NRGBoost: Energy-Based Generative Boosted Trees

Despite the rise to dominance of deep learning in unstructured data domains, tree-based methods such as Random Forests (RF) and Gradient Boosted Decision Trees (GBDT) are still the workhorses for handling discriminative tasks on tabular dat

a. We explore generative extensions of these popular algorithms with a focus on explicitly modeling the data density (up to a normalization constant), thus enabling other applications besides sampling.
As our main contribution we propose an energy-based generative boosting algorithm that is analogous to the second order boosting implemented in popular packages like XGBoost. We show that, despite producing a generative model capable of handling inference tasks over any input variable, our proposed algorithm can achieve similar discriminative performance to GBDT on a number of real world tabular datasets, outperforming alternative generative approaches. At the same time, we show that it is also competitive with neural-network-based models for sampling. Code is available at https://github.com/ajoo/nrgboost.
****
Wendi Li,Yixuan Li
Process Reward Model with Q-value Rankings
Process Reward Modeling (PRM) is critical for complex reasoning and decision-making tasks where the accuracy of intermediate steps significantly influences the overall outcome. Existing PRM approaches, primarily framed as classification problems, employ cross-entropy loss to independently evaluate each step's correctness. This method can lead to suboptimal reward distribution and does not adequately address the interdependencies among steps. To address these limitations, we introduce the Process Q-value Model (PQM), a novel framework that redefines PRM in the context of a Markov Decision Process. PQM optimizes Q-value rankings based on a novel comparative loss function, enhancing the model's ability to capture the intricate dynamics among sequential decisions. This approach provides a more granular and theoretically grounded methodology for process rewards. Our extensive empirical evaluations across various sampling policies, language model backbones, and multi-step reasoning benchmarks show that PQM outperforms classification-based PRMs. The effectiveness of the comparative loss function is highlighted in our comprehensive ablation studies, confirming PQM’s practical efficacy and theoretical advantage.
****
Abhishek Panigrahi,Bingbin Liu,Sadhika Malladi,Andrej Risteski,Surbhi Goel
Progressive distillation induces an implicit curriculum
Knowledge distillation leverages a teacher model to improve the training of a student model. A persistent challenge is that a better teacher does not always yield a better student, to which a common mitigation is to use additional supervision from several “intermediate” teachers. One empirically validated variant of this principle is progressive distillation, where the student learns from successive intermediate checkpoints of the teacher. Using sparse parity as a sandbox, we identify an implicit curriculum as one mechanism through which progressive distillation accelerates the student’s learning. This curriculum is available only through the intermediate checkpoints but not the final converged one, and imparts both empirical acceleration and a provable sample complexity benefit to the student. We then extend our investigation to Transformers trained on probabilistic context-free grammars (PCFGs) and real-world pre-training datasets (Wikipedia and Books). Through probing the teacher model, we identify an analogous implicit curriculum where the model progressively learns features that capture longer context. Our theoretical and empirical findings on sparse parity, complemented by empirical observations on more complex tasks, highlight the benefit of progressive distillation via implicit curriculum across setups.
****
Armin Toroghi,Ali Pesaranghader,Tanmana Sadhu,Scott Sanner
LLM-based Typed Hyperresolution for Commonsense Reasoning with Knowledge Bases
Large language models (LLM) are being increasingly applied to tasks requiring commonsense reasoning. Despite their outstanding potential, the reasoning process of LLMs is prone to errors and hallucinations that hinder their applicability, especially in high-stakes scenarios. Several works have attempted to enhance commonsense reasoning performance of LLMs by (i) using prompting styles that elicit more accurate reasoning, (ii) utilizing the LLM as a semantic parser for a symbolic reasoner, or (iii) enforcing the LLM to simulate a logical inference rule.

However, all these solutions have critical limitations: they are unable to lever
age the internal commonsense knowledge of the LLM in tandem with an axiomatic kn
owledge base, they lack a mechanism to reliably repair erroneous inference steps
, and their application is restricted to small knowledge bases that fit the cont
ext limit of the LLM. In this work, we present LLM-based Typed Hyperresolution (
LLM-TH), a logical commonsense reasoning framework that leverages "theory resolu
tion", a concept from classical logical inference which enables integrating LLMs
 into the "resolution" inference rule, thus mitigating reasoning errors and hall
ucinations and enabling verification of the reasoning procedure. LLM-TH is also
equipped with a mechanism for repairing erroneous inference steps supported by t
heoretical guarantees. Using "Hyperresolution" and "Typed inference" schemes, we
 show that LLM-TH can efficiently reason over large knowledge bases consisting o
f tens of thousands of rules with arbitrary predicate arities. Our experiments o
n three diverse language-based reasoning tasks—preference reasoning, multi-domai
n deductive reasoning, and geographical question answering—showcase that LLM-TH,
 using merely a BART 406M parameter NLI entailment model, significantly reduces
reasoning errors compared to baselines using Llama3-70B, Gemini1.5-Flash, GPT-3.
5-Turbo, and Mixtral-46.7B.
****

Shreyas Kapur,Erik Jenner,Stuart Russell
Diffusion On Syntax Trees For Program Synthesis
Large language models generate code one token at a time. Their autoregressive ge
neration process lacks the feedback of observing the program's output. Training
LLMs to suggest edits directly can be challenging due to the scarcity of rich ed
it data. To address these problems, we propose neural diffusion models that oper
ate on syntax trees of any context-free grammar. Similar to image diffusion mode
ls, our method also inverts "noise" applied to syntax trees. Rather than generat
ing code sequentially, we iteratively edit it while preserving syntactic validit
y, which makes it easy to combine this neural model with search. We apply our ap
proach to inverse graphics tasks, where our model learns to convert images into
programs that produce those images. Combined with search, our model is able to w
rite graphics programs, see the execution result, and debug them to meet the req
uired specifications. We additionally show how our system can write graphics pro
grams for hand-drawn sketches. Video results can be found at https://tree-diffus
ion.github.io.
****

Amirozhan Dehghani,Xinyu Qian,Asa Farahani,Pouya Bashivan
Credit-based self organizing maps: training deep topographic networks with minim
al performance degradation
In the primate neocortex, neurons with similar function are often found to be sp
atially close. Kohonen's self-organizing map (SOM) has been one of the most infl
uential approaches for simulating brain-like topographical organization in artif
icial neural network models. However, integrating these maps into deep neural ne
tworks with multitude of layers has been challenging, with self-organized deep n
eural networks suffering from substantially diminished capacity to perform visua
l recognition. We identified a key factor leading to the performance degradation
 in self-organized topographical neural network models: the discord between pred
ominantly bottom-up learning updates in the self-organizing maps, and those deri
ved from top-down, credit-based learning approaches. To address this, we propose
 an alternative self organization algorithm, tailored to align with the top-down
 learning processes in deep neural networks. This model not only emulates critic
al aspects of cortical topography but also significantly narrows the performance
 gap between non-topographical and topographical models. This advancement unders
cores the substantial importance of top-down assigned credits in shaping topogra
phical organization. Our findings are a step in reconciling topographical modeli
ng with the functional efficacy of neural network models, paving the way for mor
e brain-like neural architectures.
****

Yinan Zheng,Ruiming Liang,Kexin ZHENG,Jinliang Zheng,Liyuan Mao,Jianxiong Li,Wei
hao Gu,Rui Ai,Shengbo Eben Li,Xianyuan Zhan,Jingjing Liu

Diffusion-Based Planning for Autonomous Driving with Flexible Guidance
Achieving human-like driving behaviors in complex open-world environments is a critical challenge in autonomous driving. Contemporary learning-based planning approaches such as imitation learning methods often struggle to balance competing objectives and lack of safety assurance,due to limited adaptability and inadequacy in learning complex multi-modal behaviors commonly exhibited in human planning, not to mention their strong reliance on the fallback strategy with predefined rules. We propose a novel transformer-based Diffusion Planner for closed-loop planning, which can effectively model multi-modal driving behavior and ensure trajectory quality without any rule-based refinement. Our model supports joint modeling of both prediction and planning tasks under the same architecture, enabling cooperative behaviors between vehicles. Moreover, by learning the gradient of the trajectory score function and employing a flexible classifier guidance mechanism, Diffusion Planner effectively achieves safe and adaptable planning behaviors. Evaluations on the large-scale real-world autonomous planning benchmark nuPlan and our newly collected 200-hour delivery-vehicle driving dataset demonstrate that Diffusion Planner achieves state-of-the-art closed-loop performance with robust transferability in diverse driving styles.
****
Piotr Indyk,Michael Kapralov,Kshiteej Sheth,Tal Wagner
Improved Algorithms for Kernel Matrix-Vector Multiplication Under Sparsity Assumptions
Motivated by the problem of fast processing of attention matrices, we study fast algorithms for computing matrix-vector products for asymmetric Gaussian Kernel matrices $K\in \mathbb{R}^{n\times n}$.
K's columns are indexed by a set of n keys $k_1,k_2\ldots, k_n\in \mathbb{R}^d$, rows by a set of n queries $q_1,q_2,\ldots,q_n\in \mathbb{R}^d $, and its i,j entry is $K_{ij} = e^{-\|q_i-k_j\|_2^2/2\sigma^2}$ for some bandwidth parameter $\sigma>0$. Given a vector $x\in \mathbb{R}^n$ and error parameter $\epsilon>0$, our task is to output a $y\in \mathbb{R}^n$ such that $\|Kx-y\|_2\leq \epsilon \|x\|_2$ in time subquadratic in n and linear in d. Our algorithms rely on the following modelling assumption about the matrices K: the sum of the entries of K scales linearly in n, as opposed to worst case quadratic growth. We validate this assumption experimentally, for Gaussian kernel matrices encountered in various settings such as fast attention computation in LLMs. Under this assumption, we obtain the first subquadratic time algorithm for kernel matrix-vector multiplication for unrestricted vectors.
****
Yan Yang,Bin Gao,Ya-xiang Yuan
LancBiO: Dynamic Lanczos-aided Bilevel Optimization via Krylov Subspace
Bilevel optimization, with broad applications in machine learning, has an intricate hierarchical structure. Gradient-based methods have emerged as a common approach to large-scale bilevel problems. However, the computation of the hyper-gradient, which involves a Hessian inverse vector product, confines the efficiency and is regarded as a bottleneck. To circumvent the inverse, we construct a sequence of low-dimensional approximate Krylov subspaces with the aid of the Lanczos process. As a result, the constructed subspace is able to dynamically and incrementally approximate the Hessian inverse vector product with less effort and thus leads to a favorable estimate of the hyper-gradient. Moreover, we propose a provable subspace-based framework for bilevel problems where one central step is to solve a small-size tridiagonal linear system. To the best of our knowledge, this is the first time that subspace techniques are incorporated into bilevel optimization. This successful trial not only enjoys $\mathcal{O}(\epsilon^{-1})$ convergence rate but also demonstrates efficiency in a synthetic problem and two deep learning tasks.
****
Simla Burcu Harma,Ayan Chakraborty,Elizaveta Kostenok,Danila Mishin,Dongho Ha,Babak Falsafi,Martin Jaggi,Ming Liu,Yunho Oh,Suvinay Subramanian,Amir Yazdanbakhsh
Effective Interplay between Sparsity and Quantization: From Theory to Practice
The increasing size of deep neural networks (DNNs) necessitates effective model

compression to reduce their computational and memory footprints. Sparsity and qu
antization are two prominent compression methods that have been shown to reduce
DNNs' computational and memory footprints significantly while preserving model a
ccuracy. However, how these two methods interact when combined together remains
a key question for developers, as many tacitly assume that they are orthogonal,
meaning that their combined use does not introduce additional errors beyond thos
e introduced by each method independently. In this paper, we provide the first m
athematical proof that sparsity and quantization are non-orthogonal. We corrobor
ate these results with experiments spanning a range of large language models, in
cluding the OPT and LLaMA model families (with 125M to 8B parameters), and visio
n models like ViT and ResNet. We show that the order in which we apply these met
hods matters because applying quantization before sparsity may disrupt the relat
ive importance of tensor elements, which may inadvertently remove significant el
ements from a tensor. More importantly, we show that even if applied in the corr
ect order, the compounded errors from sparsity and quantization can significantl
y harm accuracy. Our findings extend to the efficient deployment of large models
 in resource-constrained compute platforms to reduce serving cost, offering insi
ghts into best practices for applying these compression methods to maximize hard
ware resource efficiency without compromising accuracy.
****
Beichen Li,Rundi Wu,Armando Solar-Lezama,Changxi Zheng,Liang Shi,Bernd Bickel,Wo
jciech Matusik
VLMaterial: Procedural Material Generation with Large Vision-Language Models
Procedural materials, represented as functional node graphs, are ubiquitous in c
omputer graphics for photorealistic material appearance design. They allow users
 to perform intuitive and precise editing to achieve desired visual appearances.
 However, creating a procedural material given an input image requires professio
nal knowledge and significant effort. In this work, we leverage the ability to c
onvert procedural materials into standard Python programs and fine-tune a large
pre-trained vision-language model (VLM) to generate such programs from input ima
ges. To enable effective fine-tuning, we also contribute an open-source procedur
al material dataset and propose to perform program-level augmentation by prompti
ng another pre-trained large language model (LLM). Through extensive evaluation,
 we show that our method outperforms previous methods on both synthetic and real
-world examples.
****
Jonathan Roberts,Kai Han,Samuel Albanie
Needle Threading: Can LLMs Follow Threads Through Near-Million-Scale Haystacks?
As the context limits of Large Language Models (LLMs) increase, the range of
possible applications and downstream functions broadens. In many real-world
tasks, decisions depend on details scattered across collections of often dispara
te
documents containing mostly irrelevant information. Long-context LLMs appear
well-suited to this form of complex information retrieval and reasoning, which h
as
traditionally proven costly and time-consuming. However, although the developmen
t of longer context models has seen rapid gains in recent years, our understandi
ng of how effectively LLMs use their context has not kept pace. To address
this, we conduct a set of retrieval experiments designed to evaluate the capabil
ities
of 17 leading LLMs, such as their ability to follow threads of information throu
gh
the context window. Strikingly, we find that many models are remarkably thread-
safe: capable of simultaneously following multiple threads without significant l
oss
in performance. Still, for many models, we find the effective context limit is s
ignificantly shorter than the supported context length, with accuracy decreasing
 as
the context window grows. Our study also highlights the important point that tok
en counts from different tokenizers should not be directly compared—they often

correspond to substantially different numbers of written characters. We release our code and long context experimental data.
****

Junyu Chen,Han Cai,Junsong Chen,Enze Xie,Shang Yang,Haotian Tang,Muyang Li,Song Han

Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models

We present Deep Compression Autoencoder (DC-AE), a new family of autoencoders for accelerating high-resolution diffusion models. Existing autoencodes have demonstrated impressive results at a moderate spatial compression ratio (e.g., 8x), but fail to maintain satisfactory reconstruction accuracy for high spatial compression ratios (e.g., 64x). We address this challenge by introducing two key techniques: (1) Residual Autoencoding, where we design our models to learn residuals based on the space-to-channel transformed features to alleviate the optimization difficulty of high spatial-compression autoencoders; (2) Decoupled High-Resolution Adaptation, an efficient decoupled three-phase training strategy for mitigating the generalization penalty of high spatial-compression autoencoders. With these designs, we improve the autoencoder's spatial compression ratio up to 128 while maintaining the reconstruction quality. Applying our DC-AE to latent diffusion models, we achieve significant speedup without accuracy drop. For example, on ImageNet 512x512, our DC-AE provides 19.1x inference speedup and 17.9x training speedup on H100 GPU for UViT-H while achieving a better FID, compared with the widely used SD-VAE-f8 autoencoder.
****

Yue WU,Zhaobo Qi,Yiling Wu,Junshu Sun,Yaowei Wang,Shuhui Wang

Learning Fine-Grained Representations through Textual Token Disentanglement in Composed Video Retrieval

With the explosive growth of video data, finding videos that meet detailed requirements in large datasets has become a challenge. To address this, the composed video retrieval task has been introduced, enabling users to retrieve videos using complex queries that involve both visual and textual information. However, the inherent heterogeneity between the modalities poses significant challenges. Textual data are highly abstract, while video content contains substantial redundancy. The modality gap in information representation makes existing methods struggle with the modality fusion and alignment required for fine-grained composed retrieval. To overcome these challenges, we first introduce FineCVR-1M, a fine-grained composed video retrieval dataset containing 1,010,071 video-text triplets with detailed textual descriptions. This dataset is constructed through an automated process that identifies key concept changes between video pairs to generate textual descriptions for both static and action concepts. For fine-grained retrieval methods, the key challenge lies in understanding the detailed requirements. Text description serves as clear expressions of intent, but it requires models to distinguish subtle differences in the description of video semantics. Therefore, we propose a textual Feature Disentanglement and Cross-modal Alignment framework (FDCA) that disentangles features at both the sentence and token levels. At the sequence level, we separate text features into retained and injected features. At the token level, an Auxiliary Token Disentangling mechanism is proposed to disentangle texts into retained, injected, and excluded tokens. The disentanglement at both levels extracts fine-grained features, which are aligned and fused with the reference video to extract global representations for video retrieval. Experiments on FineCVR-1M dataset demonstrate the superior performance of FDCA. Our code and dataset are available at: https://may2333.github.io/FineCVR/.
****

Teng Hu,Jiangning Zhang,Ran Yi,Hongrui Huang,Yabiao Wang,Lizhuang Ma

SaRA: High-Efficient Diffusion Model Fine-tuning with Progressive Sparse Low-Rank Adaptation

The development of diffusion models has led to significant progress in image and video generation tasks, with pre-trained models like the Stable Diffusion series playing a crucial role.

However, a key challenge remains in downstream task applications: how to effectively and efficiently adapt pre-trained diffusion models to new tasks.

Inspired by model pruning which lightens large pre-trained models by removing un important parameters, we propose a novel model fine-tuning method to make full u se of these ineffective parameters and enable the pre-trained model with new tas k-specified capabilities.
In this work, we first investigate the importance of parameters in pre-trained d iffusion models and discover that parameters with the smallest absolute values d o not contribute to the generation process due to training instabilities.
Based on this observation, we propose a fine-tuning method termed SaRA that re-u tilizes these temporarily ineffective parameters, equating to optimizing a spars e weight matrix to learn the task-specific knowledge.
To mitigate potential overfitting, we propose a nuclear-norm-based low-rank spar se training scheme for efficient fine-tuning.
Furthermore, we design a new progressive parameter adjustment strategy to make f ull use of the finetuned parameters.
Finally, we propose a novel unstructural backpropagation strategy, which signifi cantly reduces memory costs during fine-tuning.
Our method enhances the generative capabilities of pre-trained models in downstr eam applications and outperforms existing fine-tuning methods in maintaining mod el's generalization ability. Source code is available at https://sjtuplayer.git hub.io/projects/SaRA.
****

Tong Ding,Wanhua Li,Zhongqi Miao,Hanspeter Pfister
Tree of Attributes Prompt Learning for Vision-Language Models
Prompt learning has proven effective in adapting vision language models for down stream tasks. However, existing methods usually append learnable prompt tokens s olely with the category names to obtain textual features, which fails to fully l everage the rich context indicated in the category name. To address this issue, we propose the Tree of Attributes Prompt learning (TAP), which first instructs L LMs to generate a tree of attributes with a “concept - attribute - description” structure for each category, and then learn the hierarchy with vision and text p rompt tokens. Unlike existing methods that merely augment category names with a set of unstructured descriptions, our approach essentially distills structured k nowledge graphs associated with class names from LLMs. Furthermore, our approach introduces text and vision prompts designed to explicitly learn the correspondi ng visual attributes, effectively serving as domain experts. Additionally, the g eneral and diverse descriptions generated based on the class names may be wrong or absent in the specific given images. To address this misalignment, we further introduce a vision-conditional pooling module to extract instance-specific text features. Extensive experimental results demonstrate that our approach outperfo rms state-of-the-art methods on the zero-shot base-to-novel generalization, cros s-dataset transfer, as well as few-shot classification across 11 diverse dataset s. Code is available at https://github.com/HHenryD/TAP.
****

Darius Muglich,Johannes Forkel,Elise van der Pol,Jakob Nicolaus Foerster
Expected Return Symmetries
Symmetry is an important inductive bias that can improve model robustness and ge neralization across many deep learning domains. In multi-agent settings, a prior i known symmetries have been shown to address a fundamental coordination failure mode known as mutually incompatible symmetry breaking; e.g. in a game where two independent agents can choose to move "left" or "right", and where a reward of +1 or -1 is received when the agents choose the same action or different actions , respectively. However, the efficient and automatic discovery of environment sy mmetries, in particular for decentralized partially observable Markov decision p rocesses, remains an open problem. Furthermore, environmental symmetry breaking constitutes only one type of coordination failure, which motivates the search fo r a more accessible and broader symmetry class. In this paper, we introduce such a broader group of previously unexplored symmetries, which we call expected ret urn symmetries, which contains environment symmetries as a subgroup. We show tha t agents trained to be compatible under the group of expected return symmetries achieve better zero-shot coordination results than those using environment symme

tries. As an additional benefit, our method makes minimal a priori assumptions a
bout the structure of their environment and does not require access to ground tr
uth symmetries.
****

Roman Worschech,Bernd Rosenow
Analyzing Neural Scaling Laws in Two-Layer Networks with Power-Law Data Spectra
Neural scaling laws describe how the performance of deep neural networks scales
with key factors such as training data size, model complexity, and training time
, often following power-law behaviors over multiple orders of magnitude. Despite
 their empirical observation, the theoretical understanding of these scaling law
s remains limited. In this work, we employ techniques from statistical mechanics
 to analyze one-pass stochastic gradient descent within a student-teacher framew
ork, where both the student and teacher are two-layer neural networks. Our study
 primarily focuses on the generalization error and its behavior in response to d
ata covariance matrices that exhibit power-law spectra.
For linear activation functions, we derive analytical expressions for the genera
lization error, exploring different learning regimes and identifying conditions
under which power-law scaling emerges. Additionally, we extend our analysis to n
on-linear activation functions in the feature learning regime, investigating how
 power-law spectra in the data covariance matrix impact learning dynamics. Impor
tantly, we find that the length of the symmetric plateau depends on the number o
f distinct eigenvalues of the data covariance matrix and the number of hidden un
its, demonstrating how these plateaus behave under various configurations. In ad
dition, our results reveal a transition from exponential to power-law convergenc
e in the specialized phase when the data covariance matrix possesses a power-law
 spectrum. This work contributes to the theoretical understanding of neural scal
ing laws and provides insights into optimizing learning performance in practical
 scenarios involving complex data structures.
****

Vinitra Swamy,Syrielle Montariol,Julian Blackwell,Jibril Frej,Martin Jaggi,Tanja
 Käser
Intrinsic User-Centric Interpretability through Global Mixture of Experts
In human-centric settings like education or healthcare, model accuracy and model
 explainability are key factors for user adoption. Towards these two goals, intr
insically interpretable deep learning models have gained popularity, focusing on
 accurate predictions alongside faithful explanations. However, there exists a g
ap in the human-centeredness of these approaches, which often produce nuanced an
d complex explanations that are not easily actionable for downstream users. We p
resent InterpretCC (interpretable conditional computation), a family of intrinsi
cally interpretable neural networks at a unique point in the design space that o
ptimizes for ease of human understanding and explanation faithfulness, while mai
ntaining comparable performance to state-of-the-art models. InterpretCC achieves
 this through adaptive sparse activation of features before prediction, allowing
 the model to use a different, minimal set of features for each instance. We ext
end this idea into an interpretable, global mixture-of-experts (MoE) model that
allows users to specify topics of interest, discretely separates the feature spa
ce for each data point into topical subnetworks, and adaptively and sparsely act
ivates these topical subnetworks for prediction. We apply InterpretCC for text,
time series and tabular data across several real-world datasets, demonstrating c
omparable performance with non-interpretable baselines and outperforming intrins
ically interpretable baselines. Through a user study involving 56 teachers, Inte
rpretCC explanations are found to have higher actionability and usefulness over
other intrinsically interpretable approaches.
****

Yukang Chen,Fuzhao Xue,Dacheng Li,Qinghao Hu,Ligeng Zhu,Xiuyu Li,Yunhao Fang,Hao
tian Tang,Shang Yang,Zhijian Liu,Yihui He,Hongxu Yin,Pavlo Molchanov,Jan Kautz,L
inxi Fan,Yuke Zhu,Yao Lu,Song Han
LongVILA: Scaling Long-Context Visual Language Models for Long Videos
Long-context capability is critical for multi-modal foundation models, especiall
y for long video understanding. We introduce LongVILA, a full-stack solution for

long-context visual-language models by co-designing the algorithm and system. For model training, we upgrade existing VLMs to support long video understanding by incorporating two additional stages, i.e., long context extension and long video supervised fine-tuning. However, training on long video is computationally and memory intensive. We introduce the long-context Multi-Modal Sequence Parallelism (MM-SP) system that efficiently parallelizes long video training and inference, enabling 2M context length training on 256 GPUs without any gradient checkpointing. LongVILA efficiently extends the number of video frames of VILA from 8 to 2048, achieving 99.8% accuracy in 6,000-frame (more than 1 million tokens) video needle-in-a-haystack. LongVILA-7B demonstrates strong accuracy on 9 popular video benchmarks, e.g., 65.1% VideoMME with subtitle. Besides, MM-SP is 2.1x - 5.7x faster than ring style sequence parallelism and 1.1x - 1.4x faster than Megatron with a hybrid context and tensor parallelism. Moreover, it seamlessly integrates with Hugging Face Transformers.
****
Piotr Teterwak,Kuniaki Saito,Theodoros Tsiligkaridis,Bryan A. Plummer,Kate Saenko
Is Large-scale Pretraining the Secret to Good Domain Generalization?
Multi-Source Domain Generalization (DG) is the task of training on multiple source domains and achieving high classification performance on unseen target domains. Recent methods combine robust features from web-scale pretrained backbones with new features learned from source data, and this has dramatically improved benchmark results. However, it remains unclear if DG finetuning methods are becoming better over time, or if improved benchmark performance is simply an artifact of stronger pre-training. Prior studies have shown that perceptual similarity to pre-training data correlates with zero-shot performance, but we find the effect limited in the DG setting. Instead, we posit that having perceptually similar data in pretraining is not enough; and that it is how well these data were learned that determines performance. This leads us to introduce the Alignment Hypothesis, which states that the final DG performance will be high if and only if alignment of image and class label text embeddings is high. Our experiments confirm the Alignment Hypothesis is true, and we use it as an analysis tool of existing DG methods evaluated on DomainBed datasets by splitting evaluation data into In-pretraining (IP) and Out-of-pretraining (OOP). We show that all evaluated DG methods struggle on DomainBed-OOP, while recent methods excel on DomainBed-IP. Put together, our findings highlight the need for DG methods which can generalize beyond pretraining alignment. We release DomainBed-OOP at https://huggingface.co/datasets/PTeterwak/DomainBed_OOP.
****
Kathy Garcia,Emalie McMahon,Colin Conwell,Michael Bonner,Leyla Isik
Modeling dynamic social vision highlights gaps between deep learning and humans
Deep learning models trained on computer vision tasks are widely considered the most successful models of human vision to date. The majority of work that supports this idea evaluates how accurately these models predict behavior and brain responses to static images of objects and scenes. Real-world vision, however, is highly dynamic, and far less work has evaluated deep learning models on human responses to moving stimuli, especially those that involve more complicated, higher-order phenomena like social interactions. Here, we extend a dataset of natural videos depicting complex multi-agent interactions by collecting human-annotated sentence captions for each video, and we benchmark 350+ image, video, and language models on behavior and neural responses to the videos. As in prior work, we find that many vision models reach the noise ceiling in predicting visual scene features and responses along the ventral visual stream (often considered the primary neural substrate of object and scene recognition). In contrast, vision models poorly predict human action and social interaction ratings and neural responses in the lateral stream (a neural pathway theorized to specialize in dynamic, social vision), though video models show a striking advantage in predicting mid-level lateral stream regions. Language models (given human sentence captions of the videos) predict action and social ratings better than image and video models, but perform poorly at predicting neural responses in the lateral stream. Togethe

r, these results identify a major gap in AI's ability to match human social visi
on and provide insights to guide future model development for dynamic, natural c
ontexts.
****
Boyu Liu,Haoyu Huang,Linlin Yang,Yanjing Li,Guodong Guo,Xianbin Cao,Baochang Zha
ng
Efficient Low-Bit Quantization with Adaptive Scales for Multi-Task Co-Training
Co-training can achieve parameter-efficient multi-task models but remains unexpl
ored for quantization-aware training. Our investigation shows that directly intr
oducing co-training into existing quantization-aware training (QAT) methods resu
lts in significant performance degradation. Our experimental study identifies th
at the primary issue with existing QAT methods stems from the inadequate activat
ion quantization scales for the co-training framework. To address this issue, we
 propose Task-Specific Scales Quantization for Multi-Task Co-Training (TSQ-MTC)
to tackle mismatched quantization scales. Specifically, a task-specific learnabl
e multi-scale activation quantizer (TLMAQ) is incorporated to enrich the represe
ntational ability of shared features for different tasks. Additionally, we find
that in the deeper layers of the Transformer model, the quantized network suffer
s from information distortion within the attention quantizer. A structure-based
layer-by-layer distillation (SLLD) is then introduced to ensure that the quantiz
ed features effectively preserve the information from their full-precision count
erparts. Our extensive experiments in two co-training scenarios demonstrate the
effectiveness and versatility of TSQ-MTC. In particular, we successfully achieve
 a 4-bit quantized low-level visual foundation model based on IPT, which attains
 a PSNR comparable to the full-precision model while offering a $7.99\times$ com
pression ratio in the $\times4$ super-resolution task on the Set5 benchmark.
****
Xiaoling Hu,Karthik Gopinath,Peirong Liu,Malte Hoffmann,Koen Van Leemput,Oula Pu
onti,Juan Eugenio Iglesias
Hierarchical Uncertainty Estimation for Learning-based Registration in Neuroimag
ing
Over recent years, deep learning based image registration has achieved impressiv
e accuracy in many domains, including medical imaging and, specifically, human n
euroimaging with magnetic resonance imaging (MRI). However, the uncertainty esti
mation associated with these methods has been largely limited to the application
 of generic techniques (e.g., Monte Carlo dropout) that do not exploit the pecul
iarities of the problem domain, particularly spatial modeling. Here, we propose
a principled way to propagate uncertainties (epistemic or aleatoric) estimated a
t the level of spatial location by these methods, to the level of global transfo
rmation models, and further to downstream tasks. Specifically, we justify the c
hoice of a Gaussian distribution for the local uncertainty modeling, and then pr
opose a framework where uncertainties spread across hierarchical levels, dependi
ng on the choice of transformation model. Experiments on publicly available data
 sets show that Monte Carlo dropout correlates very poorly with the reference re
gistration error, whereas our uncertainty estimates correlate much better. Cruci
ally, the results also show that uncertainty-aware fitting of transformations im
proves the registration accuracy of brain MRI scans. Finally, we illustrate how
sampling from the posterior distribution of the transformations can be used to p
ropagate uncertainties to downstream neuroimaging tasks. Code is available at: h
ttps://github.com/HuXiaoling/Regre4Regis.
****
Gabriele Dominici,Pietro Barbiero,Francesco Giannini,Martin Gjoreski,Giuseppe Ma
rra,Marc Langheinrich
Counterfactual Concept Bottleneck Models
Current deep learning models are not designed to simultaneously address three fu
ndamental questions: predict class labels to solve a given classification task (
the "What?"), simulate changes in the situation to evaluate how this impacts cla
ss predictions (the "How?"), and imagine how the scenario should change to resul
t in different class predictions (the "Why not?"). While current approaches in c
ausal representation learning and concept interpretability are designed to addre

ss some of these questions individually (such as Concept Bottleneck Models, whic h address both ``what'' and ``how'' questions), no current deep learning model i s specifically built to answer all of them at the same time. To bridge this gap, we introduce CounterFactual Concept Bottleneck Models (CF-CBMs), a class of mod els designed to efficiently address the above queries all at once without the ne ed to run post-hoc searches. Our experimental results demonstrate that CF-CBMs: achieve classification accuracy comparable to black-box models and existing CBMs (“What?”), rely on fewer important concepts leading to simpler explanations (“H ow?”), and produce interpretable, concept-based counterfactuals (“Why not?”). Ad ditionally, we show that training the counterfactual generator jointly with the CBM leads to two key improvements: (i) it alters the model's decision-making pro cess, making the model rely on fewer important concepts (leading to simpler expl anations), and (ii) it significantly increases the causal effect of concept inte rventions on class predictions, making the model more responsive to these change s.
****
Apivich Hemachandra,Gregory Kang Ruey Lau,See-Kiong Ng,Bryan Kian Hsiang Low
PIED: Physics-Informed Experimental Design for Inverse Problems
In many science and engineering settings, system dynamics are characterized by g overning partial differential equations (PDEs), and a major challenge is to solv e inverse problems (IPs) where unknown PDE parameters are inferred based on obse rvational data gathered under limited budget.
Due to the high costs of setting up and running experiments, experimental design (ED) is often done with the help of PDE simulations to optimize for the most in formative design parameters (e.g., sensor placements) to solve such IPs, prior t o actual data collection. This process of optimizing design parameters is especi ally critical when the budget and other practical constraints make it infeasible to adjust the design parameters between trials during the experiments.
However, existing experimental design (ED) methods tend to require sequential an d frequent design parameter adjustments between trials. Furthermore, they also h ave significant computational bottlenecks due to the need for complex numerical simulations for PDEs, and do not exploit the advantages provided by physics info rmed neural networks (PINNs) in solving IPs for PDE-governed systems, such as it s meshless solutions, differentiability, and amortized training.
This work presents Physics-Informed Experimental Design (PIED), the first ED fra mework that makes use of PINNs in a fully differentiable architecture to perform continuous optimization of design parameters for IPs for one-shot deployments.
PIED overcomes existing methods' computational bottlenecks through parallelized computation and meta-learning of PINN parameter initialization, and proposes nov el methods to effectively take into account PINN training dynamics in optimizing the ED parameters.
Through experiments based on noisy simulated data and even real world experiment al data, we empirically show that given limited observation budget, PIED signifi cantly outperforms existing ED methods in solving IPs, including for challenging settings where the inverse parameters are unknown functions rather than just fi nite-dimensional.
****
Siyuan Qi,Bangcheng Yang,Kailin Jiang,Xiaobo Wang,Jiaqi Li,Yifan Zhong,Yaodong Y ang,Zilong Zheng
In-Context Editing: Learning Knowledge from Self-Induced Distributions
In scenarios where language models must incorporate new information efficiently without extensive retraining, traditional fine-tuning methods are prone to overf itting, degraded generalization, and unnatural language generation. To address t hese limitations, we introduce Consistent In-Context Editing (ICE), a novel appr oach leveraging the model's in-context learning capability to optimize towards a contextual distribution rather than a one-hot target. ICE introduces a simple y et effective optimization framework for the model to internalize new knowledge b y aligning its output distributions with and without additional context. This me thod enhances the robustness and effectiveness of gradient-based tuning methods, preventing overfitting and preserving the model's integrity. We analyze ICE acr

oss four critical aspects of knowledge editing: accuracy, locality, generalizati
on, and linguistic quality, demonstrating its advantages. Experimental results c
onfirm the effectiveness of ICE and demonstrate its potential for continual edit
ing, ensuring that the integrity of the model is preserved while updating inform
ation.
****
Zayne Rea Sprague,Fangcong Yin,Juan Diego Rodriguez,Dongwei Jiang,Manya Wadhwa,P
rasann Singhal,Xinyu Zhao,Xi Ye,Kyle Mahowald,Greg Durrett
To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoni
ng
Chain-of-thought (CoT) via prompting is the de facto method for eliciting reason
ing capabilities from large language models (LLMs). But for what kinds of tasks
is this extra "thinking" really helpful? To analyze this, we conducted a quantit
ative meta-analysis covering over 100 papers using CoT and ran our own evaluatio
ns of 20 datasets across 14 models. Our results show that CoT gives strong perfo
rmance benefits primarily on tasks involving math or logic, with much smaller ga
ins on other types of tasks. On MMLU, directly generating the answer without CoT
 leads to almost identical accuracy as CoT unless the question or model's respon
se contains an equals sign, indicating symbolic operations and reasoning. Follow
ing this finding, we analyze the behavior of CoT on these problems by separating
 planning and execution and comparing against tool-augmented LLMs. Much of CoT's
 gain comes from improving symbolic execution, but it underperforms relative to
using a symbolic solver. Our results indicate that CoT can be applied selectivel
y, maintaining performance while saving inference costs. Furthermore, they sugge
st a need to move beyond prompt-based CoT to new paradigms that better leverage
intermediate computation across the whole range of LLM applications.
****
Kangfu Mei,Mo Zhou,Vishal M. Patel
Field-DiT: Diffusion Transformer on Unified Video, 3D, and Game Field Generation
The probabilistic field models the distribution of continuous functions defined
over metric spaces. While these models hold great potential for unifying data ge
neration across various modalities, including images, videos, and 3D geometry, t
hey still struggle with long-context generation beyond simple examples. This lim
itation can be attributed to their MLP architecture, which lacks sufficient indu
ctive bias to capture global structures through uniform sampling.
To address this, we propose a new and simple model that incorporates a view-wise
 sampling algorithm to focus on local structure learning, along with autoregress
ive generation to preserve global geometry. It adapts cross-modality conditions,
 such as text prompts for text-to-video generation, camera poses for 3D view gen
eration, and control actions for game generation.
Experimental results across various modalities demonstrate the effectiveness of
our model, with its 675M parameter size, and highlight its potential as a founda
tional framework for scalable, modality-unified visual content generation. Our p
roject page can be found at https://kfmei.com/Field-DiT/.
****
Tobias Leemann,Periklis Petridis,Giuseppe Vietri,Dionysis Manousakas,Aaron Roth,
Sergul Aydore
Auto-GDA: Automatic Domain Adaptation for Efficient Grounding Verification in Re
trieval-Augmented Generation
While retrieval-augmented generation (RAG) has been shown to enhance factuality
of large language model (LLM) outputs, LLMs still suffer from hallucination, gen
erating incorrect or irrelevant information. A common detection strategy involve
s prompting the LLM again to assess whether its response is grounded in the retr
ieved evidence, but this approach is costly. Alternatively, lightweight natural
language inference (NLI) models for efficient grounding verification can be used
 at inference time. While existing pre-trained NLI models offer potential soluti
ons, their performance remains subpar compared to larger models on realistic RAG
 inputs. RAG inputs are more complex than most datasets used for training NLI mo
dels and have characteristics specific to the underlying knowledge base, requiri
ng adaptation of the NLI models to a specific target domain. Additionally, the l

ack of labeled instances in the target domain makes supervised domain adaptation , e.g., through fine-tuning, infeasible. To address these challenges, we introdu ce Automatic Generative Domain Adaptation (Auto-GDA). Our framework enables unsu pervised domain adaptation through synthetic data generation.
Unlike previous methods that rely on handcrafted filtering and augmentation stra tegies, Auto-GDA employs an iterative process to continuously improve the qualit y of generated samples using weak labels from less efficient teacher models and discrete optimization to select the most promising augmented samples. Experiment al results demonstrate the effectiveness of our approach, with models fine-tuned on synthetic data using Auto-GDA often surpassing the performance of the teache r model and reaching the performance level of LLMs at 10% of their computational cost.
****
Die Chen,Zhiwen Li,Mingyuan Fan,Cen Chen,Wenmeng Zhou,Yanhao Wang,Yaliang Li
Growth Inhibitors for Suppressing Inappropriate Image Concepts in Diffusion Mode ls
Despite their remarkable image generation capabilities, text-to-image diffusion models inadvertently learn inappropriate concepts from vast and unfiltered train ing data, which leads to various ethical and business risks. Specifically, model -generated images may exhibit not safe for work (NSFW) content and style copyrig ht infringements. The prompts that result in these problems often do not include explicit unsafe words; instead, they contain obscure and associative terms, whi ch are referred to as *implicit unsafe prompts*. Existing approaches directly fi ne-tune models under textual guidance to alter the cognition of the diffusion mo del, thereby erasing inappropriate concepts. This not only requires concept-spec ific fine-tuning but may also incur catastrophic forgetting. To address these is sues, we explore the representation of inappropriate concepts in the image space and guide them towards more suitable ones by injecting *growth inhibitors*, whi ch are tailored based on the identified features related to inappropriate concep ts during the diffusion process. Additionally, due to the varying degrees and sc opes of inappropriate concepts, we train an adapter to infer the corresponding s uppression scale during the injection process. Our method effectively captures t he manifestation of subtle words at the image level, enabling direct and efficie nt erasure of target concepts without the need for fine-tuning. Through extensiv e experimentation, we demonstrate that our approach achieves superior erasure re sults with little effect on other normal concepts while preserving image quality and semantics.
****
Devdhar Patel,Hava T Siegelmann
Overcoming Slow Decision Frequencies in Continuous Control: Model-Based Sequence Reinforcement Learning for Model-Free Control
Reinforcement learning (RL) is rapidly reaching and surpassing human-level contr ol capabilities. However, state-of-the-art RL algorithms often require timesteps and reaction times significantly faster than human capabilities, which is impra ctical in real-world settings and typically necessitates specialized hardware. W e introduce Sequence Reinforcement Learning (SRL), an RL algorithm designed to p roduce a sequence of actions for a given input state, enabling effective control at lower decision frequencies. SRL addresses the challenges of learning action sequences by employing both a model and an actor-critic architecture operating a t different temporal scales. We propose a "temporal recall" mechanism, where the critic uses the model to estimate intermediate states between primitive actions , providing a learning signal for each individual action within the sequence. On ce training is complete, the actor can generate action sequences independently o f the model, achieving model-free control at a slower frequency. We evaluate SRL on a suite of continuous control tasks, demonstrating that it achieves performa nce comparable to state-of-the-art algorithms while significantly reducing actor sample complexity. To better assess performance across varying decision frequen cies, we introduce the Frequency-Averaged Score (FAS) metric. Our results show t hat SRL significantly outperforms traditional RL algorithms in terms of FAS, mak ing it particularly suitable for applications requiring variable decision freque

ncies. Furthermore, we compare SRL with model-based online planning, showing tha
t SRL achieves comparable FAS while leveraging the same model during training th
at online planners use for planning.
****

Saman Forouzandeh,Parham Moradi DW,Mahdi Jalili
DistillHGNN: A Knowledge Distillation Approach for High-Speed Hypergraph Neural
Networks
In this paper, we propose a novel framework to significantly enhance the inferen
ce speed and memory efficiency of Hypergraph Neural Networks (HGNNs) while prese
rving their high accuracy. Our approach utilizes an advanced teacher-student kno
wledge distillation strategy. The teacher model, consisting of an HGNN and a Mul
ti-Layer Perceptron (MLP), not only produces soft labels but also transfers stru
ctural and high-order information to a lightweight Graph Convolutional Network (
GCN) known as TinyGCN. This dual transfer mechanism enables the student model to
 effectively capture complex dependencies while benefiting from the faster infer
ence and lower computational cost of the lightweight GCN. The student model is t
rained using both labeled data and soft labels provided by the teacher, with con
trastive learning further ensuring that the student retains high-order relations
hips. This makes the proposed method efficient and suitable for real-time applic
ations, achieving performance comparable to traditional HGNNs but with significa
ntly reduced resource requirements.
****

Jiajian Li,Qi Wang,Yunbo Wang,Xin Jin,Yang Li,Wenjun Zeng,Xiaokang Yang
Open-World Reinforcement Learning over Long Short-Term Imagination
Training visual reinforcement learning agents in a high-dimensional open world p
resents significant challenges. While various model-based methods have improved
sample efficiency by learning interactive world models, these agents tend to be
“short-sighted”, as they are typically trained on short snippets of imagined exp
eriences. We argue that the primary challenge in open-world decision-making is i
mproving the exploration efficiency across a vast state space, especially for ta
sks that demand consideration of long-horizon payoffs. In this paper, we present
 LS-Imagine, which extends the imagination horizon within a limited number of st
ate transition steps, enabling the agent to explore behaviors that potentially l
ead to promising long-term feedback. The foundation of our approach is to build
a $\textit{long short-term world model}$. To achieve this, we simulate goal-cond
itioned jumpy state transitions and compute corresponding affordance maps by zoo
ming in on specific areas within single images. This facilitates the integration
 of direct long-term values into behavior learning. Our method demonstrates sign
ificant improvements over state-of-the-art techniques in MineDojo.
****

Bodhisattwa Prasad Majumder,Harshit Surana,Dhruv Agarwal,Bhavana Dalvi Mishra,Ab
hijeetsingh Meena,Aryan Prakhar,Tirth Vora,Tushar Khot,Ashish Sabharwal,Peter Cl
ark
DiscoveryBench: Towards Data-Driven Discovery with Large Language Models
Can the rapid advances in code generation, function calling, and data analysis u
sing large language models (LLMs) help automate the search and verification of h
ypotheses purely from a set of provided datasets? To evaluate this question, we
present DiscoveryBench, the first comprehensive benchmark that formalizes the mu
lti-step process of data-driven discovery. The benchmark is designed to systemat
ically assess current model capabilities in discovery tasks and provide a useful
 resource for improving them. Our benchmark contains 264 tasks collected across
6 diverse domains, such as sociology and engineering, by manually deriving disco
very workflows from published papers to approximate the real-world challenges fa
ced by researchers, where each task is defined by a dataset, its metadata, and a
 discovery goal in natural language. We additionally provide 903 synthetic tasks
 to conduct controlled evaluations on data-driven workflows that are not covered
 in the manually collected split. Furthermore, our structured formalism of data-
driven discovery enables a facet-based evaluation that provides useful insights
into different failure modes. We evaluate several popular LLM-based reasoning fr
ameworks using both open and closed LLMs as baselines on DiscoveryBench and find

that even the best system scores only 25%. Our benchmark, thus, illustrates the challenges in autonomous data-driven discovery and serves as a valuable resource for the community to make progress.
****
Yue Jiang,Haokun Lin,Yang Bai,Bo Peng,Zhili Liu,Yueming Lyu,Yong Yang,Xingzheng, Jing Dong
Image-level Memorization Detection via Inversion-based Inference Perturbation
Recent studies have discovered that widely used text-to-image diffusion models can replicate training samples during image generation, a phenomenon known as memorization. Existing detection methods primarily focus on identifying memorized prompts. However, in real-world scenarios, image owners may need to verify whether their proprietary or personal images have been memorized by the model, even in the absence of paired prompts or related metadata. We refer to this challenge as image-level memorization detection, where current methods relying on original prompts fall short. In this work, we uncover two characteristics of memorized images after perturbing the inference procedure: lower similarity of the original images and larger magnitudes of TCNP.
Building on these insights, we propose Inversion-based Inference Perturbation (IIP), a new framework for image-level memorization detection. Our approach uses unconditional DDIM inversion to derive latent codes that contain core semantic information of original images and optimizes random prompt embeddings to introduce effective perturbation. Memorized images exhibit distinct characteristics within the proposed pipeline, providing a robust basis for detection. To support this task, we construct a comprehensive setup for the image-level memorization detection, carefully curating datasets to simulate realistic memorization scenarios. Using this setup, we evaluate our IIP framework across three different memorization settings, demonstrating its state-of-the-art performance in identifying memorized images in various settings, even in the presence of data augmentation attacks.
****
Zixuan Xiong,Guangwei Xu,Wenkai Zhang,Yuan Miao,Xuan Wu,LinHai,Ruijie Guo,Hai-Tao Zheng
Youku Dense Caption: A Large-scale Chinese Video Dense Caption Dataset and Benchmarks
With the explosive growth of video content, video captions have emerged as a crucial tool for video comprehension, significantly enhancing the ability to understand and retrieve information from videos. However, most publicly available dense video captioning datasets are in English, resulting in a scarcity of large-scale and high-quality Chinese dense video captioning datasets. To address this gap within the Chinese community and to promote the advancement of Chinese multi-modal models, we develop the first, large-scale, and high-quality Chinese dense video captioning dataset, named Youku Dense Caption. This dataset is sourced from Youku, a prominent Chinese video-sharing website. Youku Dense Caption includes 31,466 complete short videos annotated by 311,921 Chinese captions. To the best of our knowledge, it is currently the largest publicly available dataset for fine-grained Chinese video descriptions. Additionally, we establish several benchmarks for Chinese video-language tasks based on the Youku Dense Caption, including retrieval, grounding, and generation tasks. Extensive experiments and evaluations are conducted on existing state-of-the-art multi-modal models, demonstrating the dataset's utility and the potential for further research.
****
Shuofei Qiao,Runnan Fang,Zhisong Qiu,Xiaobin Wang,Ningyu Zhang,Yong Jiang,Pengjun Xie,Fei Huang,Huajun Chen
Benchmarking Agentic Workflow Generation
Large Language Models (LLMs), with their exceptional ability to handle a wide range of tasks, have driven significant advancements in tackling reasoning and planning tasks, wherein decomposing complex problems into executable workflows is a crucial step in this process. Existing workflow evaluation frameworks either focus solely on holistic performance or suffer from limitations such as restricted scenario coverage, simplistic workflow structures, and lax evaluation standards

. To this end, we introduce WorfBench, a unified workflow generation benchmark w
ith multi-faceted scenarios and intricate graph workflow structures. Additionall
y, we present WorfEval, a systemic evaluation protocol utilizing subsequence and
 subgraph matching algorithms to accurately quantify the LLM agent's workflow ge
neration capabilities. Through comprehensive evaluations across different types
of LLMs, we discover distinct gaps between the sequence planning capabilities an
d graph planning capabilities of LLM agents, with even GPT-4 exhibiting a gap of
 around 15%. We also train two open-source models and evaluate their generalizat
ion abilities on held-out tasks. Furthermore, we observe that the generated work
flows can enhance downstream tasks, enabling them to achieve superior performanc
e with less time during inference. Code and dataset are available at https://git
hub.com/zjunlp/WorfBench.
****
Zhaoning Yu,Hongyang Gao
MAGE: Model-Level Graph Neural Networks Explanations via Motif-based Graph Gener
ation
Graph Neural Networks (GNNs) have shown remarkable success in molecular tasks, y
et their interpretability remains challenging. Traditional model-level explanati
on methods like XGNN and GNNInterpreter often fail to identify valid substructur
es like rings, leading to questionable interpretability. This limitation stems f
rom XGNN's atom-by-atom approach and GNNInterpreter's reliance on average graph
embeddings, which overlook the essential structural elements crucial for molecul
es. To address these gaps, we introduce an innovative **M**otif-b**A**sed **G**N
N **E**xplainer (MAGE) that uses motifs as fundamental units for generating expl
anations. Our approach begins with extracting potential motifs through a motif d
ecomposition technique. Then, we utilize an attention-based learning method to i
dentify class-specific motifs. Finally, we employ a motif-based graph generator
for each class to create molecular graph explanations based on these class-speci
fic motifs. This novel method not only incorporates critical substructures into
the explanations but also guarantees their validity, yielding results that are h
uman-understandable. Our proposed method's effectiveness is demonstrated through
 quantitative and qualitative assessments conducted on six real-world molecular
datasets.
****
Qiong Wu,Zhaoxi Ke,Yiyi Zhou,Xiaoshuai Sun,Rongrong Ji
Routing Experts: Learning to Route Dynamic Experts in Existing Multi-modal Large
 Language Models
Recently, mixture of experts (MoE) has become a popular paradigm for achieving t
he trade-off between modal capacity and efficiency of multimodal large language
models (MLLMs). Different from previous efforts, we are dedicated to exploring t
he dynamic experts in existing MLLMs and showing that a standard MLLM can also b
e a mixture of experts. However, achieving this target is still notoriously chal
lenging. The well-trained MLLMs are more accustomed to the fixed pathway and a d
rastic change in its inference manner also greatly impedes its performance. To a
ddress these issues, we propose a novel dynamic expert routing method for existi
ng MLLMs, termed Routing Experts (RoE), which can achieve example-dependent opti
mal path routing without obvious structure tweaks. Meanwhile, a new structure sp
arsity regularization is also introduced to force the well-trained MLLMs to lear
n more short-cut pathways. In addition, we also address the alignment of the tra
ining and inference of MLLMs in terms of network routing. To validate RoE, we ap
ply it to a set of existing MLLMs, including LLaVA-1.5, LLaVA-HR and VILA, and c
onduct extensive experiments on a bunch of VL benchmarks. The experiment results
 not only show the effectiveness of our RoE in improving MLLMs' efficiency, but
also yield obvious advantages over MoE-LLaVA in both performance and speed, e.g.
, an average performance gain of 3.3% on 5 benchmarks while being 1.61 times fa
ster. Our code is anonymously released at https://github.com/DoubtedSteam/RoE
****
Masanari Kimura,Takahiro Kawashima,Tasuku Soma,Hideitsu Hino
Difference-of-submodular Bregman Divergence
The Bregman divergence, which is generated from a convex function, is commonly u

sed as a pseudo-distance for comparing vectors or functions in continuous spaces . In contrast, defining an analog of the Bregman divergence for discrete spaces is nontrivial. Iyer & Bilmes (2012b) considered Bregman divergences on discrete domains using submodular functions as generating functions, the discrete analogs of convex functions. In this paper, we further generalize this framework to cas es where the generating function is neither submodular nor supermodular, thus in creasing the flexibility and representational capacity of the resulting divergen ce, which we term the difference-of-submodular Bregman divergence. Additionally, we introduce a learnable form of this divergence using permutation-invariant ne ural networks (NNs) and demonstrate through experiments that it effectively capt ures key structural properties in discrete data. As a result, the proposed metho d significantly improves the performance of existing methods on tasks such as cl ustering and set retrieval problems. This work addresses the challenge of defini ng meaningful divergences in discrete settings and provides a new tool for tasks requiring structure-preserving distance measures.
****
Sheng Cao,Mingrui Wu,Karthik Prasad,Yuandong Tian,Zechun Liu
ParamΔ for Direct Mixing: Post-Train Large Language Model At Zero Cost
The post-training phase of large language models (LLMs) is crucial for enhancing model's ability to follow instructions, improve reasoning, and align with human preferences. It also poses significant challenges, including the need for exten sive high-quality training data and the risk of overfitting. Moreover, the model development process requires post-training and evaluation after every base mode l adjustment, which incurs significant computational costs. This paper introduce s ParamΔ, an innovative approach that streamlines the post-training proce ss by transferring knowledge and capability from an existing post-trained model to a newly upgraded base model without additional training. By capturing the lat ent knowledge embedded in model parameters, our approach effectively replicates the traditional post-training and achieves comparable performance with below 2\% performance gap (98\% of transfer efficiency). As a result, ParamΔ faci litates an efficient iterative process in base model development and enables pai nless plug-and-play testing of the updated model's performance, offering a susta inable solution for continuous retraining cycles. Furthermore, ParamΔ bri ngs a new perspective on how to fully leverage models in the open-weight communi ty, where checkpoints for base models and instruction-tuned models are readily a vailable and frequently updated, by providing a cost-effective framework that ha s the potential to accelerate the iterative cycle of model development.
****
Martin Klissarov,R Devon Hjelm,Alexander T Toshev,Bogdan Mazoure
On the Modeling Capabilities of Large Language Models for Sequential Decision Ma king
Large pretrained models are showing increasingly better performance in reasoning and planning tasks across different modalities, opening the possibility to leve rage them for complex sequential decision making problems. In this paper, we inv estigate the capabilities of Large Language Models (LLMs) for reinforcement lear ning (RL) across a diversity of interactive domains. We evaluate their ability t o produce decision-making policies, either directly, by generating actions, or i ndirectly, by first generating reward models to train an agent with RL. Our resu lts show that, even without task-specific fine-tuning, LLMs excel at reward mode ling. In particular, crafting rewards through artificial intelligence (AI) feedb ack yields the most generally applicable approach and can enhance performance by improving credit assignment and exploration. Finally, in environments with unfa miliar dynamics, we explore how fine-tuning LLMs with synthetic data can signifi cantly improve their reward modeling capabilities while mitigating catastrophic forgetting, further broadening their utility in sequential decision-making tasks .
****
Harikrishna Narasimhan,Wittawat Jitkrittum,Ankit Singh Rawat,Seungyeon Kim,Neha Gupta,Aditya Krishna Menon,Sanjiv Kumar
Faster Cascades via Speculative Decoding

Cascades and speculative decoding are two common approaches to improving languag
e models' inference efficiency. Both approaches interleave two models, but via
fundamentally distinct mechanisms: deferral rule that invokes the larger model o
nly for “hard” inputs, while speculative decoding uses speculative execution to
 primarily invoke the larger model in parallel scoring mode. These mechanisms of
fer different benefits: empirically, cascades offer compelling cost-quality trad
e-offs, often even outperforming the large model; speculative cascades offer imp
ressive speed-ups, while guaranteeing quality-neutrality. In this paper, we leve
rage the best of both these approaches by designing new speculative cascading te
chniques that implement their deferral rule through speculative execution. We ch
aracterize the optimal deferral rule for our speculative cascades, and employ a
plug-in approximation to the optimal rule. Experiments with Gemma and T5 models
 on a range of language benchmarks show that our approach yields better cost qua
lity trade-offs than cascading and speculative decoding baselines.
****
Yihuai Xu,Yongwei Wang,Yifei Bi,Huangsen Cao,Zhouhan Lin,Yu Zhao,Fei Wu
Training-free LLM-generated Text Detection by Mining Token Probability Sequences
Large language models (LLMs) have demonstrated remarkable capabilities in genera
ting high-quality texts across diverse domains. However, the potential misuse of
 LLMs has raised significant concerns, underscoring the urgent need for reliable
 detection of LLM-generated texts. Conventional training-based detectors often s
truggle with generalization, particularly in cross-domain and cross-model scenar
ios. In contrast, training-free methods, which focus on inherent discrepancies t
hrough carefully designed statistical features, offer improved generalization an
d interpretability. Despite this, existing training-free detection methods typic
ally rely on global text sequence statistics, neglecting the modeling of local d
iscriminative features, thereby limiting their detection efficacy. In this work,
 we introduce a novel training-free detector, termed \textbf{Lastde}\footnote{Th
e code and data are released at \url{https://github.com/TrustMedia-zju/Lastde_De
tector}.} that synergizes local and global statistics for enhanced detection. Fo
r the first time, we introduce time series analysis to LLM-generated text detect
ion, capturing the temporal dynamics of token probability sequences. By integrat
ing these local statistics with global ones, our detector reveals significant di
sparities between human and LLM-generated texts. We also propose an efficient al
ternative, \textbf{Lastde++} to enable real-time detection. Extensive experiment
s on six datasets involving cross-domain, cross-model, and cross-lingual detecti
on scenarios, under both white-box and black-box settings, demonstrated that our
 method consistently achieves state-of-the-art performance. Furthermore, our app
roach exhibits greater robustness against paraphrasing attacks compared to exist
ing baseline methods.
****
Haiyang Jiang,Tetsuichi Wazawa,Imari Sato,Takeharu Nagai,Yinqiang Zheng
Revolutionizing EMCCD Denoising through a Novel Physics-Based Learning Framework
 for Noise Modeling
Electron-multiplying charge-coupled device (EMCCD) has been instrumental in sens
itive observations under low-light situations including astronomy, material scie
nce, and biology.
Despite its ingenious designs to enhance target signals overcoming read-out circ
uit noises, produced images are not completely noise free, which could still cas
t a cloud on desired experiment outcomes, especially in fluorescence microscopy.
Existing studies on EMCCD's noise model have been focusing on statistical charac
teristics in theory, yet unable to incorporate latest advancements in the field
of computational photography, where physics-based noise models are utilized to g
uide deep learning processes, creating adaptive denoising algorithms for ordinar
y image sensors.
Still, those models are not directly applicable to EMCCD.
In this paper, we intend to pioneer EMCCD denoising by introducing a systematic
study on physics-based noise model calibration procedures for an EMCCD camera, a
ccurately estimating statistical features of observable noise components in expe
riments, which are then utilized to generate substantial amount of authentic tra

ining samples for one of the most recent neural networks.
A first real-world test image dataset for EMCCD is captured, containing both ima
ges of ordinary daily scenes and those of microscopic contents.
Benchmarking upon the testset and authentic microscopic images, we demonstrate d
istinct advantages of our model against previous methods for EMCCD and physics-b
ased noise modeling, forging a promising new path for EMCCD denoising.
****

Zhaoyu Liu,Kan Jiang,Murong Ma,Zhe Hou,Yun Lin,Jin Song Dong
F^3Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos
Analyzing Fast, Frequent, and Fine-grained (F^3) events presents a significant
 challenge in video analytics and multi-modal LLMs. Current methods struggle to
identify events that satisfy all the F^3 criteria with high accuracy due to ch
allenges such as motion blur and subtle visual discrepancies. To advance researc
h in video understanding, we introduce F^3Set, a benchmark that consists of vi
deo datasets for precise F^3 event detection. Datasets in F^3Set are charact
erized by their extensive scale and comprehensive detail, usually encompassing o
ver 1,000 event types with precise timestamps and supporting multi-level granula
rity. Currently, F^3Set contains several sports datasets, and this framework m
ay be extended to other applications as well. We evaluated popular temporal acti
on understanding methods on F^3Set, revealing substantial challenges for exist
ing techniques. Additionally, we propose a new method, F^3ED, for F^3 event
detections, achieving superior performance. The dataset, model, and benchmark co
de are available at https://github.com/F3Set/F3Set.
****

Quan Zhang,Yuxin Qi,Xi Tang,Jinwei Fang,Xi Lin,Ke Zhang,Chun Yuan
IMDPrompter: Adapting SAM to Image Manipulation Detection by Cross-View Automate
d Prompt Learning
Using extensive training data from SA-1B, the Segment Anything Model (SAM) has d
emonstrated exceptional generalization and zero-shot capabilities, attracting wi
despread attention in areas such as medical image segmentation and remote sensin
g image segmentation. However, its performance in the field of image manipulatio
n detection remains largely unexplored and unconfirmed. There are two main chall
enges in applying SAM to image manipulation detection: a) reliance on manual pro
mpts, and b) the difficulty of single-view information in supporting cross-datas
et generalization. To address these challenges, we develops a cross-view prompt
learning paradigm called IMDPrompter based on SAM. Benefiting from the design of
 automated prompts, IMDPrompter no longer relies on manual guidance, enabling au
tomated detection and localization. Additionally, we propose components such as
Cross-view Feature Perception, Optimal Prompt Selection, and Cross-View Prompt C
onsistency, which facilitate cross-view perceptual learning and guide SAM to gen
erate accurate masks. Extensive experimental results from five datasets (CASIA,
Columbia, Coverage, IMD2020, and NIST16) validate the effectiveness of our propo
sed method.
****

Mengcheng Lan,Chaofeng Chen,Yue Zhou,Jiaxing Xu,Yiping Ke,Xinjiang Wang,Litong F
eng,Wayne Zhang
Text4Seg: Reimagining Image Segmentation as Text Generation
Multimodal Large Language Models (MLLMs) have shown exceptional capabilities in
vision-language tasks; however, effectively integrating image segmentation into
these models remains a significant challenge. In this paper, we introduce Text4S
eg, a novel text-as-mask paradigm that casts image segmentation as a text genera
tion problem, eliminating the need for additional decoders and significantly sim
plifying the segmentation process. Our key innovation is semantic descriptors, a
 new textual representation of segmentation masks where each image patch is mapp
ed to its corresponding text label. This unified representation allows seamless
integration into the auto-regressive training pipeline of MLLMs for easier optim
ization. We demonstrate that representing an image with 16×16 semantic de
scriptors yields competitive segmentation performance. To enhance efficiency, we
 introduce the Row-wise Run-Length Encoding (R-RLE), which compresses redundant
text sequences, reducing the length of semantic descriptors by 74\% and accelera

ting inference by $3\times$, without compromising performance. Extensive experim
ents across various vision tasks, such as referring expression segmentation and
comprehension, show that Text4Seg achieves state-of-the-art performance on multi
ple datasets by fine-tuning different MLLM backbones. Our approach provides an e
fficient, scalable solution for vision-centric tasks within the MLLM framework.
****
Yuchen Liang,Xiangzhong Fang,Hanting Chen,Yunhe Wang
Linear Multistep Solver Distillation for Fast Sampling of Diffusion Models
Sampling from diffusion models can be seen as solving the corresponding
probability flow ordinary differential equation (ODE).
The solving process requires a significant number of function
evaluations (NFE), making it time-consuming.
Recently, several solver search frameworks have attempted to find
better-performing model-specific solvers. However, predicting the impact of
intermediate solving strategies on final sample quality remains challenging,
rendering the search process inefficient.
In this paper, we propose a novel method for designing
solving strategies. We first introduce a unified prediction formula
for linear multistep solvers. Subsequently, we present a solver distillation
framework, which enables a student solver to mimic the sampling trajectory
generated by a teacher solver with more steps. We utilize the mean Euclidean
distance between the student and teacher sampling trajectories as a metric,
facilitating rapid adjustment and optimization of intermediate solving strate
gies.
The design space of our framework encompasses multiple aspects,
including prediction coefficients, time step schedules, and time scaling
factors.
Our framework has the ability to complete a solver search
for Stable-Diffusion in under 12 total GPU hours.
Compared to previous reinforcement learning-based
search frameworks,
our approach achieves over a 10$\times$ increase in search efficiency.
With just 5 NFE, we achieve FID scores of 3.23 on CIFAR10, 7.16 on ImageNet-6
4,
5.44 on LSUN-Bedroom, and 12.52 on MS-COCO, resulting in a 2$\times$ sampling
acceleration ratio
compared to handcrafted solvers.
****
Milad Nasr,Javier Rando,Nicholas Carlini,Jonathan Hayase,Matthew Jagielski,A. Fe
der Cooper,Daphne Ippolito,Christopher A. Choquette-Choo,Florian Tramèr,Katherin
e Lee
Scalable Extraction of Training Data from Aligned, Production Language Models
Large language models are prone to *memorizing* some of their training data. Mem
orized (and possibly sensitive) samples can then be extracted at generation time
by adversarial or benign users. There is hope that *model alignment*---a standa
rd training process that tunes a model to harmlessly follow user instructions---
would mitigate the risk of extraction. However, we develop two novel attacks tha
t undo a language model's alignment and recover thousands of training examples f
rom popular proprietary aligned models such as OpenAI's ChatGPT. Our work highli
ghts the limitations of existing safeguards to prevent training data leakage in
production language models.
****
Hanbo Cheng,Limin Lin,Chenyu Liu,Pengcheng Xia,Pengfei Hu,Jiefeng Ma,Jun Du,Jia
Pan
DAWN: Dynamic Frame Avatar with Non-autoregressive Diffusion Framework for Talki
ng head Video Generation
Talking head generation intends to produce vivid and realistic talking head vide
os from a single portrait and speech audio clip. Although significant progress h
as been made in diffusion-based talking head generation, almost all methods rely
on autoregressive strategies, which suffer from limited context utilization bey

ond the current generation step, error accumulation, and slower generation speed . To address these challenges, we present DAWN (\textbf{D}ynamic frame \textbf{A }vatar \textbf{W}ith \textbf{N}on-autoregressive diffusion), a framework that en ables all-at-once generation of dynamic-length video sequences. Specifically, it consists of two main components: (1) audio-driven holistic facial dynamics gen eration in the latent motion space, and (2) audio-driven head pose and blink gen eration. Extensive experiments demonstrate that our method generates authentic a nd vivid videos with precise lip motions, and natural pose/blink movements. Addi tionally, with a high generation speed, DAWN possesses strong extrapolation capa bilities, ensuring the stable production of high-quality long videos. These resu lts highlight the considerable promise and potential impact of DAWN in the field of talking head video generation. Furthermore, we hope that DAWN sparks further exploration of non-autoregressive approaches in diffusion models. Our code will be publicly available at \url{https://github.com/Hanbo-Cheng/DAWN-pytorch}.
****

Sunwoo Kim,Minkyu Kim,Dongmin Park

Test-time Alignment of Diffusion Models without Reward Over-optimization

Diffusion models excel in generative tasks, but aligning them with specific obje ctives while maintaining their versatility remains challenging. Existing fine-tu ning methods often suffer from reward over-optimization, while approximate guida nce approaches fail to optimize target rewards effectively. Addressing these lim itations, we propose a training-free, test-time method based on Sequential Monte Carlo (SMC) to sample from the reward-aligned target distribution. Our approach , tailored for diffusion sampling and incorporating tempering techniques, achiev es comparable or superior target rewards to fine-tuning methods while preserving diversity and cross-reward generalization. We demonstrate its effectiveness in single-reward optimization, multi-objective scenarios, and online black-box opti mization. This work offers a robust solution for aligning diffusion models with diverse downstream objectives without compromising their general capabilities. C ode is available at https://github.com/krafton-ai/DAS.
****

Tobias Norlund,Tim Isbister,Amaru Cuba Gyllensten,Paul Gabriel dos Santos,Danila Petrelli,Ariel Ekgren,Magnus Sahlgren

SWEb: A Large Web Dataset for the Scandinavian Languages

This paper presents the hitherto largest pretraining dataset for the Scandinavia n languages: the Scandinavian WEb (SWEb), comprising over one trillion tokens. T he paper details the collection and processing pipeline, and introduces a novel model-based text extractor that significantly reduces complexity in comparison w ith rule-based approaches. We also introduce a new cloze-style benchmark for eva luating language models in Swedish, and use this test to compare models trained on the SWEb data to models trained on FineWeb, with competitive results. All dat a, models and code are shared openly.
****

Qian Yu,Peng-Tao Jiang,Hao Zhang,Jinwei Chen,Bo Li,Lihe Zhang,Huchuan Lu

High-Precision Dichotomous Image Segmentation via Probing Diffusion Capacity

In the realm of high-resolution (HR), fine-grained image segmentation, the prima ry challenge is balancing broad contextual awareness with the precision required for detailed object delineation, capturing intricate details and the finest edg es of objects. Diffusion models, trained on vast datasets comprising billions of image-text pairs, such as SD V2.1, have revolutionized text-to-image synthesis by delivering exceptional quality, fine detail resolution, and strong contextual awareness, making them an attractive solution for high-resolution image segment ation. To this end, we propose DiffDIS, a diffusion-driven segmentation model th at taps into the potential of the pre-trained U-Net within diffusion models, spe cifically designed for high-resolution, fine-grained object segmentation. By lev eraging the robust generalization capabilities and rich, versatile image represe ntation prior of the SD models, coupled with a task-specific stable one-step den oising approach, we significantly reduce the inference time while preserving hig h-fidelity, detailed generation. Additionally, we introduce an auxiliary edge ge neration task to not only enhance the preservation of fine details of the object

boundaries, but reconcile the probabilistic nature of diffusion with the determ
inistic demands of segmentation. With these refined strategies in place, DiffDIS
serves as a rapid object mask generation model, specifically optimized for gene
rating detailed binary maps at high resolutions, while demonstrating impressive
accuracy and swift processing. Experiments on the DIS5K dataset demonstrate the
superiority of DiffDIS, achieving state-of-the-art results through a streamlined
inference process. The source code will be publicly available at \href{https://
github.com/qianyu-dlut/DiffDIS}{DiffDIS}.
****

Guangyin Bao,Qi Zhang,Zixuan Gong,Zhuojia Wu,Duoqian Miao
MindSimulator: Exploring Brain Concept Localization via Synthetic fMRI
Concept-selective regions within the human cerebral cortex exhibit significant a
ctivation in response to specific visual stimuli associated with particular conc
epts. Precisely localizing these regions stands as a crucial long-term goal in n
euroscience to grasp essential brain functions and mechanisms. Conventional expe
riment-driven approaches hinge on manually constructed visual stimulus collectio
ns and corresponding brain activity recordings, constraining the support and cov
erage of concept localization. Additionally, these stimuli often consist of conc
ept objects in unnatural contexts and are potentially biased by subjective prefe
rences, thus prompting concerns about the validity and generalizability of the i
dentified regions. To address these limitations, we propose a data-driven explor
ation approach. By synthesizing extensive brain activity recordings, we statisti
cally localize various concept-selective regions. Our proposed MindSimulator lev
erages advanced generative technologies to learn the probability distribution of
brain activity conditioned on concept-oriented visual stimuli. This enables the
creation of simulated brain recordings that reflect real neural response patter
ns. Using the synthetic recordings, we successfully localize several well-studie
d concept-selective regions and validate them against empirical findings, achiev
ing promising prediction accuracy. The feasibility opens avenues for exploring n
ovel concept-selective regions and provides prior hypotheses for future neurosci
ence research.
****

Berthy Feng,Ricardo Baptista,Katherine Bouman
Neural Approximate Mirror Maps for Constrained Diffusion Models
Diffusion models excel at creating visually-convincing images, but they often st
ruggle to meet subtle constraints inherent in the training data. Such constraint
s could be physics-based (e.g., satisfying a PDE), geometric (e.g., respecting s
ymmetry), or semantic (e.g., including a particular number of objects). When the
training data all satisfy a certain constraint, enforcing this constraint on a
diffusion model makes it more reliable for generating valid synthetic data and s
olving constrained inverse problems. However, existing methods for constrained d
iffusion models are restricted in the constraints they can handle. For instance,
recent work proposed to learn mirror diffusion models (MDMs), but analytical mi
rror maps only exist for convex constraints and can be challenging to derive. We
propose *neural approximate mirror maps* (NAMMs) for general, possibly non-conv
ex constraints. Our approach only requires a differentiable distance function fr
om the constraint set. We learn an approximate mirror map that transforms data i
nto an unconstrained space and a corresponding approximate inverse that maps dat
a back to the constraint set. A generative model, such as an MDM, can then be tr
ained in the learned mirror space and its samples restored to the constraint set
by the inverse map. We validate our approach on a variety of constraints, showi
ng that compared to an unconstrained diffusion model, a NAMM-based MDM substanti
ally improves constraint satisfaction. We also demonstrate how existing diffusio
n-based inverse-problem solvers can be easily applied in the learned mirror spac
e to solve constrained inverse problems.
****

Hongye Cao,Fan Feng,Meng Fang,Shaokang Dong,Tianpei Yang,Jing Huo,Yang Gao
Towards Empowerment Gain through Causal Structure Learning in Model-Based Reinfo
rcement Learning
In Model-Based Reinforcement Learning (MBRL), incorporating causal structures in

to dynamics models provides agents with a structured understanding of the enviro
nments, enabling efficient decision.
Empowerment as an intrinsic motivation enhances the ability of agents to activel
y control their environments by maximizing the mutual information between future
 states and actions.
We posit that empowerment coupled with causal understanding can improve controll
ability, while enhanced empowerment gain can further facilitate causal reasoning
 in MBRL.
To improve learning efficiency and controllability, we propose a novel framework
, Empowerment through Causal Learning (ECL), where an agent with the awareness o
f causal dynamics models achieves empowerment-driven exploration and optimizes i
ts causal structure for task learning.
Specifically, ECL operates by first training a causal dynamics model of the envi
ronment based on collected data. We then maximize empowerment under the causal s
tructure for exploration, simultaneously using data gathered through exploration
 to update causal dynamics model to be more controllable than dense dynamics mod
el without causal structure. In downstream task learning, an intrinsic curiosity
 reward is included to balance the causality, mitigating overfitting.
Importantly, ECL is method-agnostic and is capable of integrating various causal
 discovery methods.
We evaluate ECL combined with 3 causal discovery methods across 6 environmen
ts including pixel-based tasks, demonstrating its superior performance compared
to other causal MBRL methods, in terms of causal discovery, sample efficiency, a
nd asymptotic performance.
****
Alex Iacob,Lorenzo Sani,Meghdad Kurmanji,William F. Shen,Xinchi Qiu,Dongqi Cai,Y
an Gao,Nicholas Donald Lane
DEPT: Decoupled Embeddings for Pre-training Language Models
Language Model pre-training uses broad data mixtures to enhance performance acro
ss domains and languages. However, training on such heterogeneous text corpora r
equires extensive and expensive efforts. Since these data sources vary significa
ntly in lexical, syntactic, and semantic aspects, they cause negative interferen
ce or the ``curse of multilinguality''. To address these challenges we propose a
 communication-efficient pre-training framework, DEPT. Our method decouples embe
ddings from the transformer body while simultaneously training the latter on mul
tiple data sources without requiring a shared vocabulary. DEPT can: (1) train ro
bustly and effectively under significant data heterogeneity, (2) minimize token
embedding parameters to only what the data source vocabulary requires, while cut
ting communication costs in direct proportion to both the communication frequenc
y and the reduction in parameters, (3) enhance transformer body plasticity and g
eneralization, improving both average perplexity (up to 20%) and downstream task
 performance, and (4) enable training with custom optimized vocabularies per dat
a source. We demonstrate DEPT's potential via the first vocabulary-agnostic fede
rated pre-training of billion-scale models, reducing communication costs by orde
rs of magnitude and embedding memory by 4-5x.
****
Dharmesh Tailor,Alvaro Correia,Eric Nalisnick,Christos Louizos
Approximating Full Conformal Prediction for Neural Network Regression with Gauss
-Newton Influence
Uncertainty quantification is an important prerequisite for the deployment of de
ep learning models in safety-critical areas. Yet, this hinges on the uncertainty
 estimates being useful to the extent the prediction intervals are well-calibrat
ed and sharp. In the absence of inherent uncertainty estimates (e.g. pretrained
models predicting only point estimates), popular approaches that operate post-ho
c include Laplace’s method and split conformal prediction (split-CP). However, L
aplace’s method can be miscalibrated when the model is misspecified and split-CP
 requires sample splitting, and thus comes at the expense of statistical efficie
ncy. In this work, we construct prediction intervals for neural network regresso
rs post-hoc without held-out data. This is achieved by approximating the full co
nformal prediction method (full-CP). Whilst full-CP nominally requires retrainin

g the model for every test point and candidate label, we propose to train just o
nce and locally perturb model parameters using Gauss-Newton influence to approxi
mate the effect of retraining. Coupled with linearization of the network, we exp
ress the absolute residual nonconformity score as a piecewise linear function of
 the candidate label allowing for an efficient procedure that avoids the exhaust
ive search over the output space. On standard regression benchmarks and bounding
 box localization, we show the resulting prediction intervals are locally-adapti
ve and often tighter than those of split-CP.
****
Xize Cheng,Ruofan Hu,Xiaoda Yang,Jingyu Lu,Dongjie Fu,Zehan Wang,Shengpeng Ji,Ro
ngjie Huang,Boyang Zhang,Tao Jin,Zhou Zhao
VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words?
With the rapid advancement of large models, voice assistants are gradually acqui
ring the ability to engage in open-ended daily conversations with humans. Howeve
r, current spoken dialogue systems often overlook multi-modal information in aud
io beyond text, such as speech rate, volume, emphasis, and background sounds. Re
lying solely on Automatic Speech Recognition (ASR) can lead to the loss of valua
ble auditory cues, thereby weakening the system’s ability to generate contextual
ly appropriate responses. To address this limitation, we propose \textbf{VoxDial
ogue}, a comprehensive benchmark for evaluating the ability of spoken dialogue s
ystems to understand multi-modal information beyond text. Specifically, we have
identified 12 attributes highly correlated with acoustic information beyond word
s and have meticulously designed corresponding spoken dialogue test sets for eac
h attribute, encompassing a total of 4.5K multi-turn spoken dialogue samples. Fi
nally, we evaluated several existing spoken dialogue models, analyzing their per
formance on the 12 attribute subsets of VoxDialogue. Experiments have shown that
 in spoken dialogue scenarios, many acoustic cues cannot be conveyed through tex
tual information and must be directly interpreted from the audio input. In contr
ast, while direct spoken dialogue systems excel at processing acoustic signals,
they still face limitations in handling complex dialogue tasks due to their rest
ricted context understanding capabilities. All data and code will be open source
 at \url{https://voxdialogue.github.io/}.
****
Utku Umur ACIKALIN,Aaron M Ferber,Carla P Gomes
Learning to Explore and Exploit with GNNs for Unsupervised Combinatorial Optimiz
ation
Combinatorial optimization (CO) problems are pervasive
across various domains, but their NP-hard nature often necessitates problem-spec
ific
heuristic algorithms. Recent advancements in deep learning have led to the devel
opment of learning-based heuristics, yet these approaches often struggle with li
mited search capabilities.
We introduce Explore-and-Exploit GNN (X^2GNN, pronounced x-squared GNN),
a novel unsupervised neural framework that combines exploration and exploitation
 for combinatorial search optimization:
i) Exploration - X^2GNN generates multiple solutions simultaneously, promotin
g diversity in the search space;
(ii) Exploitation - X^2GNN employs neural stochastic iterative refinement to
exploit partial existing solutions, guiding the search toward promising regions
and helping escape local optima.
By balancing exploration and exploitation, X^2GNN achieves superior performanc
e and generalization on several graph CO problems including Max Cut, Max Indepen
dent Set, and Max Clique. Notably, for large Max Clique problems, X^2GNN consi
stently generates solutions within 1.2\% of optimality, while other state-of-the
-art learning-based approaches struggle to reach within 22\% of optimal. Moreove
r, X^2GNN consistently generates better solutions than Gurobi on large graphs
for all three problems under reasonable time budgets. Furthermore, X^2GNN exhi
bits exceptional generalization capabilities. For the Maximum Independent Set pr
oblem, X^2GNN outperforms state-of-the-art methods even when trained on smalle
r or out-of-distribution graphs compared to the test set. Our framework offers

a more effective and flexible approach to neural combinatorial optimization, add
ressing a key challenge in the field and providing a promising direction for fut
ure research in learning-based heuristics for combinatorial optimization.
****

Gaojie Lin,Jianwen Jiang,Chao Liang,Tianyun Zhong,Jiaqi Yang,Zerong Zheng,Yanbo
Zheng

CyberHost: A One-stage Diffusion Framework for Audio-driven Talking Body Generat
ion

Diffusion-based video generation technology has advanced significantly, catalyzi
ng a proliferation of research in human animation. While breakthroughs have been
 made in driving human animation through various modalities for portraits, most
of current solutions for human body animation still focus on video-driven method
s, leaving audio-driven taking body generation relatively underexplored. In this
 paper, we introduce CyberHost, a one-stage audio-driven talking body generation
 framework that addresses common synthesis degradations in half-body animation,
including hand integrity, identity consistency, and natural motion.
CyberHost's key designs are twofold. Firstly, the Region Attention Module (RAM)
maintains a set of learnable, implicit, identity-agnostic latent features and co
mbines them with identity-specific local visual features to enhance the synthesi
s of critical local regions. Secondly, the Human-Prior-Guided Conditions introdu
ce more human structural priors into the model, reducing uncertainty in generate
d motion patterns and thereby improving the stability of the generated videos.
To our knowledge, CyberHost is the first one-stage audio-driven human diffusion
model capable of zero-shot video generation for the human body. Extensive experi
ments demonstrate that CyberHost surpasses previous works in both quantitative a
nd qualitative aspects. CyberHost can also be extended to video-driven and audio
-video hybrid-driven scenarios, achieving similarly satisfactory results.
****

Nathan Cloos,Moufan Li,Markus Siegel,Scott L Brincat,Earl K Miller,Guangyu Rober
t Yang,Christopher J Cueva

Differentiable Optimization of Similarity Scores Between Models and Brains

How do we know if two systems - biological or artificial - process information i
n a similar way? Similarity measures such as linear regression, Centered Kernel
Alignment (CKA), Normalized Bures Similarity (NBS), and angular Procrustes dista
nce, are often used to quantify this similarity. However, it is currently unclea
r what drives high similarity scores and even what constitutes a "good" score. H
ere, we introduce a novel tool to investigate these questions by differentiating
 through similarity measures to directly maximize the score. Surprisingly, we fi
nd that high similarity scores do not guarantee encoding task-relevant informati
on in a manner consistent with neural data; and this is particularly acute for C
KA and even some variations of cross-validated and regularized linear regression
. We find no consistent threshold for a good similarity score - it depends on bo
th the measure and the dataset. In addition, synthetic datasets optimized to max
imize similarity scores initially learn the highest variance principal component
 of the target dataset, but some methods like angular Procrustes capture lower v
ariance dimensions much earlier than methods like CKA. To shed light on this, we
 mathematically derive the sensitivity of CKA, angular Procrustes, and NBS to th
e variance of principal component dimensions, and explain the emphasis CKA place
s on high variance components. Finally, by jointly optimizing multiple similarit
y measures, we characterize their allowable ranges and reveal that some similari
ty measures are more constraining than others. While current measures offer a se
emingly straightforward way to quantify the similarity between neural systems, o
ur work underscores the need for careful interpretation. We hope the tools we de
veloped will be used by practitioners to better understand current and future si
milarity measures.
****

Muyang Li,Yujun Lin,Zhekai Zhang,Tianle Cai,Xiuyu Li,Junxian Guo,Enze Xie,Chenli
n Meng,Jun-Yan Zhu,Song Han

SVDQuant: Absorbing Outliers by Low-Rank Component for 4-Bit Diffusion Models

Diffusion models can effectively generate high-quality images. However, as they

scale, rising memory demands and higher latency pose substantial deployment cha
llenges. In this work, we aim to accelerate diffusion models by quantizing their
 weights and activations to 4 bits. At such an aggressive level, both weights an
d activations are highly sensitive, where existing post-training quantization me
thods like smoothing become insufficient. To overcome this limitation, we propos
e *SVDQuant*, a new 4-bit quantization paradigm. Different from smoothing, which
 redistributes outliers between weights and activations, our approach *absorbs*
these outliers using a low-rank branch. We first consolidate the outliers by shi
fting them from activations to weights. Then, we use a high-precision, low-rank
branch to take in the weight outliers with Singular Value Decomposition (SVD), w
hile a low-bit quantized branch handles the residuals. This process eases the qu
antization on both sides. However, naively running the low-rank branch independe
ntly incurs significant overhead due to extra data movement of activations, nega
ting the quantization speedup. To address this, we co-design an inference engine
 Nunchaku that fuses the kernels of the low-rank branch into those of the low-
bit branch to cut off redundant memory access. It can also seamlessly support of
f-the-shelf low-rank adapters (LoRAs) without re-quantization. Extensive experim
ents on SDXL, PixArt-Σ, and FLUX.1 validate the effectiveness of SVDQuant
 in preserving image quality. We reduce the memory usage for the 12B FLUX.1 mode
ls by 3.5×, achieving 3.0× speedup over the 4-bit weight-only quantization (W4A1
6) baseline on the 16GB laptop 4090 GPU with INT4 precision. On the latest RTX 5
090 desktop with Blackwell architecture, we achieve a 3.1× speedup compared to t
he W4A16 model using NVFP4 precision. Our quantization library and inference eng
ine are available at https://github.com/mit-han-lab/deepcompressor/ and https://
github.com/mit-han-lab/nunchaku/, correspondingly.
****
Jiachen Jiang,Jinxin Zhou,Zhihui Zhu
Tracing Representation Progression: Analyzing and Enhancing Layer-Wise Similarit
y
Analyzing the similarity of internal representations within and across different
 models has been an important technique for understanding the behavior of deep n
eural networks. Most existing methods for analyzing the similarity between repre
sentations of high dimensions, such as those based on Centered Kernel Alignment
(CKA), rely on statistical properties of the representations for a set of data p
oints. In this paper, we focus on transformer models and study the similarity of
 representations between the hidden layers of individual transformers. In this c
ontext, we show that a simple sample-wise cosine similarity metric is capable of
 capturing the similarity and aligns with the complicated CKA. Our experimental
results on common transformers reveal that representations across layers are pos
itively correlated, with similarity increasing when layers get closer. We provid
e a theoretical justification for this phenomenon under the geodesic curve assum
ption for the learned transformer, a property that may approximately hold for re
sidual networks. We then show that an increase in representation similarity impl
ies an increase in predicted probability when directly applying the last-layer c
lassifier to any hidden layer representation. This offers a justification for {\
it saturation events}, where the model's top prediction remains unchanged across
 subsequent layers, indicating that the shallow layer has already learned the ne
cessary knowledge. We then propose an aligned training method to improve the eff
ectiveness of shallow layer by enhancing the similarity between internal represe
ntations, with trained models that enjoy the following properties: (1) more earl
y saturation events, (2) layer-wise accuracies monotonically increase and reveal
 the minimal depth needed for the given task, (3) when served as multi-exit mode
ls, they achieve on-par performance with standard multi-exit architectures which
 consist of additional classifiers designed for early exiting in shallow layers.
 To our knowledge, our work is the first to show that one common classifier is s
ufficient for multi-exit models. We conduct experiments on both vision and NLP t
asks to demonstrate the performance of the proposed aligned training.
****
Reza Bayat,Mohammad Pezeshki,Elvis Dohmatob,David Lopez-Paz,Pascal Vincent
The Pitfalls of Memorization: When Memorization Hurts Generalization

Neural networks often learn simple explanations that fit the majority of the dat
a while memorizing exceptions that deviate from these explanations. This behavio
r leads to poor generalization when the learned explanations rely on spurious co
rrelations. In this work, we formalize $\textit{the interplay between memorizati
on and generalization}$, showing that spurious correlations would particularly l
ead to poor generalization when are combined with memorization. Memorization can
 reduce training loss to zero, leaving no incentive to learn robust, generalizab
le patterns. To address this, we propose $\textit{memorization-aware training}$
(MAT), which uses held-out predictions as a signal of memorization to shift a mo
del's logits. MAT encourages learning robust patterns invariant across distribut
ions, improving generalization under distribution shifts.
****

Qi Liu,Kai Zheng,Rui Huang,Wuchao Li,Kuo Cai,Yuan Chai,Yanan Niu,Yiqun Hui,Bing
Han,Na Mou,Hongning Wang,Wentian Bao,Yun En Yu,Guorui Zhou,Han Li,Yang Song,Defu
 Lian,Kun Gai

RecFlow: An Industrial Full Flow Recommendation Dataset

Industrial recommendation systems (RS) rely on the multi-stage pipeline to balan
ce effectiveness and efficiency when delivering items from a vast corpus to user
s. Existing RS benchmark datasets primarily focus on the exposure space, where n
ovel RS algorithms are trained and evaluated. However, when these algorithms tra
nsition to real-world industrial RS, they face two critical challenges: (1) hand
ling unexposed items—a significantly larger space than the exposed one, profound
ly impacting their practical performance; and (2) overlooking the intricate inte
rplay between multiple stages of the recommendation pipeline, resulting in subop
timal system performance. To bridge the gap between offline RS benchmarks and re
al-world online environments, we introduce RecFlow—an industrial full-flow recom
mendation dataset. Unlike existing datasets, RecFlow includes samples not only f
rom the exposure space but also from unexposed items filtered at each stage of t
he RS funnel. RecFlow comprises 38 million interactions from 42,000 users across
 nearly 9 million items with additional 1.9 billion stage samples collected from
 9.3 million online requests over 37 days and spanning 6 stages. Leveraging RecF
low, we conduct extensive experiments to demonstrate its potential in designing
novel algorithms that enhance effectiveness by incorporating stage-specific samp
les. Some of these algorithms have already been deployed online at KuaiShou, con
sistently yielding significant gains. We propose RecFlow as the first comprehens
ive whole-pipeline benchmark dataset for the RS community, enabling research on
algorithm design across the entire recommendation pipeline, including selection
bias study, debiased algorithms, multi-stage consistency and optimality, multi-t
ask recommendation, and user behavior modeling.
****

Marie-Charlotte Brandenburg,Moritz Leo Grillo,Christoph Hertrich

Decomposition Polyhedra of Piecewise Linear Functions

In this paper we contribute to the frequently studied question of how to decompo
se a continuous piecewise linear (CPWL) function into a difference of two convex
 CPWL functions. Every CPWL function has infinitely many such decompositions, bu
t for applications in optimization and neural network theory, it is crucial to f
ind decompositions with as few linear pieces as possible. This is a highly chall
enging problem, as we further demonstrate by disproving a recently proposed appr
oach by Tran and Wang [Minimal representations of tropical rational functions. A
lgebraic Statistics, 15(1):27–59, 2024]. To make the problem more tractable, we
propose to fix an underlying polyhedral complex determining the possible locus o
f nonlinearity. Under this assumption, we prove that the set of decompositions f
orms a polyhedron that arises as intersection of two translated cones. We prove
that irreducible decompositions correspond to the bounded faces of this polyhedr
on and minimal solutions must be vertices. We then identify cases with a unique
minimal decomposition, and illustrate how our insights have consequences in the
theory of submodular functions. Finally, we improve upon previous constructions
of neural networks for a given convex CPWL function and apply our framework to o
btain results in the nonconvex case.
****

Hongkang Li,Yihua Zhang,Shuai Zhang,Pin-Yu Chen,Sijia Liu,Meng Wang
When is Task Vector Provably Effective for Model Editing? A Generalization Analysis of Nonlinear Transformers
Task arithmetic refers to editing the pre-trained model by adding a weighted sum of task vectors, each of which is the weight update from the pre-trained model to fine-tuned models for certain tasks. This approach recently gained attention as a computationally efficient inference method for model editing, e.g., multi-task learning, forgetting, and out-of-domain generalization capabilities. However, the theoretical understanding of why task vectors can execute various conceptual operations remains limited, due to the highly non-convexity of training Transformer-based models. To the best of our knowledge, this paper provides the first theoretical characterization of the generalization guarantees of task vector methods on nonlinear Transformers. We consider a conceptual learning setting, where each task is a binary classification problem based on a discriminative pattern. We theoretically prove the effectiveness of task addition in simultaneously learning a set of irrelevant or aligned tasks, as well as the success of task negation in unlearning one task from irrelevant or contradictory tasks. Moreover, we prove the proper selection of linear coefficients for task arithmetic to achieve guaranteed generalization to out-of-domain tasks. All of our theoretical results hold for both dense-weight parameters and their low-rank approximations. Although established in a conceptual setting, our theoretical findings were validated on a practical machine unlearning task using the large language model Phi-1.5 (1.3B).
****
Jing He,Haodong LI,huyongzhe,Guibao Shen,Yingjie CAI,Weichao Qiu,Ying-Cong Chen
DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation
In the realm of image generation, creating customized images from visual prompt with additional textual instruction emerges as a promising endeavor. However, existing methods, both tuning-based and tuning-free, struggle with interpreting the subject-essential attributes from the visual prompt. This leads to subject-irrelevant attributes infiltrating the generation process, ultimately compromising the personalization quality in both editability and ID preservation. In this paper, we present $\textbf{DisEnvisioner}$, a novel approach for effectively extracting and enriching the subject-essential features while filtering out -irrelevant information, enabling exceptional customization performance, in a $\textbf{tuning-free}$ manner and using only $\textbf{a single image}$. Specifically, the feature of the subject and other irrelevant components are effectively separated into distinctive visual tokens, enabling a much more accurate customization. Aiming to further improving the ID consistency, we enrich the disentangled features, sculpting them into a more granular representation. Experiments demonstrate the superiority of our approach over existing methods in instruction response (editability), ID consistency, inference speed, and the overall image quality, highlighting the effectiveness and efficiency of DisEnvisioner.
****
Lucas Bandarkar,Benjamin Muller,Pritish Yuvraj,Rui Hou,Nayan Singhal,Hongjiang Lv,Bing Liu
Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models
Model merging, such as model souping, is the practice of combining different models with the same architecture together without further training. In this work, we present a model merging methodology that addresses the difficulty of fine-tuning Large Language Models (LLMs) for target tasks in non-English languages, where task-specific data is often unavailable. We focus on mathematical reasoning and without in-language math data, facilitate cross-lingual transfer by composing language and math capabilities. Starting from the same pretrained model, we fine-tune separate "experts" on math instruction data in English and on generic instruction data in the target language. We then replace the top and bottom transformer layers of the math expert directly with layers from the language expert, which consequently enhances math performance in the target language. The resulting merged models outperform the individual experts and other merging methods on the

math benchmark, MGSM, by 10% across four major languages where math instruction data is scarce. In addition, this layer swapping is simple, inexpensive, and intuitive, as it is based on an interpretative analysis of the most important parameter changes during the fine-tuning of each expert. The ability to successfully re-compose LLMs for cross-lingual transfer in this manner opens up future possibilities to combine model expertise, create modular solutions, and transfer reasoning capabilities across languages all post hoc.
****
Xinxin Zhao,Wenzhe Cai,Likun Tang,Teng Wang
ImagineNav: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination
Visual navigation is an essential skill for home-assistance robots, providing the object-searching ability to accomplish long-horizon daily tasks. Many recent approaches use Large Language Models (LLMs) for commonsense inference to improve exploration efficiency. However, the planning process of LLMs is limited within texts and it is difficult to represent the spatial occupancy and geometry layout only by texts. Both are important for making rational navigation decisions. In this work, we seek to unleash the spatial perception and planning ability of Vision-Language Models (VLMs), and explore whether the VLM, with only on-board camera captured RGB/RGB-D stream inputs, can efficiently finish the visual navigation tasks in a mapless manner. We achieve this by developing the imagination-powered navigation framework ImagineNav, which imagines the future observation images at valuable robot views and translates the complex navigation planning process into a rather simple best-view image selection problem for VLM. To generate appropriate candidate robot views for imagination, we introduce the Where2Imagine module, which is distilled to align with human navigation habits. Finally, to reach the VLM preferred views, an off-the-shelf point-goal navigation policy is utilized. Empirical experiments on the challenging open-vocabulary object navigation benchmarks demonstrates the superiority of our proposed system.
****
Berivan Isik,Natalia Ponomareva,Hussein Hazimeh,Dimitris Paparas,Sergei Vassilvitskii,Sanmi Koyejo
Scaling Laws for Downstream Task Performance in Machine Translation
Scaling laws provide important insights that can guide the design of large language models (LLMs). Existing work has primarily focused on studying scaling laws for pretraining (upstream) loss. However, in transfer learning settings, in which LLMs are pretrained on an unsupervised dataset and then finetuned on a downstream task, we often also care about the downstream performance. In this work, we study the scaling behavior in a transfer learning setting, where LLMs are finetuned for machine translation tasks. Specifically, we investigate how the choice of the pretraining data and its size affect downstream performance (translation quality) as judged by: downstream cross-entropy and translation quality metrics such as BLEU and COMET scores. Our experiments indicate that the size of the fine tuning dataset and the distribution alignment between the pretraining and downstream data significantly influence the scaling behavior. With sufficient alignment, both downstream cross-entropy and translation quality scores improve monotonically with more pretraining data. In such cases, we show that it is possible to predict the downstream translation quality metrics with good accuracy using a log-law. However, there are cases where moderate misalignment causes the downstream translation scores to fluctuate or get worse with more pretraining, whereas downstream cross-entropy monotonically improves. By analyzing these, we provide new practical insights for choosing appropriate pretraining data.
****
Xuchuang Wang,Maoli Liu,Jinhang Zuo,Xutong Liu,John C.S. Lui,Mohammad Hajiesmaili
Stochastic Bandits Robust to Adversarial Attacks
This paper investigates stochastic multi-armed bandit algorithms that are robust to adversarial attacks, where an attacker can first observe the learner's action and *then* alter their reward observation.
We study two cases of this model, with or without the knowledge of an attack bud

get C, defined as an upper bound of the summation of the difference between th
e actual and altered rewards. For both cases, we devise two types of algorithms
with regret bounds having additive or multiplicative C dependence terms.
For the known attack budget case, we prove our algorithms achieve the regret bou
nd of ${O}((K/\Delta)\log T + KC)$ and $\tilde{O}(\sqrt{KTC})$ for the additive
and multiplicative C terms, respectively, where K is the number of arms, T
 is the time horizon, Δ is the gap between the expected rewards of the op
timal arm and the second-best arm, and $\tilde{O}$ hides the logarithmic factors
.
For the unknown case, we prove our algorithms achieve the regret bound of $\tild
e{O}(\sqrt{KT} + KC^2)$ and $\tilde{O}(KC\sqrt{T})$ for the additive and multipl
icative C terms, respectively.
In addition to these upper bound results, we provide several lower bounds showin
g the tightness of our bounds and the optimality of our algorithms.
These results delineate an intrinsic separation between the bandits with attacks
 and corruption models.
****
Chris Kolb,Tobias Weber,Bernd Bischl,David Rügamer
Deep Weight Factorization: Sparse Learning Through the Lens of Artificial Symmet
ries
Sparse regularization techniques are well-established in machine learning, yet t
heir application in neural networks remains challenging due to the non-different
iability of penalties like the L_1 norm, which is incompatible with stochastic
 gradient descent. A promising alternative is shallow weight factorization, wher
e weights are decomposed into two factors, allowing for smooth optimization of $
L_1$-penalized neural networks by adding differentiable L_2 regularization to
the factors.
In this work, we introduce deep weight factorization, extending previous shallow
 approaches to more than two factors. We theoretically establish equivalence of
our deep factorization with non-convex sparse regularization and analyze its imp
act on training dynamics and optimization. Due to the limitations posed by stand
ard training practices, we propose a tailored initialization scheme and identify
 important learning rate requirements necessary for training factorized networks
.
We demonstrate the effectiveness of our deep weight factorization through experi
ments on various architectures and datasets, consistently outperforming its shal
low counterpart and widely used pruning methods.
****
Eunice Yiu,Maan Qraitem,Anisa Noor Majhi,Charlie Wong,Yutong Bai,Shiry Ginosar,A
lison Gopnik,Kate Saenko
KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models
This paper investigates visual analogical reasoning in large multimodal models (
LMMs) compared to human adults and children. A “visual analogy” is an abstract r
ule inferred from one image and applied to another. While benchmarks exist for t
esting visual reasoning in LMMs, they require advanced skills and omit basic vis
ual analogies that even young children can make. Inspired by developmental psych
ology, we propose a new benchmark of 4,300 visual transformations of everyday ob
jects to test LMMs on visual analogical reasoning and compare them to children (
ages three to five) and to adults. We structure the evaluation into three stages
: identifying what changed (e.g., color, number, etc.), how it changed (e.g., ad
ded one object), and applying the rule to new scenarios. Our findings show that
while GPT-o1, GPT-4V, LLaVA-1.5, and MANTIS identify the “what” effectively, the
y struggle with quantifying the “how” and extrapolating this rule to new objects
. In contrast, children and adults exhibit much stronger analogical reasoning at
 all three stages. Additionally, the strongest tested model, GPT-o1, performs be
tter in tasks involving simple surface-level visual attributes like color and si
ze, correlating with quicker human adult response times. Conversely, more comple
x tasks such as number, rotation, and reflection, which necessitate extensive co
gnitive processing and understanding of extrinsic spatial properties in the phys
ical world, present more significant challenges. Altogether, these findings high

light the limitations of training models on data that primarily consists of 2D i
mages and text.
****
Seungju Cho,Hongsin Lee,Changick Kim
Long-tailed Adversarial Training with Self-Distillation
Adversarial training significantly enhances adversarial robustness, yet superior
 performance is predominantly achieved on balanced datasets.
 Addressing adversarial robustness in the context of unbalanced or long-tailed d
istributions is considerably more challenging, mainly due to the scarcity of tai
l data instances.
 Previous research on adversarial robustness within long-tailed distributions ha
s primarily focused on combining traditional long-tailed natural training with e
xisting adversarial robustness methods.
 In this study, we provide an in-depth analysis for the challenge that adversari
al training struggles to achieve high performance on tail classes in long-tailed
 distributions.
 Furthermore, we propose a simple yet effective solution to advance adversarial
robustness on long-tailed distributions through a novel self-distillation techni
que.
 Specifically, this approach leverages a balanced self-teacher model, which is t
rained using a balanced dataset sampled from the original long-tailed dataset.
Our extensive experiments demonstrate state-of-the-art performance in both clean
 and robust accuracy for long-tailed adversarial robustness, with significant im
provements in tail class performance on various datasets.
We improve the accuracy against PGD attacks for tail classes by 20.3, 7.1, and 3
.8 percentage points on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, wh
ile achieving the highest robust accuracy.
****
Chaoyi Zhou,Xi Liu,Feng Luo,Siyu Huang
Latent Radiance Fields with 3D-aware 2D Representations
Latent 3D reconstruction has shown great promise in empowering 3D semantic under
standing and 3D generation by distilling 2D features into the 3D space. However,
 existing approaches struggle with the domain gap between 2D feature space and 3
D representations, resulting in degraded rendering performance. To address this
challenge, we propose a novel framework that integrates 3D awareness into the 2D
 latent space. The framework consists of three stages: (1) a correspondence-awar
e autoencoding method that enhances the 3D consistency of 2D latent representati
ons, (2) a latent radiance field (LRF) that lifts these 3D-aware 2D representati
ons into 3D space, and (3) a VAE-Radiance Field (VAE-RF) alignment strategy that
 improves image decoding from the rendered 2D representations. Extensive experim
ents demonstrate that our method outperforms the state-of-the-art latent 3D reco
nstruction approaches in terms of synthesis performance and cross-dataset genera
lizability across diverse indoor and outdoor scenes. To our knowledge, this is t
he first work showing the radiance field representations constructed from 2D lat
ent representations can yield photorealistic 3D reconstruction performance.
****
Jing-Cheng Pang,Nan Tang,Kaiyuan Li,Yuting Tang,Xin-Qiang Cai,Zhen-Yu Zhang,Gang
 Niu,Masashi Sugiyama,Yang Yu
Learning View-invariant World Models for Visual Robotic Manipulation
Robotic manipulation tasks often rely on visual inputs from cameras to perceive
the environment. However, previous approaches still suffer from performance degr
adation when the camera’s viewpoint changes during manipulation. In this paper,
we propose ReViWo (Representation learning for View-invariant World model), leve
raging multi-view data to learn robust representations for control under viewpoi
nt disturbance. ReViWo utilizes an autoencoder framework to reconstruct target i
mages by an architecture that combines view-invariant representation (VIR) and v
iew-dependent representation. To train ReViWo, we collect multi-view data in sim
ulators with known view labels, meanwhile, ReViWo is simutaneously trained on Op
en X-Embodiment datasets without view labels. The VIR is then used to train a wo
rld model on pre-collected manipulation data and a policy through interaction wi

th the world model. We evaluate the effectiveness of ReViWo in various viewpoint disturbance scenarios, including control under novel camera positions and frequ ent camera shaking, using the Meta-world & PandaGym environments. Besides, we al so conduct experiments on real world ALOHA robot. The results demonstrate that R eViWo maintains robust performance under viewpoint disturbance, while baseline m ethods suffer from significant performance degradation. Furthermore, we show tha t the VIR captures task-relevant state information and remains stable for observ ations from novel viewpoints, validating the efficacy of the ReViWo approach.
****
Dong Zhang,Rui Yan,Pingcheng Dong,Kwang-Ting Cheng
Memory Efficient Transformer Adapter for Dense Predictions
While current Vision Transformer (ViT) adapter methods have shown promising accu racy, their inference speed is implicitly hindered by inefficient memory access operations, e.g., standard normalization and frequent reshaping. In this work, w e propose META, a simple and fast ViT adapter that can improve the model's memor y efficiency and decrease memory time consumption by reducing the inefficient me mory access operations. Our method features a memory-efficient adapter block tha t enables the common sharing of layer normalization between the self-attention a nd feed-forward network layers, thereby reducing the model's reliance on normali zation operations. Within the proposed block, the cross-shaped self-attention is employed to reduce the model's frequent reshaping operations. Moreover, we augm ent the adapter block with a lightweight convolutional branch that can enhance l ocal inductive biases, particularly beneficial for the dense prediction tasks, e .g., object detection, instance segmentation, and semantic segmentation. The ada pter block is finally formulated in a cascaded manner to compute diverse head fe atures, thereby enriching the variety of feature representations. Empirically, e xtensive evaluations on multiple representative datasets validate that META subs tantially enhances the predicted quality, while achieving a new state-of-the-art accuracy-efficiency trade-off. Theoretically, we demonstrate that META exhibits superior generalization capability and stronger adaptability.
****
Shuhan Zhang,Wendi Ren,Shuang Li
Logic-Logit: A Logic-Based Approach to Choice Modeling
In this study, we propose a novel rule-based interpretable choice model, {\bf Lo gic-Logit}, designed to effectively learn and explain human choices. Choice mode ls have been widely applied across various domains—such as commercial demand for ecasting, recommendation systems, and consumer behavior analysis—typically categ orized as parametric, nonparametric, or deep network-based. While recent innovat ions have favored neural network approaches for their computational power, these flexible models often involve large parameter sets and lack interpretability, l imiting their effectiveness in contexts where transparency is essential.

Previous empirical evidence shows that individuals usually use {\it heuristic de cision rules} to form their consideration sets, from which they then choose. The se rules are often represented as {\it disjunctions of conjunctions} (i.e., OR-o f-ANDs). These rules-driven, {\it consider-then-choose} decision processes enabl e people to quickly screen numerous alternatives while reducing cognitive and se arch costs. Motivated by this insight, our approach leverages logic rules to elu cidate human choices, providing a fresh perspective on preference modeling. We i ntroduce a unique combination of column generation techniques and the Frank-Wolf e algorithm to facilitate efficient rule extraction for preference modeling—a pr ocess recognized as NP-hard. Our empirical evaluation, conducted on both synthet ic datasets and real-world data from commercial and healthcare domains, demonstr ates that Logic-Logit significantly outperforms baseline models in terms of inte rpretability and accuracy.
****
Hong Li,Nanxi Li,Yuanjie Chen,Jianbin Zhu,Qinlu Guo,Cewu Lu,Yong-Lu Li
The Labyrinth of Links: Navigating the Associative Maze of Multi-modal LLMs
Multi-modal Large Language Models (MLLMs) have exhibited impressive capability. However, recently many deficiencies of MLLMs have been found compared to human i

ntelligence, $\textit{e.g.}$, hallucination. To drive the MLLMs study, the commu
nity dedicated efforts to building larger benchmarks with complex tasks. In this
 paper, we propose benchmarking an essential but usually overlooked intelligence
: $\textbf{association}$, a human's basic capability to link observation and pri
or practice memory. To comprehensively investigate MLLM's performance on the ass
ociation, we formulate the association task and devise a standard benchmark base
d on adjective and verb semantic concepts. Instead of costly data annotation and
 curation, we propose a convenient $\textbf{annotation-free}$ construction metho
d transforming the general dataset for our association tasks. Simultaneously, we
 devise a rigorous data refinement process to eliminate confusion in the raw dat
aset. Building on this database, we establish three levels of association tasks:
 single-step, synchronous, and asynchronous associations. Moreover, we conduct a
 comprehensive investigation into the MLLMs' zero-shot association capabilities,
 addressing multiple dimensions, including three distinct memory strategies, bot
h open-source and closed-source MLLMs, cutting-edge Mixture-of-Experts (MoE) mod
els, and the involvement of human experts. Our systematic investigation shows th
at current open-source MLLMs consistently exhibit poor capability in our associa
tion tasks, even the currently state-of-the-art GPT-4V(vision) also has a signif
icant gap compared to humans. We believe our benchmark would pave the way for fu
ture MLLM studies. $\textit{Our data and code are available at:} https://mvig-r
hos.com/llm_inception.
****

Wei Wu,Can Liao,zizhen Deng,Zhengrui Guo,Jinzhuo Wang
Neuron Platonic Intrinsic Representation From Dynamics Using Contrastive Learnin
g
The Platonic Representation Hypothesis posits that behind different modalities o
f data (what we sense or detect), there exists a universal, modality-independent
 representation of reality. Inspired by this, we treat each neuron as a system,
where we can detect the neuron’s multi-segment activity data under different per
ipheral conditions. We believe that, similar to the Platonic idea, there exists
a time-invariant representation behind the different segments of the same neuron
, which reflects the intrinsic properties of the neuron’s system. Intrinsic prop
erties include the molecular profiles, brain regions and morphological structure
, etc. The optimization objective for obtaining the intrinsic representation of
neurons should satisfy two criteria: (I) segments from the same neuron should ha
ve a higher similarity than segments from different neurons; (II) the representa
tions should generalize well to out-of-domain data. To achieve this, we employ c
ontrastive learning, treating different segments from the same neuron as positiv
e pairs and segments from different neurons as negative pairs. During the implem
entation, we chose the VICReg, which uses only positive pairs for optimization b
ut indirectly separates dissimilar samples via regularization terms. To validate
 the efficacy of our method, we first applied it to simulated neuron population
dynamics data generated using the Izhikevich model. We successfully confirmed th
at our approach captures the type of each neuron as defined by preset hyperparam
eters. We then applied our method to two real-world neuron dynamics datasets, in
cluding spatial transcriptomics-derived neuron type annotations and the brain re
gions where each neuron is located. The learned representations from our model n
ot only predict neuron type and location but also show robustness when tested on
 out-of-domain data (unseen animals). This demonstrates the potential of our app
roach in advancing the understanding of neuronal systems and offers valuable ins
ights for future neuroscience research.
****

Peter Lippmann,Gerrit Gerhartz,Roman Remme,Fred A. Hamprecht
Beyond Canonicalization: How Tensorial Messages Improve Equivariant Message Pass
ing
In numerous applications of geometric deep learning, the studied systems exhibit
 spatial symmetries and it is desirable to enforce these. For the symmetry of gl
obal rotations and reflections, this means that the model should be equivariant
with respect to the transformations that form the group of $\mathrm O(d)$.
While many approaches for equivariant message passing require specialized archit

ectures, including non-standard normalization layers or non-linearities, we here present a framework based on local reference frames ("local canonicalization") which can be integrated with any architecture without restrictions.
We enhance equivariant message passing based on local canonicalization by introducing tensorial messages to communicate geometric information consistently between different local coordinate frames.
Our framework applies to message passing on geometric data in Euclidean spaces of arbitrary dimension.
We explicitly show how our approach can be adapted to make a popular existing point cloud architecture equivariant. We demonstrate the superiority of tensorial messages and achieve state-of-the-art results on normal vector regression and competitive results on other standard 3D point cloud tasks.
****
Junliang Chen,Huaiyuan Xu,Yi Wang,Lap-Pui Chau
OccProphet: Pushing the Efficiency Frontier of Camera-Only 4D Occupancy Forecasting with an Observer-Forecaster-Refiner Framework
Predicting variations in complex traffic environments is crucial for the safety of autonomous driving. Recent advancements in occupancy forecasting have enabled forecasting future 3D occupied status in driving environments by observing historical 2D images. However, high computational demands make occupancy forecasting less efficient during training and inference stages, hindering its feasibility for deployment on edge agents. In this paper, we propose a novel framework, \textit{i.e.}, OccProphet, to efficiently and effectively learn occupancy forecasting with significantly lower computational requirements while improving forecasting accuracy. OccProphet comprises three lightweight components: Observer, Forecaster, and Refiner. The Observer extracts spatio-temporal features from 3D multi-frame voxels using the proposed Efficient 4D Aggregation with Tripling-Attention Fusion, while the Forecaster and Refiner conditionally predict and refine future occupancy inferences. Experimental results on nuScenes, Lyft-Level5, and nuScenes-Occupancy datasets demonstrate that OccProphet is both training- and inference-friendly. OccProphet reduces 58\%$\sim$78\% of the computational cost with a 2.6$\times$ speedup compared with the state-of-the-art Cam4DOcc. Moreover, it achieves 4\%$\sim$18\% relatively higher forecasting accuracy. Code and models are publicly available at https://github.com/JLChen-C/OccProphet.
****
Huaguan Chen,Yang Liu,Hao Sun
PINP: Physics-Informed Neural Predictor with latent estimation of fluid flows
Accurately predicting fluid dynamics and evolution has been a long-standing challenge in physical sciences. Conventional deep learning methods often rely on the nonlinear modeling capabilities of neural networks to establish mappings between past and future states, overlooking the fluid dynamics, or only modeling the velocity field, neglecting the coupling of multiple physical quantities. In this paper, we propose a new physics-informed learning approach that incorporates coupled physical quantities into the prediction process to assist with forecasting. Central to our method lies in the discretization of physical equations, which are directly integrated into the model architecture and loss function. This integration enables the model to provide robust, long-term future predictions. By incorporating physical equations, our model demonstrates temporal extrapolation and spatial generalization capabilities. Experimental results show that our approach achieves the state-of-the-art performance in spatiotemporal prediction across both numerical simulations and real-world extreme-precipitation nowcasting benchmarks.
****
Ruifeng Li,Mingqian Li,Wei Liu,Yuhua Zhou,Xiangxin Zhou,Yuan Yao,Qiang Zhang,Hongyang Chen
UniMatch: Universal Matching from Atom to Task for Few-Shot Drug Discovery
Drug discovery is crucial for identifying candidate drugs for various diseases. However, its low success rate often results in a scarcity of annotations, posing a few-shot learning problem. Existing methods primarily focus on single-scale features, overlooking the hierarchical molecular structures that determine differ

ent molecular properties. To address these issues, we introduce Universal Matchi
ng Networks (UniMatch), a dual matching framework that integrates explicit hiera
rchical molecular matching with implicit task-level matching via meta-
learning, bridging multi-level molecular representations and task-level generali
zation. Specifically, our approach explicitly captures structural features acros
s multiple levels—atoms, substructures, and molecules—via hierarchical pooling a
nd matching, facilitating precise molecular representation and comparison. Addit
ionally, we employ a meta-learning strategy for implicit task-level matching, al
lowing the model to capture shared patterns across tasks and quickly adapt to ne
w ones. This unified matching framework ensures effective molecular alignment wh
ile leveraging shared meta-knowledge for fast adaptation. Our experimental resul
ts demonstrate that UniMatch outperforms state-of-the-art methods on the Molecul
eNet and FS-Mol benchmarks, achieving improvements of 2.87% in AUROC and 6.52% i
n ΔAUPRC. UniMatch also shows excellent generalization ability on the Meta-MolNe
t benchmark.
****

Minheng Ni,YuTao Fan,Lei Zhang,Wangmeng Zuo
Visual-O1: Understanding Ambiguous Instructions via Multi-modal Multi-turn Chain
-of-thoughts Reasoning
As large-scale models evolve, language instructions are increasingly utilized in
 multi-modal tasks. Due to human language habits, these instructions often conta
in ambiguities in real-world scenarios, necessitating the integration of visual
context or common sense for accurate interpretation. However, even highly intell
igent large models exhibit observable performance limitations on ambiguous instr
uctions, where weak reasoning abilities of disambiguation can lead to catastroph
ic errors. To address this issue, this paper proposes Visual-O1, a multi-modal m
ulti-turn chain-of-thought reasoning framework. It simulates human multi-modal m
ulti-turn reasoning, providing instantial experience for highly intelligent mode
ls or empirical experience for generally intelligent models to understand ambigu
ous instructions. Unlike traditional methods that require models to possess high
 intelligence to understand long texts or perform lengthy complex reasoning, our
 framework does not notably increase computational overhead and is more general
and effective, even for generally intelligent models. Experiments show that our
method not only enhances the performance of models of different intelligence lev
els on ambiguous instructions but also improves their performance on general dat
asets. Our work highlights the potential of artificial intelligence to work like
 humans in real-world scenarios with uncertainty and ambiguity. We release our d
ata and code at https://github.com/kodenii/Visual-O1.
****

Yuntao Du.,Kailin Jiang,Zhi Gao,Chenrui Shi,Zilong Zheng,Siyuan Qi,Qing Li
MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge
Knowledge editing techniques have emerged as essential tools for updating the fa
ctual knowledge of large language models (LLMs) and multimodal models (LMMs), al
lowing them to correct outdated or inaccurate information without retraining fro
m scratch. However, existing benchmarks for multimodal knowledge editing primari
ly focus on entity-level knowledge represented as simple triplets, which fail to
 capture the complexity of real-world multimodal information. To address this is
sue, we introduce MMKE-Bench, a comprehensive **M**ulti**M**odal **K**nowledge *
*E**diting Benchmark, designed to evaluate the ability of LMMs to edit diverse v
isual knowledge in real-world scenarios. MMKE-Bench addresses these limitations
by incorporating three types of editing tasks: visual entity editing, visual sem
antic editing, and user-specific editing. Besides, MMKE-Bench uses free-form na
tural language to represent and edit knowledge, offering a more flexible and eff
ective format. The benchmark consists of 2,940 pieces of knowledge and 8,363 ima
ges across 33 broad categories, with evaluation questions automatically generate
d and human-verified. We assess five state-of-the-art knowledge editing methods
on three prominent LMMs, revealing that no method excels across all criteria, an
d that visual and user-specific edits are particularly challenging. MMKE-Bench s
ets a new standard for evaluating the robustness of multimodal knowledge editing
 techniques, driving progress in this rapidly evolving field.

****
Jianyang Zhai,Zi-Feng Mai,Chang-Dong Wang,Feidiao Yang,Xiawu Zheng,Hui Li,Yongho ng Tian
Multimodal Quantitative Language for Generative Recommendation
Generative recommendation has emerged as a promising paradigm aiming at directly generating the identifiers of the target candidates.
Most existing methods attempt to leverage prior knowledge embedded in Pre-traine d Language Models (PLMs) to improve the recommendation performance. However, the y often fail to accommodate the differences between the general linguistic knowl edge of PLMs and the specific needs of recommendation systems. Moreover, they ra rely consider the complementary knowledge between the multimodal information of items, which represents the multi-faceted preferences of users. To facilitate e fficient recommendation knowledge transfer, we propose a novel approach called M ultimodal Quantitative Language for Generative Recommendation (MQL4GRec). Our ke y idea is to transform items from different domains and modalities into a unifie d language, which can serve as a bridge for transferring recommendation knowledg e. Specifically, we first introduce quantitative translators to convert the text and image content of items from various domains into a new and concise language , known as quantitative language, with all items sharing the same vocabulary. Th en, we design a series of quantitative language generation tasks to enrich quant itative language with semantic information and prior knowledge. Finally, we ach ieve the transfer of recommendation knowledge from different domains and modalit ies to the recommendation task through pre-training and fine-tuning. We evaluate the effectiveness of MQL4GRec through extensive experiments and comparisons wit h existing methods, achieving improvements over the baseline by 11.18\%, 14.82\% , and 7.95\% on the NDCG metric across three different datasets, respectively. O ur implementation is available at: \href{https://anonymous.4open.science/r/QL4GR ec-ED65/}{\textcolor{blue}{https://anonymous.4open.science/r/MQL4GRec-ED65/}.}
****
Minhak Song,Kwangjun Ahn,Chulhee Yun
Does SGD really happen in tiny subspaces?
Understanding the training dynamics of deep neural networks is challenging due t o their high-dimensional nature and intricate loss landscapes. Recent studies ha ve revealed that, along the training trajectory, the gradient approximately alig ns with a low-rank top eigenspace of the training loss Hessian, referred to as t he dominant subspace. Given this alignment, this paper explores whether neural n etworks can be trained within the dominant subspace, which, if feasible, could l ead to more efficient training methods. Our primary observation is that when the SGD update is projected onto the dominant subspace, the training loss does not decrease further.
This suggests that the observed alignment between the gradient and the dominant subspace is spurious. Surprisingly, projecting out the dominant subspace proves to be just as effective as the original update, despite removing the majority of the original update component. We observe similar behavior across practical set ups, including the large learning rate regime (also known as Edge of Stability), Sharpness-Aware Minimization, momentum, and adaptive optimizers. We discuss the main causes and implications of this spurious alignment, shedding light on the dynamics of neural network training.
****
Dixant Mittal,Liwei Kang,Wee Sun Lee
Learning to Search from Demonstration Sequences
Search and planning are essential for solving many real-world problems. However, in numerous learning scenarios, only action-observation sequences, such as demo nstrations or instruction sequences, are available for learning. Relying solely on supervised learning with these sequences can lead to sub-optimal performance due to the vast, unseen search space encountered during training. In this paper, we introduce Differentiable Tree Search Network (D-TSN), a novel neural network architecture that learns to construct search trees from just sequences of demon strations by performing gradient descent on a best-first search tree constructio n algorithm. D-TSN enables the joint learning of submodules, including an encode

r, value function, and world model, which are essential for planning. To constru
ct the search tree, we employ a stochastic tree expansion policy and formulate i
t as another decision-making task. Then, we optimize the tree expansion policy v
ia REINFORCE with an effective variance reduction technique for the gradient com
putation. D-TSN can be applied to problems with a known world model or to scenar
ios where it needs to jointly learn a world model with a latent state space. We
study problems from these two scenarios, including Game of 24, 2D grid navigatio
n, and Procgen games, to understand when D-TSN is more helpful. Through our expe
riments, we show that D-TSN is effective, especially when the world model with a
 latent state space is jointly learned. The code is available at https://github.
com/dixantmittal/differentiable-tree-search-network.
****
Joey Hong,Anca Dragan,Sergey Levine
Q-SFT: Q-Learning for Language Models via Supervised Fine-Tuning
Value-based reinforcement learning (RL) can in principle learn effective policie
s for a wide range of multi-turn problems, from games to dialogue to robotic con
trol, including via offline RL from static previously collected datasets. Howeve
r, despite the widespread use of policy gradient methods to train large language
 models for single turn tasks (e.g., question answering), value-based methods fo
r multi-turn RL in an off-policy or offline setting have proven particularly cha
llenging to scale to the setting of large language models. This setting requires
 effectively leveraging pretraining, scaling to large architectures with billion
s of parameters, and training on large datasets, all of which represent major ch
allenges for current value-based RL methods. In this work, we propose a novel of
fline RL algorithm that addresses these drawbacks, casting Q-learning as a modif
ied supervised fine-tuning (SFT) problem where the probabilities of tokens direc
tly translate to Q-values. In this way we obtain an algorithm that smoothly tran
sitions from maximizing the likelihood of the data during pretraining to learnin
g a near-optimal Q-function during finetuning. Our algorithm has strong theoreti
cal foundations, enjoying performance bounds similar to state-of-the-art Q-learn
ing methods, while in practice utilizing an objective that closely resembles SFT
. Because of this, our approach can enjoy the full benefits of the pretraining o
f language models, without the need to reinitialize any weights before RL finetu
ning, and without the need to initialize new heads for predicting values or adva
ntages. Empirically, we evaluate our method on both pretrained LLMs and VLMs, on
 a variety of tasks including both natural language dialogue and robotic manipul
ation and navigation from images.
****
Junchao Gong,Siwei Tu,Weidong Yang,Ben Fei,Kun Chen,zhangwenlong,Xiaokang Yang,W
anli Ouyang,LEI BAI
PostCast: Generalizable Postprocessing for Precipitation Nowcasting via Unsuperv
ised Blurriness Modeling
Precipitation nowcasting plays a pivotal role in socioeconomic sectors, especial
ly in severe convective weather warnings. Although notable progress has been ach
ieved by approaches mining the spatiotemporal correlations with deep learning, t
hese methods still suffer severe blurriness as the lead time increases, which ha
mpers accurate predictions for extreme precipitation. To alleviate blurriness, r
esearchers explore generative methods conditioned on blurry predictions. However
, the pairs of blurry predictions and corresponding ground truth need to be give
n in advance, making the training pipeline cumbersome and limiting the generalit
y of generative models within blurry modes that appear in training data. By reth
inking the blurriness in precipitation nowcasting as a blur kernel acting on pre
dictions, we propose an unsupervised postprocessing method to eliminate the blur
riness without the requirement of training with the pairs of blurry predictions
and corresponding ground truth. Specifically, we utilize blurry predictions to g
uide the generation process of a pre-trained unconditional denoising diffusion p
robabilistic model (DDPM) to obtain high-fidelity predictions with eliminated bl
urriness. A zero-shot blur kernel estimation mechanism and an auto-scale denoise
 guidance strategy are introduced to adapt the unconditional DDPM to any blurrin
ess modes varying from datasets and lead times in precipitation nowcasting. Exte

nsive experiments are conducted on 7 precipitation radar datasets, demonstrating the generality and superiority of our method.
****

Javad Aliakbari,Johan Östman,Alexandre Graell i Amat
Decoupled Subgraph Federated Learning
We address the challenge of federated learning on graph-structured data distributed across multiple clients. Specifically, we focus on the prevalent scenario of interconnected subgraphs, where inter-connections between different clients play a critical role. We present a novel framework for this scenario, named FedStruct, that harnesses deep structural dependencies. To uphold privacy, unlike existing methods, FedStruct eliminates the necessity of sharing or generating sensitive node features or embeddings among clients. Instead, it leverages explicit global graph structure information to capture inter-node dependencies. We validate the effectiveness of FedStruct through experimental results conducted on six datasets for semi-supervised node classification, showcasing performance close to the centralized approach across various scenarios, including different data partitioning methods, varying levels of label availability, and number of clients.
****

Jingnan Gao,Zhuo Chen,Xiaokang Yang,Yichao Yan
AniSDF: Fused-Granularity Neural Surfaces with Anisotropic Encoding for High-Fidelity 3D Reconstruction
Neural radiance fields have recently revolutionized novel-view synthesis and achieved high-fidelity renderings.
However, these methods sacrifice the geometry for the rendering quality, limiting their further applications including relighting and deformation.
How to synthesize photo-realistic rendering while reconstructing accurate geometry remains an unsolved problem. In this work, we present AniSDF, a novel approach that learns fused-granularity neural surfaces with physics-based encoding for high-fidelity 3D reconstruction. Different from previous neural surfaces, our fused-granularity geometry structure balances the overall structures and fine geometric details, producing accurate geometry reconstruction.
To disambiguate geometry from reflective appearance, we introduce blended radiance fields to model diffuse and specularity following the anisotropic spherical Gaussian encoding, a physics-based rendering pipeline. With these designs, AniSDF can reconstruct objects with complex structures and produce high-quality renderings.
Furthermore, our method is a unified model that does not require complex hyperparameter tuning for specific objects.
Extensive experiments demonstrate that our method boosts the quality of SDF-based methods by a great scale in both geometry reconstruction and novel-view synthesis.
****

Can Pouliquen,Mathurin Massias,Titouan Vayer
Schur's Positive-Definite Network: Deep Learning in the SPD cone with structure
Estimating matrices in the symmetric positive-definite (SPD) cone is of interest for many applications ranging from computer vision to graph learning. While there exist various convex optimization-based estimators, they remain limited in expressivity due to their model-based approach. The success of deep learning motivates the use of learning-based approaches to estimate SPD matrices with neural networks in a data-driven fashion. However, designing effective neural architectures for SPD learning is challenging, particularly when the task requires additional structural constraints, such as element-wise sparsity. Current approaches either do not ensure that the output meets all desired properties or lack expressivity. In this paper, we introduce SpodNet, a novel and generic learning module that guarantees SPD outputs and supports additional structural constraints. Notably, it solves the challenging task of learning jointly SPD and sparse matrices. Our experiments illustrate the versatility and relevance of SpodNet layers for such applications.
****

Yaniv Leviathan,Matan Kalman,Yossi Matias

Selective Attention Improves Transformer
Unneeded elements in the attention’s context degrade performance. We introduce S
elective Attention, a simple parameter-free change to the standard attention mec
hanism which reduces attention to unneeded elements. Selective attention consist
ently improves language modeling and downstream task performance in a variety of
 model sizes and context lengths. For example, transformers trained with the lan
guage modeling objective on C4 with selective attention perform language modelin
g equivalently to standard transformers with ~2X more heads and parameters in th
eir attention modules. Selective attention also allows decreasing the size of th
e attention’s context buffer, leading to meaningful reductions in the memory and
 compute requirements during inference. For example, transformers trained on C4
with context sizes of 512, 1,024, and 2,048 need 16X, 25X, and 47X less memory f
or their attention module, respectively, when equipped with selective attention,
 as those without selective attention, with the same validation perplexity.
****
Runyi Hu,Jie Zhang,Yiming Li,Jiwei Li,Qing Guo,Han Qiu,Tianwei Zhang
VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking
Artificial Intelligence Generated Content (AIGC) has advanced significantly, par
ticularly with the development of video generation models such as text-to-video
(T2V) models and image-to-video (I2V) models. However, like other AIGC types, vi
deo generation requires robust content control. A common approach is to embed wa
termarks, but most research has focused on images, with limited attention given
to videos. Traditional methods, which embed watermarks frame-by-frame in a post-
processing manner, often degrade video quality. In this paper, we propose VideoS
hield, a novel watermarking framework specifically designed for popular diffusio
n-based video generation models. Unlike post-processing methods, VideoShield emb
eds watermarks directly during video generation, eliminating the need for additi
onal training. To ensure video integrity, we introduce a tamper localization fea
ture that can detect changes both temporally (across frames) and spatially (with
in individual frames). Our method maps watermark bits to template bits, which ar
e then used to generate watermarked noise during the denoising process. Using DD
IM Inversion, we can reverse the video to its original watermarked noise, enabli
ng straightforward watermark extraction. Additionally, template bits allow preci
se detection for potential spatial and temporal modification. Extensive experime
nts across various video models (both T2V and I2V models) demonstrate that our m
ethod effectively extracts watermarks and detects tamper without compromising vi
deo quality. Furthermore, we show that this approach is applicable to image gene
ration models, enabling tamper detection in generated images as well. Codes and
models are available at https://github.com/hurunyi/VideoShield.
****
Ally Yalei Du,Dung Daniel Ngo,Steven Wu
Reconciling Model Multiplicity for Downstream Decision Making
We consider the problem of model multiplicity in downstream decision-making, a s
etting where two predictive models of equivalent accuracy cannot agree on what a
ction to take for a downstream decision-making problem. Prior work attempts to a
ddress model multiplicity by resolving prediction disagreement between models. H
owever, we show that even when the two predictive models approximately agree on
their individual predictions almost everywhere, these models can lead the downst
ream decision-maker to take actions with substantially higher losses. We address
 this issue by proposing a framework that calibrates the predictive models with
respect to both a finite set of downstream decision-making problems and the indi
vidual probability prediction. Specifically, leveraging tools from multi-calibra
tion, we provide an algorithm that, at each time-step, first reconciles the diff
erences in individual probability prediction, then calibrates the updated models
 such that they are indistinguishable from the true probability distribution to
the decision-makers. We extend our results to the setting where one does not hav
e direct access to the true probability distribution and instead relies on a set
 of i.i.d data to be the empirical distribution. Furthermore, we generalize our
results to the settings where one has more than two predictive models and an inf
initely large downstream action set. Finally, we provide a set of experiments to

evaluate our methods empirically. Compared to existing work, our proposed algor
ithm creates a pair of predictive models with improved downstream decision-makin
g losses and agrees on their best-response actions almost everywhere.
****
Jialu Li,Yuanzhen Li,Neal Wadhwa,Yael Pritch,David E. Jacobs,Michael Rubinstein,
Mohit Bansal,Nataniel Ruiz
Unbounded: A Generative Infinite Game of Character Life Simulation
We introduce the concept of a generative infinite game, a video game that transc
ends the traditional boundaries of finite, hard-coded systems by using generativ
e models. Inspired by James P. Carse's distinction between finite and infinite g
ames, we leverage recent advances in generative AI to create Unbounded: a game o
f character life simulation that is fully encapsulated in generative models. Spe
cifically, Unbounded draws inspiration from sandbox life simulations and allows
you to interact with your autonomous virtual character in a virtual world by fee
ding, playing with and guiding it - with open-ended mechanics generated by an LL
M, some of which can be emergent. In order to develop Unbounded, we propose tech
nical innovations in both the LLM and visual generation domains. Specifically, w
e present: (1) a specialized, distilled large language model (LLM) that dynamica
lly generates game mechanics, narratives, and character interactions in real-tim
e, and (2) a new dynamic regional image prompt Adapter (IP-Adapter) for vision m
odels that ensures consistent yet flexible visual generation of a character acro
ss multiple environments. We evaluate our system through both qualitative and qu
antitative analysis, showing significant improvements in character life simulati
on, user instruction following, narrative coherence, and visual consistency for
both characters and the environments compared to traditional related approaches.
****
Marcel Kollovieh,Marten Lienen,David Lüdke,Leo Schwinn,Stephan Günnemann
Flow Matching with Gaussian Process Priors for Probabilistic Time Series Forecas
ting
Recent advancements in generative modeling, particularly diffusion models, have
opened new directions for time series modeling, achieving state-of-the-art perfo
rmance in forecasting and synthesis. However, the reliance of diffusion-based mo
dels on a simple, fixed prior complicates the generative process since the data
and prior distributions differ significantly. We introduce TSFlow, a conditional
 flow matching (CFM) model for time series combining Gaussian processes, optimal
 transport paths, and data-dependent prior distributions. By incorporating (cond
itional) Gaussian processes, TSFlow aligns the prior distribution more closely w
ith the temporal structure of the data, enhancing both unconditional and conditi
onal generation. Furthermore, we propose conditional prior sampling to enable pr
obabilistic forecasting with an unconditionally trained model. In our experiment
al evaluation on eight real-world datasets, we demonstrate the generative capabi
lities of TSFlow, producing high-quality unconditional samples. Finally, we show
 that both conditionally and unconditionally trained models achieve competitive
results across multiple forecasting benchmarks.
****
Simon Welker,Matthew Le,Ricky T. Q. Chen,Wei-Ning Hsu,Timo Gerkmann,Alexander Ri
chard,YI-CHIAO WU
FlowDec: A flow-based full-band general audio codec with high perceptual quality
We propose FlowDec, a neural full-band audio codec for general audio sampled at
48 kHz that combines non-adversarial codec training with a stochastic postfilter
 based on a novel conditional flow matching method. Compared to the prior work S
coreDec which is based on score matching, we generalize from speech to general a
udio and move from 24 kbit/s to as low as 4 kbit/s, while improving output quali
ty and reducing the required postfilter DNN evaluations from 60 to 6 without any
 fine-tuning or distillation techniques. We provide theoretical insights and geo
metric intuitions for our approach in comparison to ScoreDec as well as another
recent work that uses flow matching, and conduct ablation studies on our propose
d components. We show that FlowDec is a competitive alternative to the recent GA
N-dominated stream of neural codecs, achieving FAD scores better than those of t
he established GAN-based codec DAC and listening test scores that are on par, an

d producing qualitatively more natural reconstructions for speech and harmonic s
tructures in music.
****
Miruna Cretu,Charles Harris,Ilia Igashov,Arne Schneuing,Marwin Segler,Bruno Corr
eia,Julien Roy,Emmanuel Bengio,Pietro Lio
SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints
Generative models see increasing use in computer-aided drug design. However, whi
le performing well at capturing distributions of molecular motifs, they often pr
oduce synthetically inaccessible molecules. To address this, we introduce SynFlo
wNet, a GFlowNet model whose action space uses chemical reactions and buyable re
actants to sequentially build new molecules. By incorporating forward synthesis
as an explicit constraint of the generative mechanism, we aim at bridging the ga
p between in silico molecular generation and real world synthesis capabilities.
We evaluate our approach using synthetic accessibility scores and an independent
 retrosynthesis tool to assess the synthesizability of our compounds, and motiva
te the choice of GFlowNets through considerable improvement in sample diversity
compared to baselines. Additionally, we identify challenges with reaction encodi
ngs that can complicate traversal of the MDP in the backward direction. To addre
ss this, we introduce various strategies for learning the GFlowNet backward poli
cy and thus demonstrate how additional constraints can be integrated into the GF
lowNet MDP framework. This approach enables our model to successfully identify s
ynthesis pathways for previously unseen molecules.
****
Arvind Raghavan,Elias Bareinboim
Counterfactual Realizability
It is commonly believed that, in a real-world environment, samples can only be d
rawn from observational and interventional distributions, corresponding to Layer
s 1 and 2 of the *Pearl Causal Hierarchy*. Layer 3, representing counterfactual
distributions, is believed to be inaccessible by definition. However, Bareinboim
, Forney, and Pearl (2015) introduced a procedure that allows an agent to sample
 directly from a counterfactual distribution, leaving open the question of what
other counterfactual quantities can be estimated directly via physical experimen
tation. We resolve this by introducing a formal definition of realizability, the
 ability to draw samples from a distribution, and then developing a complete alg
orithm to determine whether an arbitrary counterfactual distribution is realizab
le given fundamental physical constraints, such as the inability to go back in t
ime and subject the same unit to a different experimental condition. We illustra
te the implications of this new framework for counterfactual data collection usi
ng motivating examples from causal fairness and causal reinforcement learning. W
hile the baseline approach in these motivating settings typically follows an int
erventional or observational strategy, we show that a counterfactual strategy pr
ovably dominates both.
****
Zhibing Li,Tong Wu,Jing Tan,Mengchen Zhang,Jiaqi Wang,Dahua Lin
IDArb: Intrinsic Decomposition for Arbitrary Number of Input Views and Illuminat
ions
Capturing geometric and material information from images remains a fundamental c
hallenge in computer vision and graphics. Traditional optimization-based methods
 often require hours of computational time to reconstruct geometry, material pro
perties, and environmental lighting from dense multi-view inputs, while still st
ruggling with inherent ambiguities between lighting and material. On the other h
and, learning-based approaches leverage rich material priors from existing 3D ob
ject datasets but face challenges with maintaining multi-view consistency.
In this paper, we introduce IDArb, a diffusion-based model designed to perform i
ntrinsic decomposition on an arbitrary number of images under varying illuminati
ons. Our method achieves highly accurate and multi-view consistent estimation on
 surface normals and material properties. This is made possible through a novel
cross-view, cross-domain attention module and an illumination-augmented, view-ad
aptive training strategy. Additionally, we introduce ARB-Objaverse, a new datase
t that provides large-scale multi-view intrinsic data and renderings under diver

se lighting conditions, supporting robust training.
Extensive experiments demonstrate that IDArb outperforms state-of-the-art method s both qualitatively and quantitatively. Moreover, our approach facilitates a ra nge of downstream tasks, including single-image relighting, photometric stereo, and 3D reconstruction, highlighting its broad applicability in realistic 3D cont ent creation.
Project website: https://lizb6626.github.io/IDArb/.
****
Cornelius Emde,Francesco Pinto,Thomas Lukasiewicz,Philip Torr,Adel Bibi
Towards Certification of Uncertainty Calibration under Adversarial Attacks
Since neural classifiers are known to be sensitive to adversarial perturbations that alter their accuracy, certification methods have been developed to provide provable guarantees on the insensitivity of their predictions to such perturbati ons. On the other hand, in safety-critical applications, the frequentist interpr etation of the confidence of a classifier (also known as model calibration) can be of utmost importance. This property can be measured via the Brier score or th e expected calibration error. We show that attacks can significantly harm calibr ation, and thus propose certified calibration providing worst-case bounds on cal ibration under adversarial perturbations. Specifically, we produce analytic boun ds for the Brier score and approximate bounds via the solution of a mixed-intege r program on the expected calibration error. Finally, we propose novel calibrati on attacks and demonstrate how they can improve model calibration through advers arial calibration training. The code will be publicly released upon acceptance.
****
Yi Xiao,Qilong Jia,Kun Chen,LEI BAI,Wei Xue
VAE-Var: Variational Autoencoder-Enhanced Variational Methods for Data Assimilat ion in Meteorology
Data assimilation (DA) is an essential statistical technique for generating accu rate estimates of a physical system's states by combining prior model prediction s with observational data, especially in the realm of weather forecasting. Effec tively modeling the prior distribution while adapting to diverse observational s ources presents significant challenges for both traditional and neural network-b ased DA algorithms. This paper introduces VAE-Var, a novel neural network-based data assimilation algorithm aimed at 1) enhancing accuracy by capturing the non-Gaussian characteristics of the conditional background distribution $p(\mathbf{x }|\mathbf{x}_b)$, and 2) efficiently assimilating real-world observational data. VAE-Var utilizes a variational autoencoder to learn the background error distri bution, with its decoder creating a variational cost function to optimize the an alysis states. The advantages of VAE-Var include: 1) it maintains the framework of traditional variational assimilation, enabling it to accommodate various obse rvation operators, particularly irregular observations; 2) it lessens the depend ence on expert knowledge for constructing the background distribution, allowing for improved modeling of non-Gaussian structures; and 3) experimental results in dicate that, when applied to the FengWu weather forecasting model, VAE-Var outpe rforms DiffDA and two traditional algorithms (interpolation and 3DVar) in terms of assimilation accuracy in sparse observational contexts, and is capable of ass imilating real-world GDAS prepbufr observations over a year.
****
Yuxiang Guo,Lu Yin,Bo Jiang,Jiaqi Zhang
TODO: Enhancing LLM Alignment with Ternary Preferences
Aligning large language models (LLMs) with human intent is critical for enhancin g their performance across a variety of tasks. Standard alignment techniques, su ch as Direct Preference Optimization (DPO), often rely on the binary Bradley-Ter ry (BT) model, which can struggle to capture the complexities of human preferenc es—particularly in the presence of noisy or inconsistent labels and frequent tie s. To address these limitations, we introduce the Tie-rank Oriented Bradley-Terr y model (TOBT), an extension of the BT model that explicitly incorporates ties, enabling more nuanced preference representation. Building on this, we propose Ti e-rank Oriented Direct Preference Optimization (TODO), a novel alignment algorit hm that leverages TOBT's ternary ranking system to improve preference alignment.

In evaluations on Mistral-7B and Llama 3-8B models, TODO consistently outperforms DPO in modeling preferences across both in-distribution and out-of-distribution datasets. Additional assessments using MT Bench and benchmarks such as Piqa, ARC-c, and MMLU further demonstrate TODO's superior alignment performance. Notably, TODO also shows strong results in binary preference alignment, highlighting its versatility and potential for broader integration into LLM alignment. The code for TODO is made publicly available.
****

Jorge Loria,Anindya Bhadra

Deep Kernel Posterior Learning under Infinite Variance Prior Weights

Neal (1996) proved that infinitely wide shallow Bayesian neural networks (BNN) converge to Gaussian processes (GP), when the network weights have bounded prior variance. Cho & Saul (2009) provided a useful recursive formula for deep kernel processes for relating the covariance kernel of each layer to the layer immediately below. Moreover, they worked out the form of the layer-wise covariance kernel in an explicit manner for several common activation functions, including the ReLU. Subsequent works have made the connection between these two works, and provided useful results on the covariance kernel of a deep GP arising as wide limits of various deep Bayesian network architectures. However, recent works, including Aitchison et al. (2021), have highlighted that the covariance kernels obtained in this manner are deterministic and hence, precludes any possibility of representation learning, which amounts to learning a non-degenerate posterior of a random kernel given the data. To address this, they propose adding artificial noise to the kernel to retain stochasticity, and develop deep kernel Wishart and inverse Wishart processes. Nonetheless, this artificial noise injection could be critiqued in that it would not naturally emerge in a classic BNN architecture under an infinite-width limit. To address this, we show that a Bayesian deep neural network, where each layer width approaches infinity, and all network weights are elliptically distributed with infinite variance, converges to a process with α-stable marginals in each layer that has a conditionally Gaussian representation. These conditional random covariance kernels could be recursively linked in the manner of Cho & Saul (2009), even though marginally the process exhibits stable behavior, and hence covariances are not even necessarily defined. We also provide useful generalizations of the recent results of Loría & Bhadra (2024) on shallow networks to multi-layer networks, and remedy the prohibitive computational burden of their approach. The computational and statistical benefits over competing approaches stand out in simulations and in demonstrations on benchmark data sets.
****

Phillip Si,Peng Chen

Latent-EnSF: A Latent Ensemble Score Filter for High-Dimensional Data Assimilation with Sparse Observation Data

Accurate modeling and prediction of complex physical systems often rely on data assimilation techniques to correct errors inherent in model simulations. Traditional methods like the Ensemble Kalman Filter (EnKF) and its variants as well as the recently developed Ensemble Score Filters (EnSF) face significant challenges when dealing with high-dimensional and nonlinear Bayesian filtering problems with sparse observations, which are ubiquitous in real-world applications. In this paper, we propose a novel data assimilation method, Latent-EnSF, which leverages EnSF with efficient and consistent latent representations of the full states and sparse observations to address the joint challenges of high dimensionlity in states and high sparsity in observations for nonlinear Bayesian filtering. We introduce a coupled

Variational Autoencoder (VAE) with two encoders to encode the full states and sparse observations in a consistent way guaranteed by a latent distribution matching and regularization as well as a consistent state reconstruction. With comparison to several methods, we demonstrate the higher accuracy, faster convergence, and higher efficiency of Latent-EnSF for two challenging applications with complex models in shallow water wave propagation and medium-range weather forecasting, for highly sparse observations in both space and time.

****
Hyungho Na,Kwanghyeon Lee,Sumin Lee,Il-chul Moon
Trajectory-Class-Aware Multi-Agent Reinforcement Learning
In the context of multi-agent reinforcement learning, *generalization* is a chal
lenge to solve various tasks that may require different joint policies or coordi
nation without relying on policies specialized for each task. We refer to this t
ype of problem as a *multi-task*, and we train agents to be versatile in this mu
lti-task setting through a single training process. To address this challenge, w
e introduce TRajectory-class-Aware Multi-Agent reinforcement learning (TRAMA). I
n TRAMA, agents recognize a task type by identifying the class of trajectories t
hey are experiencing through partial observations, and the agents use this traje
ctory awareness or prediction as additional information for action policy. To th
is end, we introduce three primary objectives in TRAMA: (a) constructing a quant
ized latent space to generate trajectory embeddings that reflect key similaritie
s among them; (b) conducting trajectory clustering using these trajectory embedd
ings; and (c) building a trajectory-class-aware policy. Specifically for (c), we
 introduce a trajectory-class predictor that performs agent-wise predictions on
the trajectory class; and we design a trajectory-class representation model for
each trajectory class. Each agent takes actions based on this trajectory-class r
epresentation along with its partial observation for task-aware execution. The p
roposed method is evaluated on various tasks, including multi-task problems buil
t upon StarCraft II. Empirical results show further performance improvements ove
r state-of-the-art baselines.
****
Seongmin Lee,Marcel Boehme
How Much is Unseen Depends Chiefly on Information About the Seen
The *missing mass* refers to the proportion of data points in an *unknown* popul
ation of classifier inputs that belong to classes *not* present in the classifie
r's training data, which is assumed to be a random sample from that unknown popu
lation.
We find that *in expectation* the missing mass is entirely determined by the num
ber f_k of classes that *do* appear in the training data the same number of ti
mes *and an exponentially decaying error*.
While this is the first precise characterization of the expected missing mass in
 terms of the sample, the induced estimator suffers from an impractically high v
ariance. However, our theory suggests a large search space of nearly unbiased es
timators that can be searched effectively and efficiently. Hence, we cast distri
bution-free estimation as an optimization problem to find a distribution-specifi
c estimator with a minimized mean-squared error (MSE), given only the sample.
In our experiments, our search algorithm discovers estimators that have a substa
ntially smaller MSE than the state-of-the-art Good-Turing estimator. This holds
for over 93\% of runs when there are at least as many samples as classes. Our es
timators' MSE is roughly 80\% of the Good-Turing estimator's.
****
Qi Sun,Zhiyang Guo,Ziyu Wan,Jing Nathan Yan,Shengming Yin,Wengang Zhou,Jing Liao
,Houqiang Li
EG4D: Explicit Generation of 4D Object without Score Distillation
In recent years, the increasing demand for dynamic 3D assets in design and gamin
g applications has given rise to powerful generative pipelines capable of synthe
sizing high-quality 4D objects.
  Previous methods generally rely on score distillation sampling (SDS) algorithm
 to infer the unseen views and motion of 4D objects, thus leading to unsatisfact
ory results with defects like over-saturation and Janus problem.
  Therefore, inspired by recent progress of video diffusion models, we propose t
o optimize a 4D representation by explicitly generating multi-view videos from o
ne input image.
  However, it is far from trivial to handle practical challenges faced by such a
 pipeline, including dramatic temporal inconsistency, inter-frame geometry and t
exture diversity, and semantic defects brought by video generation results.
  To address these issues, we propose EG4D, a novel multi-stage framework that g

enerates high-quality and consistent 4D assets without score distillation.
  Specifically, collaborative techniques and solutions are developed, including an attention injection strategy to synthesize temporal-consistent multi-view vid eos, a robust and efficient dynamic reconstruction method based on Gaussian Spla tting, and a refinement stage with diffusion prior for semantic restoration.
  The qualitative comparisons and quantitative results demonstrate that our fram ework outperforms the baselines in generation quality by a considerable margin.
****
Ofir Schlisselberg,Ran Darshan
The impact of allocation strategies in subset learning on the expressive power o f neural networks
In traditional machine learning, models are defined by a set of parameters, whic h are optimized to perform specific tasks. In neural networks, these parameters correspond to the synaptic weights. However, in reality, it is often infeasible to control or update all weights. This challenge is not limited to artificial ne tworks but extends to biological networks, such as the brain, where the extent o f distributed synaptic weight modification during learning remains unclear. Moti vated by these insights, we theoretically investigate how different allocations of a fixed number of learnable weights influence the capacity of neural networks . Using a teacher-student setup, we introduce a benchmark to quantify the expres sivity associated with each allocation. We establish conditions under which allo cations have \`maximal' or \`minimal' expressive power in linear recurrent neura l networks and linear multi-layer feedforward networks. For suboptimal allocatio ns, we propose heuristic principles to estimate their expressivity. These princi ples extend to shallow ReLU networks as well. Finally, we validate our theoretic al findings with empirical experiments. Our results emphasize the critical role of strategically distributing learnable weights across the network, showing that a more widespread allocation generally enhances the network’s expressive power.
****
Mianchu Wang,Yue Jin,Giovanni Montana
Learning on One Mode: Addressing Multi-modality in Offline Reinforcement Learnin g
Offline reinforcement learning (RL) seeks to learn optimal policies from static datasets without interacting with the environment. A common challenge is handlin g multi-modal action distributions, where multiple behaviours are represented in the data. Existing methods often assume unimodal behaviour policies, leading to suboptimal performance when this assumption is violated. We propose weighted im itation Learning on One Mode (LOM), a novel approach that focuses on learning fr om a single, promising mode of the behaviour policy. By using a Gaussian mixture model to identify modes and selecting the best mode based on expected returns, LOM avoids the pitfalls of averaging over conflicting actions. Theoretically, we show that LOM improves performance while maintaining simplicity in policy learn ing. Empirically, LOM outperforms existing methods on standard D4RL benchmarks a nd demonstrates its effectiveness in complex, multi-modal scenarios.
****
Changho Shin,John Cooper,Frederic Sala
Weak-to-Strong Generalization Through the Data-Centric Lens
The weak-to-strong generalization phenomenon is the driver for important machine learning applications including highly data-efficient learning and, most recent ly, performing superalignment. While decades of research have resulted in numero us algorithms that produce strong empirical performance, understanding what aspe cts of data enable weak-to-strong generalization has been understudied. We propo se a simple data-centric mechanism that characterizes weak-to-strong generalizat ion: the overlap density. Intuitively, generalization tracks the number of point s that contain overlaps, i.e., both easy patterns (learnable by a weak model) an d challenging patterns (only learnable by a stronger model), as with such points , weak predictions can be used to learn challenging patterns by stronger models. And, we provide a practical overlap detection algorithm to find overlap density from data. Finally, we provide an algorithm to learn, among multiple sources of data, which to query when seeking to maximize overlap density and thereby enhan

ce weak-to-strong generalization. We provide a theoretical result showing that t
he generalization benefit is a function of the overlap density and a regret boun
d of our data selection algorithm. Empirically, we validate the mechanism and th
e overlap detection algorithm on a wide array of settings.
****
Lawrence Keunho Jang,Yinheng Li,Dan Zhao,Charles Ding,Justin Lin,Paul Pu Liang,R
ogerio Bonatti,Kazuhito Koishida
VideoWebArena: Evaluating Long Context Multimodal Agents with Video Understandi
ng Web Tasks
Videos are often used to learn or extract the necessary information to complete
tasks in ways different than what text or static imagery can provide. However, m
any
existing agent benchmarks neglect long-context video understanding, instead focu
s-
ing on text or static image inputs. To bridge this gap, we introduce VideoWebAre
na
(VideoWA), a benchmark for evaluating the capabilities of long-context multimoda
l
agents for video understanding. VideoWA consists of 2,021 web agent tasks based
on manually crafted video tutorials, which total almost four hours of content. F
or
our benchmark, we define a taxonomy of long-context video-based agent tasks with
two main areas of focus: skill retention and factual retention. While skill rete
ntion
tasks evaluate whether an agent can use a given human demonstration to complete
a task efficiently, the factual retention task evaluates whether an agent can re
trieve
instruction-relevant information from a video to complete a task. We find that t
he
best model achieves a 13.3% success rate on factual retention tasks and 45.8% on
factual retention QA pairs—far below human success rates of 73.9% and 79.3%,
respectively. On skill retention tasks, long-context models perform worse with
tutorials than without, exhibiting a 5% performance decrease in WebArena tasks
and a 10.3% decrease in VisualWebArena tasks. Our work highlights performance
gaps in the agentic abilities of long-context multimodal models and provides as
a
testbed for the future development of long-context video agents.
****
Jitai Hao,Yuke Zhu,Tian Wang,Jun Yu,Xin Xin,Bo Zheng,Zhaochun Ren,Sheng Guo
OmniKV: Dynamic Context Selection for Efficient Long-Context LLMs
During the inference phase of Large Language Models (LLMs) with long context, a
substantial portion of GPU memory is allocated to the KV cache, with memory usag
e increasing as the sequence length grows. To mitigate the GPU memory footprint
associate with KV cache, some previous studies have discarded less important tok
ens based on the sparsity identified in attention scores in long context scenari
os. However, we argue that attention scores cannot indicate the future importanc
e of tokens in subsequent generation iterations, because attention scores are ca
lculated based on current hidden states. Therefore, we propose OmniKV, a token-d
ropping-free and training-free inference method, which achieves a 1.68x speedup
without any loss in performance. It is well-suited for offloading, significantly
 reducing KV cache memory usage by up to 75% with it. The core innovative insigh
t of OmniKV is: Within a single generation iteration, there is a high degree of
similarity in the important tokens identified across consecutive layers. Extensi
ve experiments demonstrate that OmniKV achieves state-of-the-art performance acr
oss multiple benchmarks, with particularly advantages in chain-of-thoughts scena
rios. OmniKV extends the maximum context length supported by a single A100 for L
lama-3-8B from 128K to 450K. Our code is available at https://github.com/antgrou
p/OmniKV.git.
****
Aiwei Liu,Sheng Guan,Yiming Liu,Leyi Pan,Yifei Zhang,Liancheng Fang,Lijie Wen,Ph

ilip S. Yu,Xuming Hu
Can Watermarked LLMs be Identified by Users via Crafted Prompts?
Text watermarking for Large Language Models (LLMs) has made significant progress in detecting LLM outputs and preventing misuse. Current watermarking techniques offer high detectability, minimal impact on text quality, and robustness to text editing.

However, current researches lack investigation into the imperceptibility of watermarking techniques in LLM services.

This is crucial as LLM providers may not want to disclose the presence of watermarks in real-world scenarios, as it could reduce user willingness to use the service and make watermarks more vulnerable to attacks. This work is the first to investigate the imperceptibility of watermarked LLMs. We design an identification algorithm called Water-Probe that detects watermarks through well-designed prompts to the LLM. Our key motivation is that current watermarked LLMs expose consistent biases under the same watermark key, resulting in similar differences across prompts under different watermark keys. Experiments show that almost all mainstream watermarking algorithms are easily identified with our well-designed prompts,

while Water-Probe demonstrates a minimal false positive rate for non-watermarked LLMs.

Finally, we propose that the key to enhancing the imperceptibility of watermarked LLMs is to increase the randomness of watermark key selection. Based on this, we introduce the Water-Bag strategy, which significantly improves watermark imperceptibility by merging multiple watermark keys.
****
Zi Wang,Divyam Anshumaan,Ashish Hooda,Yudong Chen,Somesh Jha
Functional Homotopy: Smoothing Discrete Optimization via Continuous Parameters for LLM Jailbreak Attacks
Optimization methods are widely employed in deep learning to address and mitigate undesired model responses. While gradient-based techniques have proven effective for image models, their application to language models is hindered by the discrete nature of the input space. This study introduces a novel optimization approach, termed the *functional homotopy* method, which leverages the functional duality between model training and input generation. By constructing a series of easy-to-hard optimization problems, we iteratively solve these using principles derived from established homotopy methods. We apply this approach to jailbreak attack synthesis for large language models (LLMs), achieving a 20%-30% improvement in success rate over existing methods in circumventing established safe open-source models such as Llama-2 and Llama-3.
****
Leixin Zhang,Steffen Eger,Yinjie Cheng,WEIHE ZHAI,Jonas Belouadi,Fahimeh Moafian,Zhixue Zhao
ScImage: How good are multimodal large language models at scientific text-to-image generation?
Multimodal large language models (LLMs) have demonstrated impressive capabilities in generating high-quality images from textual instructions. However, their performance in generating scientific images—a critical application for accelerating scientific progress—remains underexplored. In this work, we address this gap by introducing ScImage, a benchmark designed to evaluate the multimodal capabilities of LLMs in generating scientific images from textual descriptions. ScImage assesses three key dimensions of understanding: spatial, numeric, and attribute comprehension, as well as their combinations, focusing on the relationships between scientific objects (e.g., squares, circles). We evaluate seven models, GPT-4o, Llama, AutomaTikZ, Dall-E, StableDiffusion, GPT-o1 and Qwen2.5-Coder-Instruct using two modes of output generation: code-based outputs (Python, TikZ) and direct raster image generation. Additionally, we examine four different input languages: English, German, Farsi, and Chinese. Our evaluation, conducted with 11 scientists across three criteria (correctness, relevance, and scientific accuracy), reveals that while GPT4-o produces outputs of decent quality for simpler prompts involving individual dimensions such as spatial, numeric, or attribute understanding

nding in isolation, all models face challenges in this task, especially for more complex prompts. ScImage is available: huggingface.co/datasets/casszhao/ScImage
****
Jenelle Feather,David Lipshutz,Sarah E Harvey,Alex H Williams,Eero P Simoncelli
Discriminating image representations with principal distortions
Image representations (artificial or biological) are often compared in terms of their global geometric structure; however, representations with similar global structure can have strikingly different local geometries. Here, we propose a framework for comparing a set of image representations in terms of their local geometries. We quantify the local geometry of a representation using the Fisher information matrix, a standard statistical tool for characterizing the sensitivity to local stimulus distortions, and use this as a substrate for a metric on the local geometry in the vicinity of a base image. This metric may then be used to optimally differentiate a set of models, by finding a pair of "principal distortions" that maximize the variance of the models under this metric. As an example, we use this framework to compare a set of simple models of the early visual system, identifying a novel set of image distortions that allow immediate comparison of the models by visual inspection. In a second example, we apply our method to a set of deep neural network models and reveal differences in the local geometry that arise due to architecture and training types. These examples demonstrate how our framework can be used to probe for informative differences in local sensitivities between complex models, and suggest how it could be used to compare model representations with human perception.
****
Yichi Zhang,Zhuo Chen,Lingbing Guo,yajing Xu,Binbin Hu,Ziqi Liu,Wen Zhang,Huajun Chen
Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning
Learning high-quality multi-modal entity representations is an important goal of multi-modal knowledge graph (MMKG) representation learning, which can en- hance reasoning tasks within the MMKGs, such as MMKG completion (MMKGC). The main challenge is to collaboratively model the structural information concealed in massive triples and the multi-modal features of the entities. Existing methods focus on crafting elegant entity-wise multi-modal fusion strategies, yet they over- look the utilization of multi-perspective features concealed within the modalities under diverse relational contexts. To address this issue, we introduce a novel framework with Mixture of Modality Knowledge experts (MOMOK for short) to learn adaptive multi-modal entity representations for better MMKGC. We design relation-guided modality knowledge experts to acquire relation-aware modality embeddings and integrate the predictions from multi-modalities to achieve joint decisions. Additionally, we disentangle the experts by minimizing their mutual information. Experiments on four public MMKG benchmarks demonstrate the outstanding performance of MOMOK under complex scenarios. Our code and data are available at https://github.com/zjukg/MoMoK.
****
Tian Jin,Ahmed Imtiaz Humayun,Utku Evci,Suvinay Subramanian,Amir Yazdanbakhsh,Dan Alistarh,Gintare Karolina Dziugaite
The Journey Matters: Average Parameter Count over Pre-training Unifies Sparse and Dense Scaling Laws
Pruning eliminates unnecessary parameters in neural networks; it offers a promising solution to the growing computational demands of large language models (LLMs).
While many focus on post-training pruning, sparse pre-training--which combines pruning and pre-training into a single phase--provides a simpler alternative.
In this work, we present the first systematic exploration of optimal sparse pre-training configurations for LLMs through an examination of 80 unique pruning schedules across different sparsity levels and training durations.
We find that initiating pruning at 25\% of total training compute and concluding at 75\% achieves near-optimal final evaluation loss.
These findings provide valuable insights for efficient and effective sparse pre-

training of LLMs.
Furthermore, we propose a new scaling law that modifies the Chinchilla scaling law to use the average parameter count over pre-training.
Through empirical and theoretical validation, we demonstrate that this modified scaling law accurately models evaluation loss for both sparsely and densely pre-trained LLMs, unifying scaling laws across pre-training paradigms.
Our findings indicate that while sparse pre-training achieves the same final model quality as dense pre-training for equivalent compute budgets, it provides substantial benefits through reduced model size, enabling significant potential computational savings during inference.
****
Noam Razin,Sadhika Malladi,Adithya Bhaskar,Danqi Chen,Sanjeev Arora,Boris Hanin
Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization
Direct Preference Optimization (DPO) and its variants are increasingly used for aligning language models with human preferences.
Although these methods are designed to teach a model to generate preferred responses more frequently relative to dispreferred responses, prior work has observed that the likelihood of preferred responses often decreases during training. The current work sheds light on the causes and implications of this counterintuitive phenomenon, which we term *likelihood displacement*. We demonstrate that likelihood displacement can be *catastrophic*, shifting probability mass from preferred responses to responses with an opposite meaning. As a simple example, training a model to prefer $\texttt{No}$ over $\texttt{Never}$ can sharply increase the probability of $\texttt{Yes}$. Moreover, when aligning the model to refuse unsafe prompts, we show that such displacement can *unintentionally lead to unalignment*, by shifting probability mass from preferred refusal responses to harmful responses (e.g., reducing the refusal rate of Llama-3-8B-Instruct from 74.4% to 33.4%). We theoretically characterize that likelihood displacement is driven by preferences that induce similar embeddings, as measured by a *centered hidden embedding similarity (CHES)* score. Empirically, the CHES score enables identifying which training samples contribute most to likelihood displacement in a given dataset. Filtering out these samples effectively mitigated unintentional unalignment in our experiments. More broadly, our results highlight the importance of curating data with sufficiently distinct preferences, for which we believe the CHES score may prove valuable.
****
Mostafa Karimi,Sharmi Banerjee,Tommi Jaakkola,Bella Dubrov,Shang Shang,Ron Benson
Data Distillation for extrapolative protein design through exact preference optimization
The goal of protein design typically involves increasing fitness (extrapolating) beyond what is seen during training (e.g., towards higher stability, stronger binding affinity, etc.). State-of-the-art methods assume that one can safely steer proteins towards such extrapolated regions by learning from pairs alone. We hypothesize that noisy training pairs are not sufficiently informative to capture the fitness gradient and that models learned from pairs specifically may fail to capture three-way relations important for search, e.g., how two alternatives fair relative to a seed. Building on the success of preference alignment models in large language models, we introduce a progressive search method for extrapolative protein design by directly distilling into the model relevant triplet relations. We evaluated our model's performance in designing AAV and GFP proteins and demonstrated that the proposed framework significantly improves effectiveness in extrapolation tasks.
****
Ángela López-Cardona,Carlos Segura,Alexandros Karatzoglou,Sergi Abadal,Ioannis Arapakis
Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models
Advancements in Natural Language Processing (NLP), have led to the emergence of

Large Language Models (LLMs) such as GPT, Llama, Claude, and Gemini, which excel across a range of tasks but require extensive fine-tuning to align their outputs with human expectations. A widely used method for achieving this alignment is Reinforcement Learning from Human Feedback (RLHF), which, despite its success, faces challenges in accurately modelling human preferences. In this paper, we introduce GazeReward, a novel framework that integrates implicit feedback -- and specifically eye-tracking (ET) data -- into the Reward Model (RM). In addition, we explore how ET-based features can provide insights into user preferences. Through ablation studies we test our framework with different integration methods, LLMs, and ET generator models, demonstrating that our approach significantly improves the accuracy of the RM on established human preference datasets. This work advances the ongoing discussion on optimizing AI alignment with human values, exploring the potential of cognitive data for shaping future NLP research.
****
Abdelhakim Benechehab,Youssef Attia El Hili,Ambroise Odonnat,Oussama Zekri,Albert Thomas,Giuseppe Paolo,Maurizio Filippone,Ievgen Redko,Balázs Kégl
Zero-shot Model-based Reinforcement Learning using Large Language Models
The emerging zero-shot capabilities of Large Language Models (LLMs) have led to their applications in areas extending well beyond natural language processing tasks.
In reinforcement learning, while LLMs have been extensively used in text-based environments, their integration with continuous state spaces remains understudied.
In this paper, we investigate how pre-trained LLMs can be leveraged to predict in context the dynamics of continuous Markov decision processes.
We identify handling multivariate data and incorporating the control signal as key challenges that limit the potential of LLMs' deployment in this setup and propose Disentangled In-Context Learning (DICL) to address them.
We present proof-of-concept applications in two reinforcement learning settings: model-based policy evaluation and data-augmented off-policy reinforcement learning, supported by theoretical analysis of the proposed methods.
Our experiments further demonstrate that our approach produces well-calibrated uncertainty estimates. We release the code at https://github.com/abenechehab/dicl.
****
Justin Deschenaux,Caglar Gulcehre
Beyond Autoregression: Fast LLMs via Self-Distillation Through Time
Autoregressive (AR) Large Language Models (LLMs) have demonstrated significant success across numerous tasks. However, the AR modeling paradigm presents certain limitations; for instance, contemporary autoregressive LLMs are trained to generate one token at a time, which can result in noticeable latency. Recent advances have indicated that search and repeated sampling can enhance performance in various applications, such as theorem proving, code generation, and alignment, by utilizing greater computational resources during inference. In this study, we demonstrate that diffusion language models are capable of generating at least 32 tokens simultaneously, while exceeding the performance of AR models in text quality and on the LAMBADA natural language understanding benchmark. This outcome is achieved through a novel distillation method for discrete diffusion models, which reduces the number of inference steps by a factor of 32-64. Practically, at the 1.3B parameters scale, diffusion models, even without caching, can generate tokens at a rate that is up to 8 times faster than AR models employing KV-caching, and we anticipate further improvements with the inclusion of caching. Moreover, we demonstrate the efficacy of our approach for diffusion language models with up to 860M parameters.
****
Fangxun Shu,Yue Liao,Lei Zhang,Le Zhuo,Chenning Xu,Guanghao Zhang,Haonan Shi,Long Chan,TaoZhong,Zhelun Yu,Wanggui He,Siming Fu,Haoyuan Li,Si Liu,Hongsheng Li,Hao Jiang
LLaVA-MoD: Making LLaVA Tiny via MoE-Knowledge Distillation
We introduce LLaVA-MoD, a novel framework designed to enable the efficient train

ing of small-scale Multimodal Language Models (s-MLLM) distilling knowledge fr
om large-scale MLLM (l-MLLM). Our approach tackles two fundamental challenges
in MLLM distillation. First, we optimize the network structure of s-MLLM by in
tegrating a sparse Mixture of Experts (MoE) architecture into the language model
, striking a balance between computational efficiency and model expressiveness.
Second, we propose a progressive knowledge transfer strategy for comprehensive k
nowledge transfer. This strategy begins with mimic distillation, where we minimi
ze the Kullback-Leibler (KL) divergence between output distributions to enable $
s$-MLLM to emulate s-MLLM's understanding. Following this, we introduce prefer
ence distillation via Preference Optimization (PO), where the key lies in treati
ng l-MLLM as the reference model. During this phase, the s-MLLM's ability to
 discriminate between superior and inferior examples is significantly enhanced b
eyond l-MLLM, leading to a better s-MLLM that surpasses l-MLLM, particular
ly in hallucination benchmarks.
Extensive experiments demonstrate that LLaVA-MoD surpasses existing works across
 various benchmarks while maintaining a minimal activated parameters and low com
putational costs. Remarkably, LLaVA-MoD-2B surpasses Qwen-VL-Chat-7B with an ave
rage gain of 8.8\%, using merely 0.3% of the training data and 23\% trainable
 parameters. The results underscore LLaVA-MoD's ability to effectively distill c
omprehensive knowledge from its teacher model, paving the way for developing eff
icient MLLMs.
****
Shumao Zhai,Ning Ge
Learning Splitting Heuristics in Divide-and-Conquer SAT Solvers with Reinforceme
nt Learning
We propose RDC-SAT, a novel approach to optimize splitting heuristics in Divide-
and-Conquer SAT solvers using deep reinforcement learning. Our method dynamical
ly extracts features from the current solving state whenever a split is required
. These features, such as learned clauses, variable activity scores, and clause
 LBD (Literal Block Distance) values, are represented as a graph. A GNN integra
ted with an Actor-Critic model processes this graph to determine the optimal spl
it variable. Unlike traditional linear state transitions characterized by Marko
v processes, divide-and-conquer challenges involve tree-like state transitions.
 To address this, we developed a reinforcement learning environment based on the
 Painless framework that efficiently handles these transitions. Additionally, w
e designed different discounted reward functions for satisfiable and unsatisfiab
le SAT problems, capable of handling tree-like state transitions. We trained ou
r model using the Decentralized Proximal Policy Optimization (DPPO) algorithm on
 phase transition random 3-SAT problems and implemented the RDC-SAT solver, whic
h operates in both GPU-accelerated and non-GPU modes. Evaluations show that RDC
-SAT significantly improves the performance of D\&C solvers on phase transition
random 3-SAT datasets and generalizes well to the SAT Competition 2023 dataset,
substantially outperforming traditional splitting heuristics.
****
Jiawen Qin,Haonan Yuan,Qingyun Sun,Lyujin Xu,Jiaqi Yuan,Pengfeng Huang,Zhaonan W
ang,Xingcheng Fu,Hao Peng,Jianxin Li,Philip S. Yu
IGL-Bench: Establishing the Comprehensive Benchmark for Imbalanced Graph Learnin
g
Deep graph learning has gained grand popularity over the past years due to its v
ersatility and success in representing graph data across a wide range of domains
. However, the pervasive issue of imbalanced graph data distributions, where cer
tain parts exhibit disproportionally abundant data while others remain sparse, u
ndermines the efficacy of conventional graph learning algorithms, leading to bia
sed outcomes. To address this challenge, Imbalanced Graph Learning (IGL) has gar
nered substantial attention, enabling more balanced data distributions and bette
r task performance. Despite the proliferation of IGL algorithms, the absence of
consistent experimental protocols and fair performance comparisons pose a signif
icant barrier to comprehending advancements in this field. To bridge this gap, w
e introduce **IGL-Bench**, a foundational comprehensive benchmark for imbalanced
 graph learning, embarking on **17** diverse graph datasets and **24** distinct

IGL algorithms with uniform data processing and splitting strategies. Specifically, IGL-Bench systematically investigates state-of-the-art IGL algorithms in terms of **effectiveness**, **robustness**, and **efficiency** on node-level and graph-level tasks, with the scope of class-imbalance and topology-imbalance. Extensive experiments demonstrate the potential benefits of IGL algorithms on various imbalanced conditions, offering insights and opportunities in the IGL field. Further, we have developed an open-sourced and unified package to facilitate reproducible evaluation and inspire further innovative research, available at: https://github.com/RingBDStack/IGL-Bench.
****
Toshimitsu Uesaka,Taiji Suzuki,Yuhta Takida,Chieh-Hsin Lai,Naoki Murata,Yuki Mitsufuji
Weighted Point Set Embedding for Multimodal Contrastive Learning Toward Optimal Similarity Metric
In typical multimodal contrastive learning, such as CLIP, encoders produce one point in the latent representation space for each input. However, one-point representation has difficulty in capturing the relationship and the similarity structure of a huge amount of instances in the real world. For richer classes of the similarity, we propose the use of weighted point sets, namely, sets of pairs of weight and vector, as representations of instances. In this work, we theoretically show the benefit of our proposed method through a new understanding of the contrastive loss of CLIP, which we call symmetric InfoNCE. We clarify that the optimal similarity that minimizes symmetric InfoNCE is the pointwise mutual information, and show an upper bound of excess risk on downstream classification tasks of representations that achieve the optimal similarity. In addition, we show that our proposed similarity based on weighted point sets consistently achieves the optimal similarity. To verify the effectiveness of our proposed method, we demonstrate pretraining of text-image representation models and classification tasks on common benchmarks.
****
Gennadiy Averkov,Christopher Hojny,Maximilian Merkert
On the Expressiveness of Rational ReLU Neural Networks With Bounded Depth
To confirm that the expressive power of ReLU neural networks grows with their depth, the function $F_n = \max (0,x_1,\ldots,x_n)$ has been considered in the literature.

A conjecture by Hertrich, Basu, Di Summa, and Skutella [NeurIPS 2021] states that any ReLU network that exactly represents F_n has at least $\lceil \log_2 (n+1) \rceil$ hidden layers.

The conjecture has recently been confirmed for networks with integer weights by Haase, Hertrich, and Loho [ICLR 2023].

We follow up on this line of research and show that, within ReLU networks whose weights are decimal fractions, F_n can only be represented by networks with at least $\lceil \log_3 (n+1) \rceil$ hidden layers.

Moreover, if all weights are N-ary fractions, then F_n can only be represented by networks with at least $\Omega(\frac{\ln n}{\ln \ln N})$ layers.

These results are a partial confirmation of the above conjecture for rational ReLU networks, and provide the first non-constant lower bound on the depth of practically relevant ReLU networks.
****
Shaorong Chen,Jun Xia,Jingbo Zhou,Lecheng Zhang,Zhangyang Gao,Bozhen Hu,Cheng Tan,Wenjie Du,Stan Z. Li
ReNovo: Retrieval-Based \emph{De Novo} Mass Spectrometry Peptide Sequencing
Proteomics is the large-scale study of proteins. Tandem mass spectrometry, as the only high-throughput technique for protein sequence identification, plays a pi

votal role in proteomics research. One of the long-standing challenges in this f
ield is peptide identification, which entails determining the specific peptide (
sequence of amino acids) that corresponds to each observed mass spectrum. The co
nventional approach involves database searching, wherein the observed mass spect
rum is scored against a pre-constructed peptide database. However, the reliance
on pre-existing databases limits applicability in scenarios where the peptide is
 absent from existing databases. Such circumstances necessitate \emph{de novo} p
eptide sequencing, which derives peptide sequence solely from input mass spectru
m, independent of any peptide database. Despite ongoing advancements in \emph{de
 novo} peptide sequencing, its performance still has considerable room for impro
vement, which limits its application in large-scale experiments. In this study,
we introduce a novel \textbf{Re}trieval-based \emph{De \textbf{Novo}} peptide se
quencing methodology, termed \textbf{ReNovo}, which draws inspiration from datab
ase search methods. Specifically, by constructing a datastore from training data
, ReNovo can retrieve information from the datastore during the inference stage
to conduct retrieval-based inference, thereby achieving improved performance. Th
is innovative approach enables ReNovo to effectively combine the strengths of bo
th methods: utilizing the assistance of the datastore while also being capable o
f predicting novel peptides that are not present in pre-existing databases. A se
ries of experiments have confirmed that ReNovo outperforms state-of-the-art mode
ls across multiple widely-used datasets, incurring only minor storage and time c
onsumption, representing a significant advancement in proteomics. Supplementary
materials include the code.
****

Koichi Namekata,Sherwin Bahmani,Ziyi Wu,Yash Kant,Igor Gilitschenski,David B. Li
ndell
SG-I2V: Self-Guided Trajectory Control in Image-to-Video Generation
Methods for image-to-video generation have achieved impressive, photo-realistic
quality.
However, adjusting specific elements in generated videos, such as object motion
or camera movement, is often a tedious process of trial and error, e.g., involvi
ng re-generating videos with different random seeds.
Recent techniques address this issue by fine-tuning a pre-trained model to follo
w conditioning signals, such as bounding boxes or point trajectories.
Yet, this fine-tuning procedure can be computationally expensive, and it require
s datasets with annotated object motion, which can be difficult to procure.
In this work, we introduce SG-I2V, a framework for controllable image-to-video g
eneration that is self-guided—offering zero-shot control by relying solely on th
e knowledge present in a pre-trained image-to-video diffusion model without the
need for fine-tuning or external knowledge.
Our zero-shot method outperforms unsupervised baselines while significantly narr
owing down the performance gap with supervised models in terms of visual quality
 and motion fidelity.
Additional details and video results are available on our project page: https://
kmcode1.github.io/Projects/SG-I2V
****

Yunfei Teng,Yuxuan Ren,Kai Chen,Xi Chen,Zhaoming Chen,Qiwei Ye
CryoGEN: Generative Energy-based Models for Cryogenic Electron Tomography Recons
truction
Cryogenic electron tomography (Cryo-ET) is a powerful technique for visualizing
subcellular structures in their native states. Nonetheless, its effectiveness is
 compromised by anisotropic resolution artifacts caused by the missing-wedge eff
ect. To address this, IsoNet, a deep learning-based method, proposes iteratively
 reconstructing the missing-wedge information. While successful, IsoNet's depend
ence on recursive prediction updates often leads to training instability and mod
el divergence. In this study, we introduce CryoGEN—an energy-based probabilistic
 model that not only mitigates resolution anisotropy but also removes the need f
or recursive subtomogram averaging, delivering an approximate *10*$\times$ speed
up for training. Evaluations across various biological datasets, including immat
ure HIV-1 virions and ribosomes, demonstrate that CryoGEN significantly enhances

structural completeness and interpretability of the reconstructed samples.
****
Mohamed Amine Ketata,Nicholas Gao,Johanna Sommer,Tom Wollschläger,Stephan Günnemann
Lift Your Molecules: Molecular Graph Generation in Latent Euclidean Space
We introduce a new framework for 2D molecular graph generation using 3D molecule generative models. Our Synthetic Coordinate Embedding (SyCo) framework maps 2D molecular graphs to 3D Euclidean point clouds via synthetic coordinates and learns the inverse map using an E(n)-Equivariant Graph Neural Network (EGNN). The induced point cloud-structured latent space is well-suited to apply existing 3D molecule generative models. This approach simplifies the graph generation problem into a point cloud generation problem followed by node and edge classification tasks, without relying on molecular fragments nor autoregressive decoding. Further, we propose a novel similarity-constrained optimization scheme for 3D diffusion models based on inpainting and guidance. As a concrete implementation of our framework, we develop EDM-SyCo based on the E(3) Equivariant Diffusion Model (EDM). EDM-SyCo achieves state-of-the-art performance in distribution learning of molecular graphs, outperforming the best non-autoregressive methods by more than 26\% on ZINC250K and 16\% on the GuacaMol dataset while improving conditional generation by up to 3.9 times.
****
Jiachen Tu,Yaokun Shi,Fan Lam
Score-based Self-supervised MRI Denoising
Magnetic resonance imaging (MRI) is a powerful noninvasive diagnostic imaging tool that provides unparalleled soft tissue contrast and anatomical detail. Noise contamination, especially in accelerated and/or low-field acquisitions, can significantly degrade image quality and diagnostic accuracy. Supervised learning based denoising approaches have achieved impressive performance but require high signal-to-noise ratio (SNR) labels, which are often unavailable. Self-supervised learning holds promise to address the label scarcity issue, but existing self-supervised denoising methods tend to oversmooth fine spatial features and often yield inferior performance than supervised methods. We introduce Corruption2Self (C2S), a novel score-based self-supervised framework for MRI denoising. At the core of C2S is a generalized denoising score matching (GDSM) loss, which extends denoising score matching to work directly with noisy observations by modeling the conditional expectation of higher-SNR images given further corrupted observations. This allows the model to effectively learn denoising across multiple noise levels directly from noisy data. Additionally, we incorporate a reparameterization of noise levels to stabilize training and enhance convergence, and introduce a detail refinement extension to balance noise reduction with the preservation of fine spatial features. Moreover, C2S can be extended to multi-contrast denoising by leveraging complementary information across different MRI contrasts. We demonstrate that our method achieves state-of-the-art performance among self-supervised methods and competitive results compared to supervised counterparts across varying noise conditions and MRI contrasts on the M4Raw and fastMRI dataset. The project website is available at: https://jiachentu.github.io/Corruption2Self-Self-Supervised-Denoising/.
****
Cunxiang Wang,Ruoxi Ning,Boqi Pan,Tonghui Wu,Qipeng Guo,Cheng Deng,Guangsheng Bao,Xiangkun Hu,Zheng Zhang,Qian Wang,Yue Zhang
NovelQA: Benchmarking Question Answering on Documents Exceeding 200K Tokens
Recent advancements in Large Language Models (LLMs) have pushed the boundaries of natural language processing, especially in long-context understanding. However, the evaluation of these models' long-context abilities remains a challenge due to the limitations of current benchmarks. To address this gap, we introduce NovelQA, a benchmark tailored for evaluating LLMs with complex, extended narratives. NovelQA, constructed from English novels, offers a unique blend of complexity, length, and narrative coherence, making it an ideal tool for assessing deep textual understanding in LLMs. This paper details the design and construction of NovelQA, focusing on its comprehensive manual annotation process and the variety o

f question types aimed at evaluating nuanced comprehension. Our evaluation of l
ong-context LLMs on NovelQA reveals significant insights into their strengths an
d weaknesses. Notably, the models struggle with multi-hop reasoning, detail-orie
nted questions, and handling extremely long inputs, averaging over 200,000 token
s. Results highlight the need for substantial advancements in LLMs to enhance th
eir long-context comprehension and contribute effectively to computational liter
ary analysis.
****

Anthony Gruber,Kookjin Lee,Haksoo Lim,Noseong Park,Nathaniel Trask
Efficiently Parameterized Neural Metriplectic Systems
Metriplectic systems are learned from data in a way that scales quadratically in
 both the size of the state and the rank of the metriplectic operators. In addi
tion to being provably energy-conserving and entropy-stable, the proposed neural
 metriplectic systems (NMS) approach includes approximation results that demonst
rate its ability to accurately learn metriplectic dynamics from data, along with
 an error estimate that indicates its potential for generalization to unseen tim
escales when the approximation error is low. Examples are provided to illustrate
 performance both with full state information available and when entropic variab
les are unknown, confirming that the NMS approach exhibits superior accuracy and
 scalability without compromising on model expressivity.
****

Mario Lino Valencia,Tobias Pfaff,Nils Thuerey
Learning Distributions of Complex Fluid Simulations with Diffusion Graph Network
s
Physical systems with complex unsteady dynamics, such as fluid flows, are often
poorly represented by a single mean solution. For many practical applications, i
t is crucial to access the full distribution of possible states, from which rele
vant statistics (e.g., RMS and two-point correlations) can be derived. Here, we
propose a graph-based latent diffusion model that enables direct sampling of sta
tes from their equilibrium distribution, given a mesh discretization of the syst
em and its physical parameters. This allows for the efficient computation of flo
w statistics without running long and expensive numerical simulations. The graph
-based structure enables operations on unstructured meshes, which is critical fo
r representing complex geometries with spatially localized high gradients, while
 latent-space diffusion modeling with a multi-scale GNN allows for efficient lea
rning and inference of entire distributions of solutions. A key finding of our w
ork is that the proposed networks can accurately learn full distributions even w
hen trained on incomplete data from relatively short simulations. We apply this
method to a range of fluid dynamics tasks, such as predicting pressure distribut
ions on 3D wing models in turbulent flow, demonstrating both accuracy and comput
ational efficiency in challenging scenarios. The ability to directly sample accu
rate solutions, and capturing their diversity from short ground-truth simulation
s, is highly promising for complex scientific modeling tasks.
****

Fanghua Yu,Jinjin Gu,Jinfan Hu,Zheyuan Li,Chao Dong
UniCon: Unidirectional Information Flow for Effective Control of Large-Scale Dif
fusion Models
We introduce UniCon, a novel architecture designed to enhance control and effici
ency in training adapters for large-scale diffusion models. Unlike existing meth
ods that rely on bidirectional interaction between the diffusion model and contr
ol adapter, UniCon implements a unidirectional flow from the diffusion network t
o the adapter, allowing the adapter alone to generate the final output. UniCon r
educes computational demands by eliminating the need for the diffusion model to
compute and store gradients during adapter training. Our results indicate that U
niCon reduces GPU memory usage by one-third and increases training speed by 2.3
times, while maintaining the same adapter parameter size. Additionally, without
requiring extra computational resources, UniCon enables the training of adapters
 with double the parameter volume of existing ControlNets. In a series of image
conditional generation tasks, UniCon has demonstrated precise responsiveness to
control inputs and exceptional generation capabilities.

****
Jongwoo Ko,Saket Dingliwal,Bhavana Ganesh,Sailik Sengupta,Sravan Babu Bodapati,Aram Galstyan
SeRA: Self-Reviewing and Alignment of LLMs using Implicit Reward Margins
Direct alignment algorithms (DAAs), such as direct preference optimization (DPO), have become popular alternatives to Reinforcement Learning from Human Feedback (RLHF) due to their simplicity, efficiency, and stability. However, the preferences used by DAAs are usually collected before alignment training begins and remain unchanged (off-policy). This design leads to two problems where the policy model (1) picks up on spurious correlations in the dataset (as opposed to only learning alignment to human preferences), and (2) overfits to feedback on off-policy trajectories that have less likelihood of being generated by the updated policy model. To address these issues, we introduce Self-Reviewing and Alignment (SeRA), a cost-efficient and effective method that can be readily combined with existing DAAs. SeRA comprises of two components: (1) sample selection using implicit reward margin to alleviate over-optimization on such undesired features, and (2) preference bootstrapping using implicit rewards to augment preference data with updated policy models in a cost-efficient manner. Extensive experiments, including on instruction-following tasks, demonstrate the effectiveness and generality of SeRA in training LLMs with diverse offline preference datasets and and DAAs.
****
Jiachen T. Wang,Dawn Song,James Zou,Prateek Mittal,Ruoxi Jia
Capturing the Temporal Dependence of Training Data Influence
Traditional data influence estimation methods, like influence function, assume that learning algorithms are permutation-invariant with respect to training data. However, modern training paradigms—especially for foundation models using stochastic algorithms and non-convergent, multi-stage curricula—are sensitive to data ordering, thus violating this assumption. This mismatch renders influence functions inadequate for answering some critical questions in current machine learning: How can we differentiate the influence of the same data contributing at different stages of training? More generally, how can we capture the dependence of data influence on the optimization trajectory during training? To address this gap, we formalize the concept of \emph{trajectory-specific leave-one-out (LOO) influence}, which quantifies the impact of removing a data point from a specific iteration during training, accounting for the exact sequence of data encountered and the model's optimization trajectory. However, exactly evaluating the trajectory-specific LOO presents a significant computational challenge. To address this, we propose \emph{data value embedding}, a novel technique enabling efficient approximation of trajectory-specific LOO. Specifically, we compute a training data embedding that encapsulates the cumulative interactions between data and the evolving model parameters. The LOO can then be efficiently approximated through a simple dot-product between the data value embedding and the gradient of the given test data. As data value embedding captures training data ordering, it offers valuable insights into model training dynamics. In particular, we uncover distinct phases of data influence, revealing that data points in the early and late stages of training exert a greater impact on the final model. These insights translate into actionable strategies for managing the computational overhead of data selection by strategically timing the selection process, potentially opening new avenues in data curation research.
****
Hyungi Lee,Chaeyun Jang,Dong Bok Lee,Juho Lee
Dimension Agnostic Neural Processes
Meta-learning aims to train models that can generalize to new tasks with limited labeled data by extracting shared features across diverse task datasets. Additionally, it accounts for prediction uncertainty during both training and evaluation, a concept known as uncertainty-aware meta-learning. Neural Process (NP) is a well-known uncertainty-aware meta-learning method that constructs implicit stochastic processes using parametric neural networks, enabling rapid adaptation to new tasks. However, existing NP methods face challenges in accommodating diverse

input dimensions and learned features, limiting their broad applicability acros s regression tasks. To address these limitations and advance the utility of NP m odels as general regressors, we introduce Dimension Agnostic Neural Process (DAN P). DANP incorporates Dimension Aggregator Block (DAB) to transform input featur es into a fixed-dimensional space, enhancing the model's ability to handle diver se datasets. Furthermore, leveraging the Transformer architecture and latent enc oding layers, DANP learns a wider range of features that are generalizable acros s various tasks. Through comprehensive experimentation on various synthetic and practical regression tasks, we empirically show that DANP outperforms previous N P variations, showcasing its effectiveness in overcoming the limitations of trad itional NP models and its potential for broader applicability in diverse regress ion scenarios.
****

Shengwei Xu,Yuxuan Lu,Grant Schoenebeck,Yuqing Kong
Benchmarking LLMs' Judgments with No Gold Standard
We introduce the GEM (Generative Estimator for Mutual Information), an evaluatio n metric for assessing language generation by large language models (LLMs), part icularly in generating informative judgments, without the need for a gold standa rd reference. GEM broadens the scenarios where we can benchmark LLM generation p erformance-from traditional ones, like machine translation and summarization, wh ere gold standard references are readily available, to subjective tasks without clear gold standards, such as academic peer review.

GEM uses a generative model to estimate mutual information between candidate and reference responses, without requiring the reference to be a gold standard. In experiments on two human-annotated datasets, GEM demonstrates competitive correl ations with human scores compared to the state-of-the-art GPT-4o Examiner, and o utperforms all other baselines. Additionally, GEM is more robust against strateg ic manipulation, such as rephrasing or elongation, which can artificially inflat e scores under a GPT-4o Examiner.

We also present GRE-bench (Generating Review Evaluation Benchmark) which evaluat es LLMs based on how well they can generate high-quality peer reviews for academ ic research papers. Because GRE-bench is based upon GEM, it inherits its robust ness properties. Additionally, GRE-bench circumvents data contamination problem s (or data leakage) by using the continuous influx of new open-access research p apers and peer reviews each year. We show GRE-bench results of various popular L LMs on their peer review capabilities using the ICLR2023 dataset.
****

Gathika Ratnayaka,James Nichols,Qing Wang
Learning Partial Graph Matching via Optimal Partial Transport
Partial graph matching extends traditional graph matching by allowing some nodes to remain unmatched, enabling applications in more complex scenarios. However, this flexibility introduces additional complexity, as both the subset of nodes t o match and the optimal mapping must be determined. While recent studies have ex plored deep learning techniques for partial graph matching, a significant limita tion remains: the absence of an optimization objective that fully captures the p roblem’s intrinsic nature while enabling efficient solutions. In this paper, we propose a novel optimization framework for partial graph matching, inspired by o ptimal partial transport. Our approach formulates an objective that enables part ial assignments while incorporating matching biases, using weighted total variat ion as the divergence function to guarantee optimal partial assignments. Our met hod can achieve efficient, exact solutions within cubic worst case time complexi ty. Our contributions are threefold: (i) we introduce a novel optimization objec tive that balances matched and unmatched nodes; (ii) we establish a connection b etween partial graph matching and linear sum assignment problem, enabling effici ent solutions; (iii) we propose a deep graph matching architecture with a novel partial matching loss, providing an end-to-end solution. The empirical evaluatio ns on standard graph matching benchmarks demonstrate the efficacy of the propose d approach.

****
Qingren Yao,Chao-Han Huck Yang,Renhe Jiang,Yuxuan Liang,Ming Jin,Shirui Pan
Towards Neural Scaling Laws for Time Series Foundation Models
Scaling laws offer valuable insights into the design of time series foundation m odels (TSFMs). However, previous research has largely focused on the scaling law s of TSFMs for in-distribution (ID) data, leaving their out-of-distribution (OOD) scaling behavior and the influence of model architectures less explored. In th is work, we examine two common TSFM architectures—encoder-only and decoder-only Transformers—and investigate their scaling behavior on both ID and OOD data. The se models are trained and evaluated across varying parameter counts, compute bud gets, and dataset sizes. Our experiments reveal that the log-likelihood loss of TSFMs exhibits similar scaling behavior in both OOD and ID settings. We further compare the scaling properties across different architectures, incorporating two state-of-the-art TSFMs as case studies, showing that model architecture plays a significant role in scaling. The encoder-only Transformers demonstrate better s calability than the decoder-only Transformers, while the architectural enhanceme nts in the two advanced TSFMs primarily improve ID performance but reduce OOD sc alability. While scaling up TSFMs is expected to drive performance breakthroughs , the lack of a comprehensive understanding of TSFM scaling laws has hindered th e development of a robust framework to guide model scaling. We fill this gap in this work by synthesizing our findings and providing practical guidelines for de signing and scaling larger TSFMs with enhanced model capabilities.
****
Manuel Gloeckler,Shoji Toyota,Kenji Fukumizu,Jakob H. Macke
Compositional simulation-based inference for time series
Amortized simulation-based inference (SBI) methods train neural networks on simu lated data to perform Bayesian inference. While this strategy avoids the need fo r tractable likelihoods, it often requires a large number of simulations and has been challenging to scale to time series data. Scientific simulators frequently emulate real-world dynamics through thousands of single-state transitions over time. We propose an SBI approach that can exploit such Markovian simulators by l ocally identifying parameters consistent with individual state transitions. We t hen compose these local results to obtain a posterior over parameters that align with the entire time series observation. We focus on applying this approach to neural posterior score estimation but also show how it can be applied, e.g., to neural likelihood (ratio) estimation. We demonstrate that our approach is more s imulation-efficient than directly estimating the global posterior on several syn thetic benchmark tasks and simulators used in ecology and epidemiology. Finally, we validate scalability and simulation efficiency of our approach by applying i t to a high-dimensional Kolmogorov flow simulator with around one million data d imensions.
****
Zhihui Xie,Jiahui Gao,Lei Li,Zhenguo Li,Qi Liu,Lingpeng Kong
Jailbreaking as a Reward Misspecification Problem
The widespread adoption of large language models (LLMs) has raised concerns abou t their safety and reliability, particularly regarding their vulnerability to ad versarial attacks. In this paper, we propose a new perspective that attributes t his vulnerability to reward misspecification during the alignment process. This misspecification occurs when the reward function fails to accurately capture the intended behavior, leading to misaligned model outputs. We introduce a metric R eGap to quantify the extent of reward misspecification and demonstrate its effec tiveness and robustness in detecting harmful backdoor prompts. Building upon the se insights, we present ReMiss, a system for automated red teaming that generate s adversarial prompts in a reward-misspecified space. ReMiss achieves state-of-t he-art attack success rates on the AdvBench benchmark against various target al igned LLMs while preserving the human readability of the generated prompts. Furt hermore, these attacks on open-source models demonstrate high transferability to closed-source models like GPT-4o and out-of-distribution tasks from HarmBench. Detailed analysis highlights the unique advantages of the proposed reward misspe cification objective compared to previous methods, offering new insights for imp

roving LLM safety and robustness.
****
Hoang V. Tran,Thieu Vo,An Nguyen The,Tho Tran Huu,Minh-Khoi Nguyen-Nhat,Thanh Tr
an,Duy-Tung Pham,Tan Minh Nguyen
Equivariant Neural Functional Networks for Transformers
This paper systematically explores neural functional networks (NFN) for transfor
mer architectures. NFN are specialized neural networks that treat the weights, g
radients, or sparsity patterns of a deep neural network (DNN) as input data and
have proven valuable for tasks such as learnable optimizers, implicit data repre
sentations, and weight editing. While NFN have been extensively developed for ML
P and CNN, no prior work has addressed their design for transformers, despite th
e importance of transformers in modern deep learning. This paper aims to address
 this gap by providing a systematic study of NFN for transformers. We first dete
rmine the maximal symmetric group of the weights in a multi-head attention modul
e as well as a necessary and sufficient condition under which two sets of hyperp
arameters of the multi-head attention module define the same function. We then d
efine the weight space of transformer architectures and its associated group act
ion, which leads to the design principles for NFN in transformers. Based on thes
e, we introduce Transformer-NFN, an NFN that is equivariant under this group act
ion. Additionally, we release a dataset of more than 125,000 Transformers model
checkpoints trained on two datasets with two different tasks, providing a benchm
ark for evaluating Transformer-NFN and encouraging further research on transform
er training and performance.
****
Roberto Garcia,Jerry Weihong Liu,Daniel Sorvisto,Sabri Eyuboglu
Adaptive Rank Allocation: Speeding Up Modern Transformers with RaNA Adapters
Large Language Models (LLMs) are computationally intensive, particularly during
inference. Neuron-adaptive techniques, which selectively activate neurons in Mul
ti-Layer Perceptron (MLP) layers, offer some speedups but suffer from limitation
s in modern Transformers. These include reliance on sparse activations, incompat
ibility with attention layers, and the use of costly neuron masking techniques.
To address these issues, we propose the Adaptive Rank Allocation framework and i
ntroduce the Rank and Neuron Allocator (RaNA) adapter. RaNA adapters leverage ra
nk adapters, which operate on linear layers by applying both low-rank matrix dec
ompositions and adaptive masking to efficiently allocate compute without dependi
ng on activation sparsity. This enables RaNA to be generally applied to MLPs and
 linear components of attention modules, while eliminating the need for expensiv
e maskers found in neuron-adaptive methods. Notably, when compared to neuron ada
pters, RaNA improves perplexity by up to 7 points and increases accuracy by up t
o 8 percentage-points when reducing FLOPs by $\sim$44\% in state-of-the-art Tran
sformer architectures. These results position RaNA as a robust solution for impr
oving inference efficiency in modern Transformer architectures.
****
Simon Schrodi,David T. Hoffmann,Max Argus,Volker Fischer,Thomas Brox
Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imba
lance in Contrastive Vision-Language Models
Contrastive vision-language models (VLMs), like CLIP, have gained popularity for
 their versatile applicability to various downstream tasks. Despite their succes
ses in some tasks, like zero-shot object recognition, they perform surprisingly
poor on other tasks, like attribute recognition. Previous work has attributed th
ese challenges to the modality gap, a separation of image and text in the shared
 representation space, and to a bias towards objects over other factors, such as
 attributes. In this analysis paper, we investigate both phenomena thoroughly. W
e evaluated off-the-shelf VLMs and while the gap's influence on performance is t
ypically overshadowed by other factors, we find indications that closing the gap
 indeed leads to improvements. Moreover, we find that, contrary to intuition, on
ly few embedding dimensions drive the gap and that the embedding spaces are diff
erently organized. To allow for a clean study of object bias, we introduce a def
inition and a corresponding measure of it. Equipped with this tool, we find that
 object bias does not lead to worse performance on other concepts, such as attri

butes per se. However, why do both phenomena, modality gap and object bias, emerge in the first place? To answer this fundamental question and uncover some of the inner workings of contrastive VLMs, we conducted experiments that allowed us to control the amount of shared information between the modalities. These experiments revealed that the driving factor behind both the modality gap and the object bias, is an information imbalance between images and captions, and unveiled an intriguing connection between the modality gap and entropy of the logits.
****
Xuan Liu,Jie ZHANG,HaoYang Shang,Song Guo,Yang Chengxu,Quanyan Zhu
Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View
Large language models (LLMs) have been shown to face hallucination issues due to the data they trained on often containing human bias; whether this is reflected in the decision-making process of LLM agents remains under-explored. As LLM Agents are increasingly employed in intricate social environments, a pressing and natural question emerges: Can we utilize LLM Agents' systematic hallucinations to mirror human cognitive biases, thus exhibiting irrational social intelligence? In this paper, we probe the irrational behavior among contemporary LLM agents by melding practical social science experiments with theoretical insights. Specifically, we propose CogMir, an open-ended Multi-LLM Agents framework that utilizes hallucination properties to assess and enhance LLM Agents’ social intelligence through cognitive biases. Experimental results on CogMir subsets show that LLM Agents and humans exhibit high consistency in irrational and prosocial decision-making under uncertain conditions, underscoring the prosociality of LLM Agents as social entities and highlighting the significance of hallucination properties. Additionally, CogMir framework demonstrates its potential as a valuable platform for encouraging more research into the social intelligence of LLM Agents.
****
Mathieu Chevalley,Patrick Schwab,Arash Mehrjou
Deriving Causal Order from Single-Variable Interventions: Guarantees & Algorithm
Targeted and uniform interventions to a system are crucial for unveiling causal relationships. While several methods have been developed to leverage interventional data for causal structure learning, their practical application in real-world scenarios often remains challenging. Recent benchmark studies have highlighted these difficulties, even when large numbers of single-variable intervention samples are available. In this work, we demonstrate, both theoretically and empirically, that such datasets contain a wealth of causal information that can be effectively extracted under realistic assumptions about the data distribution. More specifically, we introduce a novel variant of interventional faithfulness, which relies on comparisons between the marginal distributions of each variable across observational and interventional settings, and we introduce a score on causal orders. Under this assumption, we are able to prove strong theoretical guarantees on the optimum of our score that also hold for large-scale settings. To empirically verify our theory, we introduce Intersort, an algorithm designed to infer the causal order from datasets containing large numbers of single-variable interventions by approximately optimizing our score. Intersort outperforms baselines (GIES, DCDI, PC and EASE) on almost all simulated data settings replicating common benchmarks in the field. Our proposed novel approach to modeling interventional datasets thus offers a promising avenue for advancing causal inference, highlighting significant potential for further enhancements under realistic assumptions.
****
Rasoul Shafipour,David Harrison,Maxwell Horton,JEFFREY MARKER,Houman Bedayat,Sachin Mehta,Mohammad Rastegari,Mahyar Najibi,Saman Naderiparizi
SeedLM: Compressing LLM Weights into Seeds of Pseudo-Random Generators
Large Language Models (LLMs) have transformed natural language processing, but face significant challenges in widespread deployment due to their high runtime cost. In this paper, we introduce SeedLM, a novel post-training compression method that uses seeds of a pseudo-random generator to encode and compress model weights. Specifically, for each block of weights, we find a seed that is fed into a Linear Feedback Shift Register (LFSR) during inference to efficiently generate a

random matrix. This matrix is then linearly combined with compressed coefficient
s to reconstruct the weight block. SeedLM reduces memory access and leverages id
le compute cycles during inference, effectively speeding up memory-bound tasks b
y trading compute for fewer memory accesses. Unlike state-of-the-art methods tha
t rely on calibration data, our approach is data-free and generalizes well acros
s diverse tasks. Our experiments with Llama3 70B, which is particularly challeng
ing, show zero-shot accuracy retention at 4- and 3-bit compression to be on par
with or better than state-of-the-art methods, while maintaining performance comp
arable to FP16 baselines. Additionally, FPGA-based tests demonstrate that 4-bit
SeedLM, as model size increases, approaches a 4x speed-up over an FP16 Llama 2/3
 baseline.
****
Masao Someki,Yifan Peng,Siddhant Arora,Markus Müller,Athanasios Mouchtaris,Grant
 Strimel,Jing Liu,Shinji Watanabe
Context-aware Dynamic Pruning for Speech Foundation Models
Foundation models, such as large language models, have achieved remarkable succe
ss in natural language processing and are evolving into models capable of handli
ng multiple modalities.
Listening ability, in particular, is crucial for many applications, leading to r
esearch on building speech foundation models. However, the high computational co
st of these large models presents a significant challenge for real-world applica
tions. Although substantial efforts have been made to reduce computational costs
, such as through pruning techniques, the majority of these approaches are appli
ed primarily during the training phase for specific downstream tasks. In this st
udy, we hypothesize that optimal pruned networks may vary based on contextual fa
ctors such as speaker characteristics, languages, and tasks. To address this, we
 propose a dynamic pruning technique that adapts to these contexts during infere
nce without altering the underlying model. We demonstrated that we could success
fully reduce inference time by approximately 30\% while maintaining accuracy in
multilingual/multi-task scenarios. We also found that the obtained pruned struct
ure offers meaningful interpretations based on the context, e.g., task-related i
nformation emerging as the dominant factor for efficient pruning.
****
Lvmin Zhang,Anyi Rao,Maneesh Agrawala
Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and
Editing by Imposing Consistent Light Transport
Diffusion-based image generators are becoming unique methods for illumination ha
rmonization and editing. The current bottleneck in scaling up the training of di
ffusion-based illumination editing models is mainly in the difficulty of preserv
ing the underlying image details and maintaining intrinsic properties, such as a
lbedos, unchanged. Without appropriate constraints, directly training the latest
 large image models with complex, varied, or in-the-wild data is likely to produ
ce a structure-guided random image generator, rather than achieving the intended
 goal of precise illumination manipulation. We propose Imposing Consistent Light
 (IC-Light) transport during training, rooted in the physical principle that the
 linear blending of an object's appearances under different illumination conditi
ons is consistent with its appearance under mixed illumination. This consistency
 allows for stable and scalable illumination learning, uniform handling of vario
us data sources, and facilitates a physically grounded model behavior that modif
ies only the illumination of images while keeping other intrinsic properties unc
hanged. Based on this method, we can scale up the training of diffusion-based il
lumination editing models to large data quantities (> 10 million), across all av
ailable data types (real light stages, rendered samples, in-the-wild synthetic a
ugmentations, etc), and using strong backbones (SDXL, Flux, etc). We also demons
trate that this approach reduces uncertainties and mitigates artifacts such as m
ismatched materials or altered albedos.
****
Xianliang Li,Jun Luo,Zhiwei Zheng,Hanxiao Wang,Li Luo,Lingkun Wen,Linlong Wu,She
ng Xu
On the Performance Analysis of Momentum Method: A Frequency Domain Perspective

Momentum-based optimizers are widely adopted for training neural networks. However, the optimal selection of momentum coefficients remains elusive. This uncertainty impedes a clear understanding of the role of momentum in stochastic gradient methods. In this paper, we present a frequency domain analysis framework that interprets the momentum method as a time-variant filter for gradients, where adjustments to momentum coefficients modify the filter characteristics. Our experiments support this perspective and provide a deeper understanding of the mechanism involved. Moreover, our analysis reveals the following significant findings: high-frequency gradient components are undesired in the late stages of training; preserving the original gradient in the early stages, and gradually amplifying low-frequency gradient components during training both enhance performance. Based on these insights, we propose Frequency Stochastic Gradient Descent with Momentum (FSGDM), a heuristic optimizer that dynamically adjusts the momentum filtering characteristic with an empirically effective dynamic magnitude response. Experimental results demonstrate the superiority of FSGDM over conventional momentum optimizers.
****
Marianne Arriola,Aaron Gokaslan,Justin T Chiu,Zhihan Yang,Zhixuan Qi,Jiaqi Han,Subham Sekhar Sahoo,Volodymyr Kuleshov
Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models
Diffusion language models offer unique benefits over autoregressive models due to their potential for parallelized generation and controllability, yet they lag in likelihood modeling and are limited to fixed-length generation. In this work, we introduce a class of block diffusion language models that interpolate between discrete denoising diffusion and autoregressive models. Block diffusion overcomes key limitations of both approaches by supporting flexible-length generation and improving inference efficiency with KV caching and parallel token sampling. We propose a recipe for building effective block diffusion models that includes an efficient training algorithm, estimators of gradient variance, and data-driven noise schedules to minimize the variance. Block diffusion sets a new state-of-the-art performance among diffusion models on language modeling benchmarks and enables generation of arbitrary-length sequences. We provide the code, along with the model weights and blog post on the project page: https://m-arriola.com/bd3lms/
****
Liang CHEN,Li Shen,Yang Deng,Xiaoyan Zhao,Bin Liang,Kam-Fai Wong
PEARL: Towards Permutation-Resilient LLMs
The in-context learning (ICL) capability of large language models (LLMs) enables them to perform challenging tasks using provided demonstrations. However, ICL is highly sensitive to the ordering of demonstrations, leading to instability in predictions. This paper shows that this vulnerability can be exploited to design a natural attack—difficult for model providers to detect—that achieves nearly 80% success rate on LLaMA-3 by simply permuting the demonstrations. Existing mitigation methods primarily rely on post-processing and fail to enhance the model's inherent robustness to input permutations, raising concerns about safety and reliability of LLMs. To address this issue, we propose Permutation-resilient learning (PEARL), a novel framework based on distributionally robust optimization (DRO), which optimizes model performance against the worst-case input permutation. Specifically, PEARL consists of a permutation-proposal network (P-Net) and the LLM. The P-Net generates the most challenging permutations by treating it as an optimal transport problem, which is solved using an entropy-constrained Sinkhorn algorithm. Through minimax optimization, the P-Net and the LLM iteratively optimize against each other, progressively improving the LLM's robustness. Experiments on synthetic pre-training and real-world instruction tuning tasks demonstrate that PEARL effectively mitigates permutation attacks and enhances performance. Notably, despite being trained on fewer shots and shorter contexts, PEARL achieves performance gains of up to 40% when scaled to many-shot and long-context scenarios, highlighting its efficiency and generalization capabilities.
****

Tanqiu Jiang,Changjiang Li,Fenglong Ma,Ting Wang
RAPID: Retrieval Augmented Training of Differentially Private Diffusion Models
Differentially private diffusion models (DPDMs) harness the remarkable generativ
e capabilities of diffusion models while enforcing differential privacy (DP) for
 sensitive data. However, existing DPDM training approaches often suffer from si
gnificant utility loss, large memory footprint, and expensive inference cost, im
peding their practical uses.
To overcome such limitations, we present RAPID: Retrieval Augmented PrIvate Diff
usion model, a novel approach that integrates retrieval augmented generation (RA
G) into DPDM training. Specifically, RAPID leverages available public data to bu
ild a knowledge base of sample trajectories; when training the diffusion model o
n private data, RAPID computes the early sampling steps as queries, retrieves si
milar trajectories from the knowledge base as surrogates, and focuses on trainin
g the later sampling steps in a differentially private manner. Extensive evaluat
ion using benchmark datasets and models demonstrates that, with the same privacy
 guarantee, RAPID significantly outperforms state-of-the-art approaches by large
 margins in generative quality, memory footprint, and inference cost, suggesting
 that retrieval-augmented DP training represents a promising direction for devel
oping future privacy-preserving generative models. The code is available at: ht
tps://github.com/TanqiuJiang/RAPID
****
Christopher A. Choquette-Choo,Arun Ganesh,Saminul Haque,Thomas Steinke,Abhradeep
 Guha Thakurta
Near-Exact Privacy Amplification for Matrix Mechanisms
We study the problem of computing the privacy parameters for DP machine learning
 when using privacy amplification via random batching and noise correlated acros
s rounds via a correlation matrix $\textbf{C}$ (i.e., the matrix mechanism). Pas
t work on this problem either only applied to banded $\textbf{C}$, or gave loose
 privacy parameters. In this work, we give a framework for computing near-exact
privacy parameters for any lower-triangular, non-negative $\textbf{C}$. Our fram
ework allows us to optimize the correlation matrix $\textbf{C}$ while accounting
 for amplification, whereas past work could not. Empirically, we show this lets
us achieve smaller RMSE on prefix sums than the previous state-of-the-art (SOTA)
. We also show that we can improve on the SOTA performance on deep learning task
s. Our two main technical tools are (i) using Monte Carlo accounting to bypass c
omposition, which was the main technical challenge for past work, and (ii) a ``b
alls-in-bins'' batching scheme that enables easy privacy analysis and is closer
to practical random batching than Poisson sampling.
****
Zhiyong Wang,Dongruo Zhou,John C.S. Lui,Wen Sun
Model-based RL as a Minimalist Approach to Horizon-Free and Second-Order Bounds
Learning a transition model via Maximum Likelihood Estimation (MLE) followed by
planning inside the learned model is perhaps the most standard and simplest Mode
l-based Reinforcement Learning (RL) framework. In this work, we show that such a
 simple Model-based RL scheme, when equipped with optimistic and pessimistic pla
nning procedures, achieves strong regret and sample complexity bounds in online
and offline RL settings. Particularly, we demonstrate that under the conditions
where the trajectory-wise reward is normalized between zero and one and the tran
sition is time-homogenous, it achieves nearly horizon-free and second-order boun
ds.
****
Yida Yin,Zhiqiu Xu,Zhiyuan Li,Trevor Darrell,Zhuang Liu
A Coefficient Makes SVRG Effective
Stochastic Variance Reduced Gradient (SVRG), introduced by Johnson & Zhang (2013
), is a theoretically compelling optimization method. However, as Defazio & Bott
ou (2019) highlight, its effectiveness in deep learning is yet to be proven. In
this work, we demonstrate the potential of SVRG in optimizing real-world neural
networks. Our empirical analysis finds that, for deeper neural networks, the str
ength of the variance reduction term in SVRG should be smaller and decrease as t
raining progresses. Inspired by this, we introduce a multiplicative coefficient

α to control the strength and adjust it through a linear decay schedule. We name our method α-SVRG. Our results show α-SVRG better optimizes models, consistently reducing training loss compared to the baseline and standard SVRG across various model architectures and multiple image classification datasets. We hope our findings encourage further exploration into variance reduction techniques in deep learning. Code is available at github.com/davidyyd/alpha-SVRG.
****
Haoqi Yuan,Bohan Zhou,Yuhui Fu,Zongqing Lu
Cross-Embodiment Dexterous Grasping with Reinforcement Learning
Dexterous hands exhibit significant potential for complex real-world grasping tasks. While recent studies have primarily focused on learning policies for specific robotic hands, the development of a universal policy that controls diverse dexterous hands remains largely unexplored.
In this work, we study the learning of cross-embodiment dexterous grasping policies using reinforcement learning (RL). Inspired by the capability of human hands to control various dexterous hands through teleoperation, we propose a universal action space based on the human hand's eigengrasps. The policy outputs eigengrasp actions that are then converted into specific joint actions for each robot hand through a retargeting mapping. We simplify the robot hand's proprioception to include only the positions of fingertips and the palm, offering a unified observation space across different robot hands. Our approach demonstrates an 80\% success rate in grasping objects from the YCB dataset across four distinct embodiments using a single vision-based policy. Additionally, our policy exhibits zero-shot generalization to two previously unseen embodiments and significant improvement in efficient finetuning. For further details and videos, visit our project page (https://sites.google.com/view/crossdex).
****
Yunyang Li,Zaishuo Xia,Lin Huang,Xinran Wei,Samuel Harshe,Han Yang,Erpai Luo,Zun Wang,Jia Zhang,Chang Liu,Bin Shao,Mark Gerstein
Enhancing the Scalability and Applicability of Kohn-Sham Hamiltonians for Molecular Systems
Density Functional Theory (DFT) is a pivotal method within quantum chemistry and materials science, with its core involving the construction and solution of the Kohn-Sham Hamiltonian. Despite its importance, the application of DFT is frequently limited by the substantial computational resources required to construct the Kohn-Sham Hamiltonian. In response to these limitations, current research has employed deep-learning models to efficiently predict molecular and solid Hamiltonians, with roto-translational symmetries encoded in their neural networks. However, the scalability of prior models may be problematic when applied to large molecules, resulting in non-physical predictions of ground-state properties. In this study, we generate a substantially larger training set (PubChemQH) than used previously and use it to create a scalable model for DFT calculations with physical accuracy. For our model, we introduce a loss function derived from physical principles, which we call Wavefunction Alignment Loss (WALoss). WALoss involves performing a basis change on the predicted Hamiltonian to align it with the observed one; thus, the resulting differences can serve as a surrogate for orbital energy differences, allowing models to make better predictions for molecular orbitals and total energies than previously possible. WALoss also substantially accelerates self-consistent-field (SCF) DFT calculations. Here, we show it achieves a reduction in total energy prediction error by a factor of 1347 and an SCF calculation speed-up by a factor of 18\%. These substantial improvements set new benchmarks for achieving accurate and applicable predictions in larger molecular systems.
****
Rei Higuchi,Pierre-Louis Poirion,Akiko Takeda
Improving Convergence Guarantees of Random Subspace Second-order Algorithm for Nonconvex Optimization
In recent years, random subspace methods have been actively studied for large-dimensional nonconvex problems. Recent subspace methods have improved theoretical

guarantees such as iteration complexity and local convergence rate while reducin
g computational costs by deriving descent directions in randomly selected low-di
mensional subspaces. This paper proposes the Random Subspace Homogenized Trust R
egion (RSHTR) method with the best theoretical guarantees among random subspace
algorithms for nonconvex optimization. RSHTR achieves an ε-approxima
te first-order stationary point in $O(\varepsilon^{-3/2})$ iterations, convergin
g locally at a linear rate. Furthermore, under rank-deficient conditions, RSHTR
satisfies ε-approximate second-order necessary conditions in $O(\var
epsilon^{-3/2})$ iterations and exhibits a local quadratic convergence. Experime
nts on real-world datasets verify the benefits of RSHTR.
****

Jayden Teoh,Pradeep Varakantham,Peter Vamplew
On Generalization Across Environments In Multi-Objective Reinforcement Learning
Real-world sequential decision-making tasks often require balancing trade-offs b
etween multiple conflicting objectives, making Multi-Objective Reinforcement Lea
rning (MORL) an increasingly prominent field of research. Despite recent advance
s, existing MORL literature has narrowly focused on performance within static en
vironments, neglecting the importance of generalizing across diverse settings. C
onversely, existing research on generalization in RL has always assumed scalar r
ewards, overlooking the inherent multi-objectivity of real-world problems. Gener
alization in the multi-objective context is fundamentally more challenging, as i
t requires learning a Pareto set of policies addressing varying preferences acro
ss multiple objectives. In this paper, we formalize the concept of generalizatio
n in MORL and how it can be evaluated. We then contribute a novel benchmark feat
uring diverse multi-objective domains with parameterized environment configurati
ons to facilitate future studies in this area. Our baseline evaluations of state
-of-the-art MORL algorithms on this benchmark reveals limited generalization cap
abilities, suggesting significant room for improvement. Our empirical findings a
lso expose limitations in the expressivity of scalar rewards, emphasizing the ne
ed for multi-objective specifications to achieve effective generalization. We fu
rther analyzed the algorithmic complexities within current MORL approaches that
could impede the transfer in performance from the single- to multiple-environmen
t settings. This work fills a critical gap and lays the groundwork for future re
search that brings together two key areas in reinforcement learning: solving mul
ti-objective decision-making problems and generalizing across diverse environmen
ts. We make our code available at [https://github.com/JaydenTeoh/MORL-Generaliza
tion](https://github.com/JaydenTeoh/MORL-Generalization).
****

Arinbjörn Kolbeinsson,Kyle O'Brien,Tianjin Huang,Shanghua Gao,Shiwei Liu,Jonatha
n Richard Schwarz,Anurag Jayant Vaidya,Faisal Mahmood,Marinka Zitnik,Tianlong Ch
en,Thomas Hartvigsen
Composable Interventions for Language Models
Test-time interventions for language models can enhance factual accuracy, mitiga
te harmful outputs, and improve model efficiency without costly retraining. But
despite a flood of new methods, different types of interventions are largely dev
eloping independently.
In practice, multiple interventions must be applied sequentially to the same mod
el, yet we lack standardized ways to study how interventions interact. We fill t
his gap by introducing composable interventions, a framework to study the effect
s of using multiple interventions on the same language models, featuring new met
rics and a unified codebase. Using our framework, we conduct extensive experimen
ts and compose popular methods from three emerging intervention categories---kno
wledge editing, model compression, and machine unlearning. Our results over 417
different compositions uncover meaningful interactions: compression hinders edit
ing and unlearning, composing interventions hinges on their order of application
, and popular general-purpose metrics are inadequate for assessing composability
. Taken together, our findings showcase clear gaps in composability, suggesting
a need for new multi-objective interventions.
****

Yoshiaki Kitazawa

Bounds on L_p Errors in Density Ratio Estimation via f-Divergence Loss Functions

Density ratio estimation (DRE) is a core technique in machine learning used to capture relationships between two probability distributions.
f-divergence loss functions, which are derived from variational representations of f-divergence, have become a standard choice in DRE for achieving cutting-edge performance.
This study provides novel theoretical insights into DRE by deriving upper and lower bounds on the L_p errors through f-divergence loss functions.
These bounds apply to any estimator belonging to a class of Lipschitz continuous estimators, irrespective of the specific f-divergence loss function employed.
The derived bounds are expressed as a product involving the data dimensionality and the expected value of the density ratio raised to the p-th power.
Notably, the lower bound includes an exponential term that depends on the Kullback--Leibler (KL) divergence, revealing that the L_p error increases significantly as the KL divergence grows when $p > 1$.
This increase becomes even more pronounced as the value of p grows.
The theoretical insights are validated through numerical experiments.
****

Pei Liu,Luping Ji,Jiaxiang Gou,Bo Fu,Mao Ye

Interpretable Vision-Language Survival Analysis with Ordinal Inductive Bias for Computational Pathology

Histopathology Whole-Slide Images (WSIs) provide an important tool to assess cancer prognosis in computational pathology (CPATH). While existing survival analysis (SA) approaches have made exciting progress, they are generally limited to adopting highly-expressive network architectures and only coarse-grained patient-level labels to learn visual prognostic representations from gigapixel WSIs. Such learning paradigm suffers from critical performance bottlenecks, when facing present scarce training data and standard multi-instance learning (MIL) framework in CPATH. To overcome it, this paper, for the first time, proposes a new Vision-Language-based SA (**VLSA**) paradigm. Concretely, (1) VLSA is driven by pathology VL foundation models. It no longer relies on high-capability networks and shows the advantage of *data efficiency*. (2) In vision-end, VLSA encodes textual prognostic prior and then employs it as *auxiliary signals* to guide the aggregating of visual prognostic features at instance level, thereby compensating for the weak supervision in MIL. Moreover, given the characteristics of SA, we propose i) *ordinal survival prompt learning* to transform continuous survival labels into textual prompts; and ii) *ordinal incidence function* as prediction target to make SA compatible with VL-based prediction. Notably, VLSA's predictions can be interpreted intuitively by our Shapley values-based method. The extensive experiments on five datasets confirm the effectiveness of our scheme. Our VLSA could pave a new way for SA in CPATH by offering weakly-supervised MIL an effective means to learn valuable prognostic clues from gigapixel WSIs. Our source code is available at https://github.com/liupei101/VLSA.
****

Hongjue Zhao,Yuchen Wang,Hairong Qi,Zijie Huang,Han Zhao,Lui Sha,Huajie Shao

Accelerating Neural ODEs: A Variational Formulation-based Approach

Neural Ordinary Differential Equations (Neural ODEs or NODEs) excel at modeling continuous dynamical systems from observational data, especially when the data is irregularly sampled. However, existing training methods predominantly rely on numerical ODE solvers, which are time-consuming and prone to accumulating numerical errors over time due to autoregression. In this work, we propose VF-NODE, a novel approach based on the variational formulation (VF) to accelerate the training of NODEs. Unlike existing training methods, the proposed VF-NODEs implement a series of global integrals, thus evaluating Deep Neural Network (DNN)--based vector fields only at specific observed data points. This strategy drastically reduces the number of function evaluations (NFEs). Moreover, our method eliminates the use of autoregression, thereby reducing error accumulations for modeling dynamical systems. Nevertheless, the VF loss introduces oscillatory terms into the integrals when using the Fourier basis. We incorporate Filon's method to addres

s this issue. To further enhance the performance for noisy and incomplete data, we employ the natural cubic spline regression to estimate a closed-form approximation. We provide a fundamental analysis of how our approach minimizes computational costs. Extensive experiments demonstrate that our approach accelerates NODE training by 10 to 1000 times compared to existing NODE-based methods, while achieving higher or comparable accuracy in dynamical systems. The code is available at https://github.com/ZhaoHongjue/VF-NODE-ICLR2025.
****
Hanyang Zhao,Genta Indra Winata,Anirban Das,Shi-Xiong Zhang,David Yao,Wenpin Tang,Sambit Sahu
RainbowPO: A Unified Framework for Combining Improvements in Preference Optimization
Recently, numerous preference optimization algorithms have been introduced as extensions to the Direct Preference Optimization (DPO) family. While these methods have successfully aligned models with human preferences, there is a lack of understanding regarding the contributions of their additional components. Moreover, fair and consistent comparisons are scarce, making it difficult to discern which components genuinely enhance downstream performance. In this work, we propose RainbowPO, a unified framework that demystifies the effectiveness of existing DPO methods by categorizing their key components into seven broad directions. We integrate these components into a single cohesive objective, enhancing the performance of each individual element. Through extensive experiments, we demonstrate that RainbowPO outperforms existing DPO variants. Additionally, we provide insights to guide researchers in developing new DPO methods and assist practitioners in their implementations.
****
Zhiwei Zhang,Minhua Lin,Junjie Xu,Zongyu Wu,Enyan Dai,Suhang Wang
Robustness Inspired Graph Backdoor Defense
Graph Neural Networks (GNNs) have achieved promising results in tasks such as node classification and graph classification. However, recent studies reveal that GNNs are vulnerable to backdoor attacks, posing a significant threat to their real-world adoption. Despite initial efforts to defend against specific graph backdoor attacks, there is no work on defending against various types of backdoor attacks where generated triggers have different properties. Hence, we first empirically verify that prediction variance under edge dropping is a crucial indicator for identifying poisoned nodes. With this observation, we propose using random edge dropping to detect backdoors and theoretically show that it can efficiently distinguish poisoned nodes from clean ones. Furthermore, we introduce a novel robust training strategy to efficiently counteract the impact of the triggers. Extensive experiments on real-world datasets show that our framework can effectively identify poisoned nodes, significantly degrade the attack success rate, and maintain clean accuracy when defending against various types of graph backdoor attacks with different properties. Our code is available at: https://github.com/zzwjames/RIGBD.
****
Jan-Hendrik Bastek,WaiChing Sun,Dennis Kochmann
Physics-Informed Diffusion Models
Generative models such as denoising diffusion models are quickly advancing their ability to approximate highly complex data distributions. They are also increasingly leveraged in scientific machine learning, where samples from the implied data distribution are expected to adhere to specific governing equations. We present a framework that unifies generative modeling and partial differential equation fulfillment by introducing a first-principle-based loss term that enforces generated samples to fulfill the underlying physical constraints. Our approach reduces the residual error by up to two orders of magnitudes compared to previous work in a fluid flow case study and outperforms task-specific frameworks in relevant metrics for structural topology optimization. We also present numerical evidence that our extended training objective acts as a natural regularization mechanism against overfitting. Our framework is simple to implement and versatile in its applicability for imposing equality and inequality constraints as well as au

xiliary optimization objectives. Code is available at https://github.com/jhbastek/PhysicsInformedDiffusionModels.
****
Hao Di,Tong He,Haishan Ye,Yinghui Huang,Xiangyu Chang,Guang Dai,Ivor Tsang
ProAdvPrompter: A Two-Stage Journey to Effective Adversarial Prompting for LLMs
As large language models (LLMs) are increasingly being integrated into various real-world applications, the identification of their vulnerabilities to jailbreaking attacks becomes an essential component of ensuring the safety and reliability of LLMs.
Previous studies have developed LLM assistants, known as the adversarial prompter, to automatically generate suffixes that manipulate target LLMs into generating harmful and undesirable outputs.
However, these approaches often suffer from low performance or generate semantically meaningless prompts, which can be easily identified by perplexity-based defenses.
In this paper, we introduce a novel two-stage method, $\texttt{ProAdvPrompter}$, that significantly improves the performance of adversarial prompters.
In $\texttt{ProAdvPrompter}$, the first stage (Exploration) utilizes the loss information to guide the adversarial prompter in generating suffixes that are more likely to elicit harmful responses.
Then the second stage (Exploitation) iteratively fine-tunes the prompter using high-quality generated adversarial suffixes to further boost performance.
Additionally, we incorporate the prompt template to aid in the Exploration stage and propose a filtering mechanism to accelerate the training process in the Exploitation stage.
We evaluate $\texttt{ProAdvPrompter}$ against the well-aligned LLMs (i.e., Llama 2-Chat-7B and Llama3-chat-8B), achieving attack success rates of 99.68% and 97.12% respectively after 10 trials on the AdvBench dataset, thereby enhancing performance by ~ 2 times compared to previous works.
Moreover, $\texttt{ProAdvPrompter}$ reduces training time by 20% on Llama3-Instruct-8B, generates more generalized adversarial suffixes, and demonstrates resilience against the perplexity defense.
An ablation study further evaluates the effects of key components in $\texttt{ProAdvPrompter}$ (the prompt template and the filtering mechanism).
****
Xinran Zhao,Hanie Sedghi,Bernd Bohnet,Dale Schuurmans,Azade Nova
Improving Large Language Model Planning with Action Sequence Similarity
Planning is essential for artificial intelligence systems to look ahead and proactively determine a course of actions to reach objectives in the virtual and real world. Recent work on large language models (LLMs) sheds light on their planning capability in various tasks. However, it remains unclear what signals in the context influence the model performance. In this work, we explore how to improve the model planning capability through in-context learning (ICL), specifically, what signals can help select the exemplars. Through extensive experiments, we observe that commonly used problem similarity may result in false positives with drastically different plans, which can mislead the model. In response, we propose to sample and filter exemplars leveraging plan side action sequence similarity (AS). We propose GRASE-DC: a two-stage pipeline that first re-samples high AS exemplars and then curates the selected exemplars with dynamic clustering on AS to achieve a balance of relevance and diversity. Our experimental result confirms that GRASE-DC achieves significant performance improvement on various planning tasks (up to ~11-40 point absolute accuracy improvement with 27.3% fewer exemplars needed on average). With GRASE-DC* + VAL, where we iteratively apply GRASE-DC with a validator, we are able to even boost the performance by 18.9% more.
Extensive analysis validates the consistent performance improvement of GRASE-DC with various backbone LLMs and on both classical planning and natural language planning benchmarks. GRASE-DC can further boost the planning accuracy by ~24 absolute points on harder problems using simpler problems as exemplars over a random baseline. This demonstrates its ability to generalize to out-of-distribution problems.

****
Minghan Chen,Guikun Chen,Wenguan Wang,Yi Yang
Hydra-SGG: Hybrid Relation Assignment for One-stage Scene Graph Generation
DETR introduces a simplified one-stage framework for scene graph generation (SGG) but faces challenges of sparse supervision and false negative samples. The for mer occurs because each image typically contains fewer than 10 relation annotati ons, while DETR-based SGG models employ over 100 relation queries. Each ground t ruth relation is assigned to only one query during training. The latter arises w hen one ground truth relation may have multiple queries with similar matching sc ores, leading to suboptimally matched queries being treated as negative samples. To address these, we propose Hydra-SGG, a one-stage SGG method featuring a Hybr id Relation Assignment. This approach combines a One-to-One Relation Assignment with an IoU-based One-to-Many Relation Assignment, increasing positive training samples and mitigating sparse supervision. In addition, we empirically demonstra te that removing self-attention between relation queries leads to duplicate pred ictions, which actually benefits the proposed One-to-Many Relation Assignment. W ith this insight, we introduce Hydra Branch, an auxiliary decoder without self-a ttention layers, to further enhance One-to-Many Relation Assignment by promoting different queries to make the same relation prediction. Hydra-SGG achieves stat e-of-the-art performance on multiple datasets, including VG150 (16.0 mR@50), Ope n Images V6 (50.1 weighted score), and GQA (12.7 mR@50). Our code and pre-traine d models will be released on Hydra-SGG.
****
Kushagra Pandey,Jaideep Pathak,Yilun Xu,Stephan Mandt,Michael Pritchard,Arash Va hdat,Morteza Mardani
Heavy-Tailed Diffusion Models
Diffusion models achieve state-of-the-art generation quality across many applica tions, but their ability to capture rare or extreme events in heavy-tailed distr ibutions remains unclear. In this work, we show that traditional diffusion and f low-matching models with standard Gaussian priors fail to capture heavy-tailed b ehavior. We address this by repurposing the diffusion framework for heavy-tail e stimation using multivariate Student-t distributions. We develop a tailored pert urbation kernel and derive the denoising posterior based on the conditional Stud ent-t distribution for the backward process. Inspired by γ-divergence for heavy-tailed distributions, we derive a training objective for heavy-tailed den oisers. The resulting framework introduces controllable tail generation using on ly a single scalar hyperparameter, making it easily tunable for diverse real-wor ld distributions. As specific instantiations of our framework, we introduce t-ED M and t-Flow, extensions of existing diffusion and flow models that employ a Stu dent-t prior. Remarkably, our approach is readily compatible with standard Gauss ian diffusion models and requires only minimal code changes. Empirically, we sho w that our t-EDM and t-Flow outperform standard diffusion models in heavy-tail e stimation on high-resolution weather datasets in which generating rare and extre me events is crucial.
****
Zhaolong Du,Shasha Mao,Xuequan Lu,Mengnan Qi,Yimeng Zhang,Jing Gu,Licheng Jiao
Rethinking Multiple-Instance Learning From Feature Space to Probability Space
Multiple-instance learning (MIL) was initially proposed to identify key instance s within a set (bag) of instances when only one bag-level label is provided. Cur rent deep MIL models mostly solve multi-instance problem in feature space. Never theless, with the increasing complexity of data, we found this paradigm faces si gnificant risks in representation learning stage, which could lead to algorithm degradation in deep MIL models. We speculate that the degradation issue stems fr om the persistent drift of instances in feature space during learning. In this p aper, we propose a novel Probability-Space MIL network (PSMIL) as a countermeasu re. In PSMIL, a self-training alignment strategy is introduced in probability sp ace to cope with the drift problem in feature space, and the alignment target ob jective is proven mathematically optimal. Furthermore, we reveal that the widely -used attention-based pooling mechanism in current deep MIL models is easily aff ected by the perturbation in feature space and further introduce an alternative

called probability-space attention pooling. It effectively captures the key inst
ance in each bag from feature space to probability space, and further eliminates
 the impact of selection drift in the pooling stage. To summarize, PSMIL seeks t
o solve a MIL problem in probability space rather than feature space. Experiment
al results illustrate that PSMIL could potentially achieve performance close to
supervised learning level in complex tasks (gap within 5\%), with the incrementa
l alignment in propability space bring more than 19\% accuracy improvements for
current existing mainstream models in simulated CIFAR datasets. For existing pu
blicly available MIL benchmarks/datasets, attention in probability space also ac
hieves competitive performance to the state-of-the-art deep MIL models. Codes ar
e available at \url{https://github.com/LMBDA-design/PSAMIL}.
****
Xiaorui Su,Yibo Wang,Shanghua Gao,Xiaolong Liu,Valentina Giunchiglia,Djork-Arné
Clevert,Marinka Zitnik
KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA
Biomedical reasoning integrates structured, codified knowledge with tacit, exper
ience-driven insights. Depending on the context, quantity, and nature of availab
le evidence, researchers and clinicians use diverse strategies, including rule-b
ased, prototype-based, and case-based reasoning. Effective medical AI models mus
t handle this complexity while ensuring reliability and adaptability. We introdu
ce KGARevion, a knowledge graph-based agent that answers knowledge-intensive que
stions. Upon receiving a query, KGARevion generates relevant triplets by leverag
ing the latent knowledge embedded in a large language model. It then verifies th
ese triplets against a grounded knowledge graph, filtering out errors and retain
ing only accurate, contextually relevant information for the final answer. This
multi-step process strengthens reasoning, adapts to different models of medical
inference, and outperforms retrieval-augmented generation-based approaches that
lack effective verification mechanisms. Evaluations on medical QA benchmarks sho
w that KGARevion improves accuracy by over 5.2% over 15 models in handling compl
ex medical queries. To further assess its effectiveness, we curated three new me
dical QA datasets with varying levels of semantic complexity, where KGARevion im
proved accuracy by 10.4%. The agent integrates with different LLMs and biomedica
l knowledge graphs for broad applicability across knowledge-intensive tasks. We
evaluated KGARevion on AfriMed-QA, a newly introduced dataset focused on African
 healthcare, demonstrating its strong zero-shot generalization to underrepresent
ed medical contexts.
****
Murong Yue,Wenlin Yao,Haitao Mi,Dian Yu,Ziyu Yao,Dong Yu
DOTS: Learning to Reason Dynamically in LLMs via Optimal Reasoning Trajectories
Search
Enhancing the capability of large language models (LLMs) in reasoning has gained
 significant attention in recent years. Previous studies have demonstrated the e
ffectiveness of various prompting strategies in aiding LLMs in reasoning (called
 "reasoning actions"), such as step-by-step thinking, reflecting before answerin
g, solving with programs, and their combinations. However, these approaches ofte
n applied static, predefined reasoning actions uniformly to all questions, witho
ut considering the specific characteristics of each question or the capability o
f the task-solving LLM. In this paper, we propose DOTS, an approach enabling LLM
s to reason Dynamically via Optimal reasoning Trajectories Search, tailored to t
he specific characteristics of each question and the inherent capability of the
task-solving LLM.
Our approach involves three key steps: i) defining atomic reasoning action modul
es that can be composed into various reasoning action trajectories; ii) searchin
g for the optimal action trajectory for each training question through iterative
 exploration and evaluation for the specific task-solving LLM; and iii) using th
e collected optimal trajectories to train an LLM to plan for the reasoning traje
ctories of unseen questions. In particular, we propose two learning paradigms, i
.e., fine-tuning an external LLM as a planner to guide the task-solving LLM, or
directly fine-tuning the task-solving LLM with an internalized capability for re
asoning actions planning. Our experiments across eight reasoning tasks show that

our method consistently outperforms static reasoning techniques and the vanilla instruction tuning approach. Further analysis reveals that our method enables LLMs to adjust their computation based on problem complexity, allocating deeper thinking and reasoning to harder problems.
****
Yusen Wu,Bujiao Wu,Yanqi Song,Xiao Yuan,Jingbo Wang
Learning the Complexity of Weakly Noisy Quantum States
Quantifying the complexity of quantum states is a longstanding key problem in various subfields of science, ranging from quantum computing to the black-hole theory. The lower bound on quantum pure state complexity has been shown to grow linearly with system size [J. Haferkamp et al., 2022, *Nat. Phys.*]. However, extending this result to noisy circuit environments, which better reflect real quantum devices, remains an open challenge. In this paper, we explore the complexity of weakly noisy quantum states via the quantum learning method. We present an efficient learning algorithm, that leverages the classical shadow representation of target quantum states, to predict the circuit complexity of weakly noisy quantum states. Our algorithm is proved to be optimal in terms of sample complexity accompanied with polynomial classical processing time. Our result builds a bridge between the learning algorithm and quantum state complexity, meanwhile highlighting the power of learning algorithm in characterizing intrinsic properties of quantum states.
****
Hanlin Tang,Yang Lin,Jing Lin,Qingsen Han,Danning Ke,Shikuan Hong,Yiwu Yao,Gongyi Wang
RazorAttention: Efficient KV Cache Compression Through Retrieval Heads
The memory and computational demands of Key-Value (KV) cache present significant challenges for deploying long-context language models. Previous approaches attempt to mitigate this issue by selectively dropping tokens, which irreversibly erases critical information that might be needed for future queries. In this paper, we propose a novel compression technique for KV cache that preserves all token information. Our investigation reveals that: i) Most attention heads primarily focus on the local context; ii) Only a few heads, denoted as retrieval heads, can essentially pay attention to all input tokens. These key observations motivate us to use separate caching strategy for attention heads.Therefore, we propose RazorAttention, a training-free KV cache compression algorithm, which maintains a full cache for these crucial retrieval heads and discards the remote tokens in non-retrieval heads. Furthermore, we introduce a novel mechanism involving a “compensation token” to further recover the information in the dropped tokens. Extensive evaluations across a diverse set of large language models (LLMs) demonstrate that RazorAttention achieves a reduction in KV cache size by over 70% without noticeable impacts on performance. Additionally, RazorAttention is compatible with FlashAttention, rendering it an efficient and plug-and-play solution that enhances LLM inference efficiency without overhead or retraining of the original model.
****
Duy Kien Nguyen,Mido Assran,Unnat Jain,Martin R. Oswald,Cees G. M. Snoek,Xinlei Chen
An Image is Worth More Than 16x16 Patches: Exploring Transformers on Individual Pixels
This work does not introduce a new method. Instead, we present an interesting finding that questions the necessity of the inductive bias of locality in modern computer vision architectures. Concretely, we find that vanilla Transformers can operate by directly treating each individual pixel as a token and achieve highly performant results. This is substantially different from the popular design in Vision Transformer, which maintains the inductive bias from ConvNets towards local neighborhoods (e.g., by treating each 16x16 patch as a token). We showcase the effectiveness of pixels-as-tokens across three well-studied computer vision tasks: supervised learning for classification and regression, self-supervised learning via masked autoencoding, and image generation with diffusion models. Although it's computationally less practical to directly operate on individual pixels,

we believe the community must be made aware of this surprising piece of knowledge when devising the next generation of neural network architectures for computer vision.
****
Xinyu He,Dongqi Fu,Hanghang Tong,Ross Maciejewski,Jingrui He
Temporal Heterogeneous Graph Generation with Privacy, Utility, and Efficiency
Nowadays, temporal heterogeneous graphs attract much research and industrial attention for building the next-generation Relational Deep Learning models and applications, due to their informative structures and features. While providing timely and precise services like personalized recommendations and question answering, this rich information also introduces extra exposure risk for each node in the graph. The distinctive local topology, the abundant heterogeneous features, and the time dimension of the graph data are more prone to expose sensitive information and narrow down the scope of victim candidates, which calls for well-defined protection techniques on graphs. To this end, we propose a Temporal Heterogeneous Graph Generator balancing Privacy, Utility, and Efficiency, named THePUff. More specifically, we first propose a differential privacy algorithm to perturb the input temporal heterogeneous graph for protecting privacy, and then utilize both the perturbed graph and the original one in a generative adversarial setting for THePUff to learn and generate privacy-guaranteed and utility-preserved graph data in an efficient manner. We further propose 6 new metrics in the temporal setting to measure heterogeneous graph utility and privacy. Finally, based on temporal heterogeneous graph datasets with up to 1 million nodes and 20 million edges, the experiments show that THePUff generates utilizable temporal heterogeneous graphs with privacy protected, compared with state-of-the-art baselines.
****
Aly Lidayan,Michael D Dennis,Stuart Russell
BAMDP Shaping: a Unified Framework for Intrinsic Motivation and Reward Shaping
Intrinsic motivation and reward shaping guide reinforcement learning (RL) agents by adding pseudo-rewards, which can lead to useful emergent behaviors. However, they can also encourage counterproductive exploits, e.g., fixation with noisy TV screens. Here we provide a theoretical model which anticipates these behaviors, and provides broad criteria under which adverse effects can be bounded.
We characterize all pseudo-rewards as reward shaping in Bayes-Adaptive Markov Decision Processes (BAMDPs), which formulates the problem of learning in MDPs as an MDP over the agent's knowledge.
Pseudo-rewards guide exploration by incentivizing RL agents to go to states with higher BAMDP value, which comprises the value of information gathered and the prior value of the physical state, while they mislead exploration when they align poorly with this value.
We extend potential-based shaping theory to prove BAMDP Potential-based shaping Functions (BAMPFs) are immune to reward-hacking (convergence to behaviors maximizing composite rewards to the detriment of real rewards) in meta-RL, and show empirically how a BAMPF helps a meta-RL agent learn optimal RL algorithms for a Bernoulli Bandit domain.
We finally prove that BAMPFs with bounded monotone increasing potentials also resist reward-hacking in the regular RL setting. We show that it is straightforward to retrofit or design new pseudo-reward terms in this form to avoid unintended side effects, and provide an empirical demonstration in the Mountain Car environment.
****
Guibin Zhang,Yiyan Qi,Ziyang Cheng,Yanwei Yue,Dawei Cheng,Jian Guo
Rationalizing and Augmenting Dynamic Graph Neural Networks
Graph data augmentation (GDA) has shown significant promise in enhancing the performance, generalization, and robustness of graph neural networks (GNNs). However, contemporary methodologies are often limited to static graphs, whose applicability on dynamic graphs—more prevalent in real-world applications—remains unexamined. In this paper, we empirically highlight the challenges faced by static GDA methods when applied to dynamic graphs, particularly their inability to maintain temporal consistency. In light of this limitation, we propose a dedicated augm

entation framework for dynamic graphs, termed $\texttt{DyAug}$, which adaptively augments the evolving graph structure with temporal consistency awareness. Specifically, we introduce the paradigm of graph rationalization for dynamic GNNs, progressively distinguishing between causal subgraphs (\textit{rationale}) and the non-causal complement (\textit{environment}) across snapshots. We develop three types of environment replacement, including, spatial, temporal, and spatial-temporal, to facilitate data augmentation in the latent representation space, thereby improving the performance, generalization, and robustness of dynamic GNNs. Extensive experiments on six benchmarks and three GNN backbones demonstrate that $\texttt{DyAug}$ can \textbf{(I)} improve the performance of dynamic GNNs by $0.89\\%\sim3.13\\%\uparrow$; \textbf{(II)} effectively counter targeted and non-targeted adversarial attacks with $6.2\\%\sim12.2\\%\\uparrow$ performance boost; \textbf{(III)} make stable predictions under temporal distribution shifts.
****
Yuheng Chen,Pengfei Cao,Yubo Chen,Kang Liu,Jun Zhao
Knowledge Localization: Mission Not Accomplished? Enter Query Localization!
Large language models (LLMs) store extensive factual knowledge, but the mechanisms behind how they store and express this knowledge remain unclear.
The Knowledge Neuron (KN) thesis is a prominent theory for explaining these mechanisms. This theory is based on the **Knowledge Localization (KL)** assumption, which suggests that a fact can be localized to a few knowledge storage units, namely knowledge neurons.
However, this assumption has two limitations: first, it may be too rigid regarding knowledge storage, and second, it neglects the role of the attention module in knowledge expression.

In this paper, we first re-examine the KL assumption and demonstrate that its limitations do indeed exist. To address these, we then present two new findings, each targeting one of the limitations: one focusing on knowledge storage and the other on knowledge expression.
We summarize these findings as **Query Localization** assumption and argue that the KL assumption can be viewed as a simplification of the QL assumption.
Based on QL assumption, we further propose the Consistency-Aware KN modification method, which improves the performance of knowledge modification, further validating our new assumption. We conduct 39 sets of experiments, along with additional visualization experiments, to rigorously confirm our conclusions. Code will be made public soon.
****
Po-Hung Yeh,Kuang-Huei Lee,Jun-cheng Chen
Training-Free Diffusion Model Alignment with Sampling Demons
Aligning diffusion models with user preferences has been a key challenge.
Existing methods for aligning diffusion models either require retraining or are limited to differentiable reward functions.
To address these limitations, we propose a stochastic optimization approach, dubbed *Demon*, to guide the denoising process at inference time without backpropagation through reward functions or model retraining.
Our approach works by controlling noise distribution in denoising steps to concentrate density on regions corresponding to high rewards through stochastic optimization.
We provide comprehensive theoretical and empirical evidence to support and validate our approach, including experiments that use non-differentiable sources of rewards such as Visual-Language Model (VLM) APIs and human judgements.
To the best of our knowledge, the proposed approach is the first inference-time, backpropagation-free preference alignment method for diffusion models.
Our method can be easily integrated with existing diffusion models without further training.
Our experiments show that the proposed approach significantly improves the average aesthetics scores for text-to-image generation. Implementation is available at [this URL](https://github.com/aiiu-lab/DemonSampling).
****

Chenyu Zhang,Xu Chen,Xuan Di
Stochastic Semi-Gradient Descent for Learning Mean Field Games with Population-Aware Function Approximation
Mean field games (MFGs) model interactions in large-population multi-agent systems through population distributions. Traditional learning methods for MFGs are based on fixed-point iteration (FPI), where policy updates and induced population distributions are computed separately and sequentially. However, FPI-type methods may suffer from inefficiency and instability due to potential oscillations caused by this forward-backward procedure. In this work, we propose a novel perspective that treats the policy and population as a unified parameter controlling the game dynamics. By applying stochastic parameter approximation to this unified parameter, we develop SemiSGD, a simple stochastic gradient descent (SGD)-type method, where an agent updates its policy and population estimates simultaneously and fully asynchronously. Building on this perspective, we further apply linear function approximation (LFA) to the unified parameter, resulting in the first population-aware LFA (PA-LFA) for learning MFGs on continuous state-action spaces. A comprehensive finite-time convergence analysis is provided for SemiSGD with PA-LFA, including its convergence to the equilibrium for linear MFGs—a class of MFGs with a linear structure concerning the population—under the standard contractivity condition, and to a neighborhood of the equilibrium under a more practical condition. We also characterize the approximation error for non-linear MFGs. We validate our theoretical findings with six experiments on three MFGs.
****
Yonghao Zhuang,Lanxiang Hu,Longfei Yun,Souvik Kundu,Zhengzhong Liu,Eric P. Xing,Hao Zhang
Scaling Long Context Training Data by Long-Distance Referrals
Training large language models for long context understanding faces the challenge of data shortage.
Previous data engineering approaches mechanically concatenate short documents, which may create many pseudo long documents but raise concerns about data quality.
In this paper, we study the core attribute of high quality data for long context training, and provide a data pipeline, LongPack, to scale
such data.
We found that long distance referrals, which occur in natural long documents, are crucial for long-context training.
However, simply concatenating short documents does not reliably generate these relations.
We further show that the density of long-distance referrals, which is higher in longer documents, has a key role in training efficiency, making previous upsampling methods suboptimal.
To enrich long documents, we propose LongPack, a data pipeline that constructs long documents by packing shorter ones based on referral relationships.
Specifically, for web pages, which are the primary source for language model training, we found hyper-link a native signal for such a relation.
By packing web pages through their hyper-link connection, we can create longer, high-quality documents.
Our experiments demonstrate that LongPackis highly scalable, generating a corpus of long documents equivalent in size to an entire pretraining dataset using just 0.5% root documents.
Furthermore, the constructed documents have a ‘near-natural’ quality as innate long documents for long context training, reaching a 32.7% higher score than previous state-of-the-art methods.
****
Chaoran Cheng,Boran Han,Danielle C. Maddix,Abdul Fatir Ansari,Andrew Stuart,Michael W. Mahoney,Bernie Wang
Gradient-Free Generation for Hard-Constrained Systems
Generative models that satisfy hard constraints are critical in many scientific and engineering applications, where physical laws or system requirements must be strictly respected. Many existing constrained generative models, especially tho

se developed for computer vision, rely heavily on gradient information, which is often sparse or computationally expensive in some fields, e.g., partial differential equations (PDEs).
In this work, we introduce a novel framework for adapting pre-trained, unconstrained flow-matching models to satisfy constraints exactly in a zero-shot manner without requiring expensive gradient computations or fine-tuning.
Our framework, *ECI sampling*, alternates between extrapolation (E), correction (C), and interpolation (I) stages during each iterative sampling step of flow matching sampling to ensure accurate integration of constraint information while preserving the validity of the generation.
We demonstrate the effectiveness of our approach across various PDE systems, showing that ECI-guided generation strictly adheres to physical constraints and accurately captures complex distribution shifts induced by these constraints.
Empirical results demonstrate that our framework consistently outperforms baseline approaches in various zero-shot constrained generation tasks and also achieves competitive results in the regression tasks without additional fine-tuning.
****
Ahmed H. Salamah,Kaixiang Zheng,Yiwen Liu,EN-HUI YANG
JPEG Inspired Deep Learning
Although it is traditionally believed that lossy image compression, such as JPEG compression, has a negative impact on the performance of deep neural networks (DNNs), it is shown by recent works that well-crafted JPEG compression can actually improve the performance of deep learning (DL). Inspired by this, we propose JPEG-DL, a novel DL framework that prepends any underlying DNN architecture with a trainable JPEG compression layer. To make the quantization operation in JPEG compression trainable, a new differentiable soft quantizer is employed at the JPEG layer, and then the quantization operation and underlying DNN are jointly trained. Extensive experiments show that in comparison with the standard DL, JPEG-DL delivers significant accuracy improvements across various datasets and model architectures while enhancing robustness against adversarial attacks. Particularly, on some fine-grained image classification datasets, JPEG-DL can increase prediction accuracy by as much as 20.9%. Our code is available on https://github.com/AhmedHussKhalifa/JPEG-Inspired-DL.git.
****
Neta Shaul,Itai Gat,Marton Havasi,Daniel Severo,Anuroop Sriram,Peter Holderrieth,Brian Karrer,Yaron Lipman,Ricky T. Q. Chen
Flow Matching with General Discrete Paths: A Kinetic-Optimal Perspective
The design space of discrete-space diffusion or flow generative models are significantly less well-understood than their continuous-space counterparts, with many works focusing only on a simple masked construction.
In this work, we aim to take a holistic approach to the construction of discrete generative models based on continuous-time Markov chains, and for the first time, allow the use of arbitrary discrete probability paths, or colloquially, corruption processes.
Through the lens of optimizing the symmetric kinetic energy, we propose velocity formulas that can be applied to any given probability path, completely decoupling the probability and velocity, and giving the user the freedom to specify any desirable probability path based on expert knowledge specific to the data domain.
Furthermore, we find that a special construction of mixture probability paths optimizes the symmetric kinetic energy for the discrete case.
We empirically validate the usefulness of this new design space across multiple modalities: text generation, inorganic material generation, and image generation. We find that we can outperform the mask construction even in text with kinetic-optimal mixture paths, while we can make use of domain-specific constructions of the probability path over the visual domain.
****
Leo Gao,Tom Dupre la Tour,Henk Tillman,Gabriel Goh,Rajan Troll,Alec Radford,Ilya Sutskever,Jan Leike,Jeffrey Wu
Scaling and evaluating sparse autoencoders

Sparse autoencoders provide a promising unsupervised approach for extracting interpretable features from a language model by reconstructing activations from a sparse bottleneck layer. Since language models learn many concepts, autoencoders need to be very large to recover all relevant features. However, studying the properties of autoencoder scaling is difficult due to the need to balance reconstruction and sparsity objectives and the presence of dead latents. We propose using k-sparse autoencoders [Makhzani and Frey, 2013] to directly control sparsity, simplifying tuning and improving the reconstruction-sparsity frontier. Additionally, we find modifications that result in few dead latents, even at the largest scales we tried. Using these techniques, we find clean scaling laws with respect to autoencoder size and sparsity. We also introduce several new metrics for evaluating feature quality based on the recovery of hypothesized features, the explainability of activation patterns, and the sparsity of downstream effects. These metrics all generally improve with autoencoder size. To demonstrate the scalability of our approach, we train a 16 million latent autoencoder on GPT-4 activations for 40 billion tokens. We release training code and autoencoders for open-source models, as well as a visualizer.
****

Andy K Zhang,Neil Perry,Riya Dulepet,Joey Ji,Celeste Menders,Justin W Lin,Eliot Jones,Gashon Hussein,Samantha Liu,Donovan Julian Jasper,Pura Peetathawatchai,Ari Glenn,Vikram Sivashankar,Daniel Zamoshchin,Leo Glikbarg,Derek Askaryar,Haoxiang Yang,Aolin Zhang,Rishi Alluri,Nathan Tran,Rinnara Sangpisit,Kenny O Oseleononmen,Dan Boneh,Daniel E. Ho,Percy Liang
Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models
Language Model (LM) agents for cybersecurity that are capable of autonomously identifying vulnerabilities and executing exploits have potential to cause real-world impact. Policymakers, model providers, and researchers in the AI and cybersecurity communities are interested in quantifying the capabilities of such agents to help mitigate cyberrisk and investigate opportunities for penetration testing. Toward that end, we introduce Cybench, a framework for specifying cybersecurity tasks and evaluating agents on those tasks. We include 40 professional-level Capture the Flag (CTF) tasks from 4 distinct CTF competitions, chosen to be recent, meaningful, and spanning a wide range of difficulties. Each task includes its own description, starter files, and is initialized in an environment where an agent can execute commands and observe outputs. Since many tasks are beyond the capabilities of existing LM agents, we introduce subtasks for each task, which break down a task into intermediary steps for a more detailed evaluation. To evaluate agent capabilities, we construct a cybersecurity agent and evaluate 8 models: GPT-4o, OpenAI o1-preview, Claude 3 Opus, Claude 3.5 Sonnet, Mixtral 8x22b Instruct, Gemini 1.5 Pro, Llama 3 70B Chat, and Llama 3.1 405B Instruct. For the top performing models (GPT-4o and Claude 3.5 Sonnet), we further investigate performance across 4 agent scaffolds (structured bash, action-only, pseudoterminal, and web search). Without subtask guidance, agents leveraging Claude 3.5 Sonnet, GPT-4o, OpenAI o1-preview, and Claude 3 Opus successfully solved complete tasks that took human teams up to 11 minutes to solve. In comparison, the most difficult task took human teams 24 hours and 54 minutes to solve. Anonymized code and data are available at https://drive.google.com/file/d/1kp3H0pw1WMAH-Qyyn9WA0ZKmEa7Cr4D4 and https://drive.google.com/file/d/1BcTQ02BBR0m5LYTiK-tQmIK17_TxijIy.
****

Shicheng Xu,Liang Pang,Huawei Shen,Xueqi Cheng
A Theory for Token-Level Harmonization in Retrieval-Augmented Generation
Retrieval-augmented generation (RAG) utilizes retrieved texts to enhance large language models (LLMs). Studies show that while RAG provides valuable external information (benefit), it may also mislead LLMs (detriment) with noisy or incorrect retrieved texts. Although many existing methods attempt to preserve benefit and avoid detriment, they lack a theoretical explanation for RAG. The benefit and detriment in the next token prediction of RAG remain a 'black box' that cannot be quantified or compared in an explainable manner, so existing methods are data-driven, need additional utility evaluators or post-hoc. This paper takes the fir

st step towards providing a theory to explain and trade off the benefit and detr
iment in RAG. We model RAG as the fusion between distributions of LLMs' knowledg
e and distributions of retrieved texts. Then, we formalize the trade-off between
 the value of external knowledge (benefit) and its potential risk of misleading
LLMs (detriment) in next token prediction of RAG by distribution difference in t
his fusion. Finally, we prove that the actual effect of RAG on the token, which
is the comparison between benefit and detriment, can be predicted without any tr
aining or accessing the utility of retrieval. Based on our theory, we propose a
practical novel method, Tok-RAG, which achieves collaborative generation between
 the pure LLM and RAG at token level to preserve benefit and avoid detriment. Ex
periments in real-world tasks using LLMs such as OPT, LLaMA-2, and Mistral show
the effectiveness of our method and support our theoretical findings. Code is in
 supplemental material and will be released on GitHub after acceptance.
****
Wangbo Zhao,Yizeng Han,Jiasheng Tang,Kai Wang,Yibing Song,Gao Huang,Fan Wang,Yan
g You
Dynamic Diffusion Transformer
Diffusion Transformer (DiT), an emerging diffusion model for image generation,
has demonstrated superior performance but suffers from substantial computational
costs. Our investigations reveal that these costs stem from the static inference
paradigm, which inevitably introduces redundant computation in certain diffusion
timesteps and spatial regions. To address this inefficiency, we propose Dynamic
Diffusion Transformer (DyDiT), an architecture that dynamically adjusts its comp
u-
tation along both timestep and spatial dimensions during generation. Specificall
y,
we introduce a Timestep-wise Dynamic Width (TDW) approach that adapts model
width conditioned on the generation timesteps. In addition, we design a Spatial-
wise Dynamic Token (SDT) strategy to avoid redundant computation at unnecessary
spatial locations. Extensive experiments on various datasets and different-sized
models verify the superiority of DyDiT. Notably, with <3% additional fine-tuning
 it-
erations, our method reduces the FLOPs of DiT-XL by 51%, accelerates generation
by 1.73×, and achieves a competitive FID score of 2.07 on ImageNet.
****
Weimin Lyu,Jiachen Yao,Saumya Gupta,Lu Pang,Tao Sun,Lingjie Yi,Lijie Hu,Haibin L
ing,Chao Chen
Backdooring Vision-Language Models with Out-Of-Distribution Data
The emergence of Vision-Language Models (VLMs) represents a significant advancem
ent in integrating computer vision with Large Language Models (LLMs) to generate
 detailed text descriptions from visual inputs. Despite their growing importance
, the security of VLMs, particularly against backdoor attacks, is under explored
. Moreover, prior works often assume attackers have access to the original train
ing data, which is often unrealistic. In this paper, we address a more practical
 and challenging scenario where attackers must rely solely on Out-Of-Distributio
n (OOD) data. We introduce VLOOD (Backdoor Vision-Language Models using Out-of-D
istribution Data), a novel approach with two key contributions: (1) demonstratin
g backdoor attacks on VLMs in complex image-to-text tasks while minimizing degra
dation of the original semantics under poisoned inputs, and (2) proposing innova
tive techniques for backdoor injection without requiring any access to the origi
nal training data. Our evaluation on image captioning and visual question answer
ing (VQA) tasks confirms the effectiveness of VLOOD, revealing a critical securi
ty vulnerability in VLMs and laying the foundation for future research on securi
ng multimodal models against sophisticated threats.
****
Anh Tuan Bui,Thuy-Trang Vu,Long Tung Vuong,Trung Le,Paul Montague,Tamas Abraham,
Junae Kim,Dinh Phung
Fantastic Targets for Concept Erasure in Diffusion Models and Where To Find Them
Concept erasure has emerged as a promising technique for mitigating the risk of
harmful content generation in diffusion models by selectively unlearning undesir

able concepts. The common principle of previous works to remove a specific conce
pt is to map it to a fixed generic concept, such as a neutral concept or just an
 empty text prompt. In this paper, we demonstrate that this fixed-target strateg
y is suboptimal, as it fails to account for the impact of erasing one concept on
 the others. To address this limitation, we model the concept space as a graph a
nd empirically analyze the effects of erasing one concept on the remaining conce
pts. Our analysis uncovers intriguing geometric properties of the concept space,
 where the influence of erasing a concept is confined to a local region. Buildin
g on this insight, we propose the Adaptive Guided Erasure (AGE) method, which dy
namically selects optimal target concepts tailored to each undesirable concept,
minimizing unintended side effects. Experimental results show that AGE significa
ntly outperforms state-of-the-art erasure methods on preserving unrelated concep
ts while maintaining effective erasure performance. Our code is published at {ht
tps://github.com/tuananhbui89/Adaptive-Guided-Erasure}.
****

Jiachun Li,Pengfei Cao,Zhuoran Jin,Yubo Chen,Kang Liu,Jun Zhao
MIRAGE: Evaluating and Explaining Inductive Reasoning Process in Language Models
Inductive reasoning is an essential capability for large language models (LLMs)
to achieve higher intelligence, which requires the model to generalize rules fro
m observed facts and then apply them to unseen examples. We present {\scshape Mi
rage}, a synthetic dataset that addresses the limitations of previous work, spec
ifically the lack of comprehensive evaluation and flexible test data. In it, we
evaluate LLMs' capabilities in both the inductive and deductive stages, allowing
 for flexible variation in input distribution, task scenario, and task difficult
y to analyze the factors influencing LLMs' inductive reasoning. Based on these m
ulti-faceted evaluations, we demonstrate that the LLM is a poor rule-based reaso
ner. In many cases, when conducting inductive reasoning, they do not rely on a c
orrect rule to answer the unseen case. From the perspectives of different prompt
ing methods, observation numbers, and task forms, models tend to consistently co
nduct correct deduction without correct inductive rules. Besides, we find that L
LMs are good neighbor-based reasoners. In the inductive reasoning process, the m
odel tends to focus on observed facts that are close to the current test example
 in feature space. By leveraging these similar examples, the model maintains str
ong inductive capabilities within a localized region, significantly improving it
s deductive performance.
****

Yudong Zhang,Xuan Yu,Xu Wang,Zhaoyang Sun,Chen Zhang,Pengkun Wang,Yang Wang
COFlowNet: Conservative Constraints on Flows Enable High-Quality Candidate Gener
ation
Generative flow networks (GFlowNets) have been considered as powerful tools for
generating candidates with desired properties. Given that evaluating the propert
y of candidates can be complex and time-consuming, existing GFlowNets train prox
y models for efficient online evaluation. However, the performance of proxy mode
ls is heavily dependent on the amount of data and is of considerable uncertainty
. Therefore, it is of great interest that how to develop an offline GFlowNet tha
t does not rely on online evaluation. Under the offline setting, the limited dat
a results in an insufficient exploration of state space. The insufficient explor
ation means that offline GFlowNets can hardly generate satisfying candidates out
 of the distribution of training data. Therefore, it is critical to restrict the
 offline model to act in the distribution of training data. The distinctive trai
ning goal of GFlownets poses a unique challenge for making such restrictions. Ta
ckling the challenge, we propose Conservative Offline GFlowNet (COFlowNet) in th
is paper. We define unsupported flow, edges containing unseen states in training
 data. Models can learn extremely little knowledge about unsupported flow from t
raining data. By constraining the model from exploring unsupported flows, we res
trict COFlowNet to explore as optimal trajectories on the training set as possib
le, thus generating better candidates. In order to improve the diversity of cand
idates, we further introduce a quantile version of unsupported flow restriction.
 Experimental results on several widely-used datasets validate the effectiveness
 of COFlowNet in generating high-scored and diverse candidates. All implementati

ons are available at https://github.com/yuxuan9982/COflownet.
****
Hongliang Chi,Wei Jin,Charu C. Aggarwal,Yao Ma
Precedence-Constrained Winter Value for Effective Graph Data Valuation
Data valuation is essential for quantifying data’s worth, aiding in assessing da
ta quality and determining fair compensation. While existing data valuation meth
ods have proven effective in evaluating the value of Euclidean data, they face l
imitations when applied to the increasingly popular graph-structured data. Parti
cularly, graph data valuation introduces unique challenges, primarily stemming f
rom the intricate dependencies among nodes and the exponential growth in value e
stimation costs. To address the challenging problem of graph data valuation, we
put forth an innovative solution, Precedence-Constrained Winter (PC-Winter) Valu
e, to account for the complex graph structure. Furthermore, we develop a variety
 of strategies to address the computational challenges and enable efficient appr
oximation of PC-Winter. Extensive experiments demonstrate the effectiveness of P
C-Winter across diverse datasets and tasks.
****
Pit Neitemeier,Björn Deiseroth,Constantin Eichenberg,Lukas Balles
Hierarchical Autoregressive Transformers: Combining Byte- and Word-Level Process
ing for Robust, Adaptable Language Models
Tokenization is a fundamental step in natural language processing, breaking text
 into units that computational models can process. While learned subword tokeniz
ers have become the de-facto standard, they present challenges such as large voc
abularies, limited adaptability to new domains or languages, and sensitivity to
spelling errors and variations. To overcome these limitations, we investigate a
hierarchical architecture for autoregressive language modelling that combines ch
aracter-level and word-level processing. It employs a lightweight character-leve
l encoder to convert character sequences into word embeddings, which are then pr
ocessed by a word-level backbone model and decoded back into characters via a co
mpact character-level decoder. This method retains the sequence compression bene
fits of word-level tokenization without relying on a rigid, predefined vocabular
y. We demonstrate, at scales up to 7 billion parameters, that hierarchical trans
formers match the downstream task performance of subword-tokenizer-based models
while exhibiting significantly greater robustness to input perturbations. Additi
onally, during continued pretraining on an out-of-domain language, our model tra
ins almost twice as fast, achieves superior performance on the target language,
and retains more of its previously learned knowledge. Hierarchical transformers
pave the way for NLP systems that are more robust, flexible, and generalizable a
cross languages and domains.
****
Tiansheng Huang,Sihao Hu,Fatih Ilhan,Selim Furkan Tekin,Ling Liu
Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating
Harmful Perturbation
Harmful fine-tuning attack poses serious safety concerns for large language mode
ls' fine-tuning-as-a-service. While existing defenses have been proposed to miti
gate the issue, their performances are still far away from satisfactory, and the
 root cause of the problem has not been fully recovered. To this end, we in this
 paper show that \textit{harmful perturbation} over the model weights could be a
 probable cause of alignment-broken. In order to attenuate the negative impact o
f harmful perturbation, we propose an alignment-stage solution, dubbed Booster.
Technically, along with the original alignment loss, we append a loss regulariz
er in the alignment stage's optimization. The regularizer ensures that the model
's harmful loss reduction after the simulated harmful perturbation is attenuated
, thereby mitigating the subsequent fine-tuning risk. Empirical results show
 that Booster can effectively reduce the harmful score of the fine-tuned models
while maintaining the performance of downstream tasks. Our code is available at
https://github.com/git-disl/Booster
****
Wenhao Chai,Enxin Song,Yilun Du,Chenlin Meng,Vashisht Madhavan,Omer Bar-Tal,Jenq
-Neng Hwang,Saining Xie,Christopher D Manning

AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark
Video detailed captioning is a key task which aims to generate comprehensive and coherent textual descriptions of video content, benefiting both video understanding and generation. In this paper, we propose AuroraCap, a video captioner based on a large multimodal model. We follow the simplest architecture design without additional parameters for temporal modeling. To address the overhead caused by lengthy video sequences, we implement the token merging strategy, reducing the number of input visual tokens. Surprisingly, we found that this strategy results in little performance loss. AuroraCap shows superior performance on various video and image captioning benchmarks, for example, obtaining a CIDEr of 88.9 on Flickr30k, beating GPT-4V (55.3) and Gemini-1.5 Pro (82.2). However, existing video caption benchmarks only include simple descriptions, consisting of a few dozen words, which limits research in this field. Therefore, we develop VDC, a video detailed captioning benchmark with over one thousand carefully annotated structured captions. In addition, we propose a new LLM-assisted metric VDCscore for bettering evaluation, which adopts a divide-and-conquer strategy to transform long caption evaluation into multiple short question-answer pairs. With the help of human Elo ranking, our experiments show that this benchmark better correlates with human judgments of video detailed captioning quality.
****
Ce Zhang,Zifu Wan,Zhehan Kan,Martin Q. Ma,Simon Stepputtis,Deva Ramanan,Russ Salakhutdinov,Louis-Philippe Morency,Katia P. Sycara,Yaqi Xie
Self-Correcting Decoding with Generative Feedback for Mitigating Hallucinations in Large Vision-Language Models
While recent Large Vision-Language Models (LVLMs) have shown remarkable performance in multi-modal tasks, they are prone to generating hallucinatory text responses that do not align with the given visual input, which restricts their practical applicability in real-world scenarios. In this work, inspired by the observation that the text-to-image generation process is the inverse of image-conditioned response generation in LVLMs, we explore the potential of leveraging text-to-image generative models to assist in mitigating hallucinations in LVLMs. We discover that generative models can offer valuable self-feedback for mitigating hallucinations at both the response and token levels. Building on this insight, we introduce self-correcting Decoding with Generative Feedback (DeGF), a novel training-free algorithm that incorporates feedback from text-to-image generative models into the decoding process to effectively mitigate hallucinations in LVLMs. Specifically, DeGF generates an image from the initial response produced by LVLMs, which acts as an auxiliary visual reference and provides self-feedback to verify and correct the initial response through complementary or contrastive decoding. Extensive experimental results validate the effectiveness of our approach in mitigating diverse types of hallucinations, consistently surpassing state-of-the-art methods across six benchmarks. Code is available at https://github.com/zhangce01/DeGF.
****
Xuehai He,Weixi Feng,Kaizhi Zheng,Yujie Lu,Wanrong Zhu,Jiachen Li,Yue Fan,Jianfeng Wang,Linjie Li,Zhengyuan Yang,Kevin Lin,William Yang Wang,Lijuan Wang,Xin Eric Wang
MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos
Multimodal Language Language Models (MLLMs) demonstrate the emerging abilities of "world models"---interpreting and reasoning about complex real-world dynamics. To assess these abilities, we posit videos are the ideal medium, as they encapsulate rich representations of real-world dynamics and causalities. To this end, we introduce MMWorld, a new benchmark for multi-discipline, multi-faceted multimodal video understanding. MMWorld distinguishes itself from previous video understanding benchmarks with two unique advantages: (1) multi-discipline, covering various disciplines that often require domain expertise for comprehensive understanding; (2) multi-faceted reasoning, including explanation, counterfactual thinking, future prediction, etc. MMWorld consists of a human-annotated dataset to evaluate MLLMs with questions about the whole videos and a synthetic dataset to analyze MLLMs within a single modality of perception. Together, MMWorld encompasse

s 1,910 videos across seven broad disciplines and 69 subdisciplines, complete wi
th 6,627 question-answer pairs and associated captions. The evaluation includes
4 proprietary and 11 open-source MLLMs, which struggle on MMWorld (e.g., GPT-4o
performs the best with only 62.5% accuracy), showing large room for improvement.
Further ablation studies reveal other interesting findings such as models' diff
erent skill sets from humans. We hope MMWorld can serve as an essential step tow
ards world model evaluation in videos.
****
Jun Hyeong Kim,Seonghwan Kim,Seokhyun Moon,Hyeongwoo Kim,Jeheon Woo,Woo Youn Kim
Discrete Diffusion Schrödinger Bridge Matching for Graph Transformation
Transporting between arbitrary distributions is a fundamental goal in generative
modeling.
Recently proposed diffusion bridge models provide a potential solution, but they
rely on a joint distribution that is difficult to obtain in practice.
Furthermore, formulations based on continuous domains limit their applicability
to discrete domains such as graphs.
To overcome these limitations, we propose Discrete Diffusion Schrödinger Bridge
Matching (DDSBM), a novel framework that utilizes continuous-time Markov chains
to solve the SB problem in a high-dimensional discrete state space.
Our approach extends Iterative Markovian Fitting to discrete domains, and we hav
e proved its convergence to the SB.
Furthermore, we adapt our framework for the graph transformation, and show that
our design choice of underlying dynamics characterized by independent modificati
ons of nodes and edges can be interpreted as the entropy-regularized version of
optimal transport with a cost function described by the graph edit distance.
To demonstrate the effectiveness of our framework, we have applied DDSBM to mole
cular optimization in the field of chemistry.
Experimental results demonstrate that DDSBM effectively optimizes molecules' pro
perty-of-interest with minimal graph transformation, successfully retaining othe
r features. Source code is available [here](https://github.com/junhkim1226/DDSBM
).
****
Sang-Hoon Lee,Ha-Yeong Choi,Seong-Whan Lee
PeriodWave: Multi-Period Flow Matching for High-Fidelity Waveform Generation
Recently, universal waveform generation tasks have been investigated conditioned
on various out-of-distribution scenarios. Although one-step GAN-based methods h
ave shown their strength in fast waveform generation, they are vulnerable to tra
in-inference mismatch scenarios such as two-stage text-to-speech. Meanwhile, dif
fusion-based models have shown their powerful generative performance in other do
mains; however, they stay out of the limelight due to slow inference speed in wa
veform generation tasks. Above all, there is no generator architecture that can
explicitly disentangle the natural periodic features of high-resolution waveform
signals. In this paper, we propose PeriodWave, a novel universal waveform gener
ation model from Mel-spectrogram and neural audio codec. First, we introduce a p
eriod-aware flow matching estimator that effectively captures the periodic featu
res of the waveform signal when estimating the vector fields. Additionally, we u
tilize a multi-period estimator that avoids overlaps to capture different period
ic features of waveform signals. Although increasing the number of periods can i
mprove the performance significantly, this requires more computational costs. To
reduce this issue, we also propose a single period-conditional universal estima
tor that can feed-forward parallel by period-wise batch inference. Additionally,
we first introduce FreeU to reduce the high-frequency noise for waveform genera
tion. Furthermore, we demonstrate the effectiveness of the proposed method in ne
ural audio codec decoding task, and present the streaming generation framework o
f non-autoregressive model for speech language models. The experimental results
demonstrated that our model outperforms the previous models in reconstruction ta
sks from Mel-spectrogram and discrete token, and text-to-speech tasks. Source co
de is available at https://github.com/sh-lee-prml/PeriodWave
****
Yi Ren,Danica J. Sutherland

Learning Dynamics of LLM Finetuning
Learning dynamics, which describes how the learning of specific training example s influences the model's predictions on other examples,
gives us a powerful tool for understanding the behavior of deep learning systems . We study the learning dynamics of large language models during different types of finetuning, by analyzing the step-wise decomposition of how influence accumu lates among different potential responses. Our framework allows a uniform interp retation of many interesting observations about the training of popular algorith ms for both instruction tuning and preference tuning. In particular, we propose a hypothetical explanation of why specific types of hallucination are strengthen ed after finetuning, e.g., the model might use phrases or facts in the response for question B to answer question A, or the model might keep repeating similar s imple phrases when generating responses. We also extend our framework and highli ght a unique ``squeezing effect'' to explain a previously observed phenomenon in off-policy direct preference optimization (DPO), where running DPO for too long makes even the desired outputs less likely. This framework also provides insigh ts into where the benefits of on-policy DPO and other variants come from. The an alysis not only provides a novel perspective of understanding LLM's finetuning b ut also inspires a simple, effective method to improve alignment performance.
****

Samet Demir,Zafer Dogan
Asymptotic Analysis of Two-Layer Neural Networks after One Gradient Step under G aussian Mixtures Data with Structure
In this work, we study the training and generalization performance of two-layer neural networks (NNs) after one gradient descent step under structured data mode led by Gaussian mixtures. While previous research has extensively analyzed this model under isotropic data assumption, such simplifications overlook the complex ities inherent in real-world datasets. Our work addresses this limitation by ana lyzing two-layer NNs under Gaussian mixture data assumption in the asymptoticall y proportional limit, where the input dimension, number of hidden neurons, and s ample size grow with finite ratios. We characterize the training and generalizat ion errors by leveraging recent advancements in Gaussian universality. Specifica lly, we prove that a high-order polynomial model performs equivalent to the non-linear neural networks under certain conditions. The degree of the equivalent mo del is intricately linked to both the "data spread" and the learning rate employ ed during one gradient step. Through extensive simulations, we demonstrate the e quivalence between the original model and its polynomial counterpart across vari ous regression and classification tasks. Additionally, we explore how different properties of Gaussian mixtures affect learning outcomes. Finally, we illustrate experimental results on Fashion-MNIST classification, indicating that our findi ngs can translate to realistic data.
****

Han Wang,Yanjie Wang,Yang Li,Can Huang
GLOMA: Global Video Text Spotting with Morphological Association
Video Text Spotting (VTS) is a fundamental visual task that aims to predict the trajectories and content of texts in a video. Previous works usually conduct loc al associations and apply IoU-based distance and complex post-processing procedu res to boost performance, ignoring the abundant temporal information and the mor phological characteristics in VTS. In this paper, we propose \model{} to model t he tracking problem as global associations and utilize the Gaussian Wasserstein distance to guide the morphological correlation between frames. Our main contrib utions can be summarized as three folds. 1). We propose a Transformer-based glob al tracking method \model{} for VTS and associate multiple frames simultaneously . 2). We introduce a Wasserstein distance-based method to conduct positional ass ociations between frames. 3). We conduct extensive experiments on public dataset s. On the ICDAR2015 video dataset, \model{} achieves \textbf{56.0} MOTA with \te xtbf{4.6} absolute improvement compared with the previous SOTA method and outper forms the previous Transformer-based method by a significant \textbf{8.3} MOTA.
****

Jihwan Yoon,Sangbeom Han,Jaeseok Oh,Minsik Lee

SplineGS: Learning Smooth Trajectories in Gaussian Splatting for Dynamic Scene Reconstruction
Reconstructing complex scenes with deforming objects for novel view synthesis is a challenging task. Recent works have addressed this with 3D Gaussian Splatting, which effectively reconstructs static scenes with high quality in short training time, by adding specialized modules for the deformations of Gaussian blobs. However, designing an effective deformation module that incorporates appropriate spatiotemporal inductive biases still remains unresolved. To address this issue, we propose SplineGS in this paper, which utilizes non-uniform rational B-splines (NURBS), an extension of B-spline, to represent temporally smooth deformation. A set of representative trajectories are learned based on NURBS, and the individual trajectories of Gaussian blobs are represented as linear combinations of these trajectories for spatial smoothness. The weights of the combinations are trained based on a multi-resolution hash table and an MLP, with the positions of the Gaussian blobs as the keys. Thanks to this design, the proposed method does not need any regularizers for trajectories, which enables efficient training. Experiments demonstrate that the proposed method provides competitive performance over the existing methods with much shorter training time.
****
Wenxuan Wang,Quan Sun,Fan Zhang,Yepeng Tang,Jing Liu,Xinlong Wang
Diffusion Feedback Helps CLIP See Better
Contrastive Language-Image Pre-training (CLIP), which excels at abstracting open-world representations across domains and modalities, has become a foundation for a variety of vision and multimodal tasks. However, recent studies reveal that CLIP has severe visual shortcomings, such as which can hardly distinguish orientation, quantity, color, structure, etc. These visual shortcomings also limit the perception capabilities of multimodal large language models (MLLMs) built on CLIP. The main reason could be that the image-text pairs used to train CLIP are inherently biased, due to the lack of the distinctiveness of the text and the diversity of images. In this work, we present a simple post-training approach for CLIP models, which largely overcomes its visual shortcomings via a self-supervised diffusion process. We introduce DIVA, which uses the DIffusion model as a Visual Assistant for CLIP. Specifically, DIVA leverages generative feedback from text-to-image diffusion models to optimize CLIP representations, with only images (without corresponding text). We demonstrate that DIVA improves CLIP's performance on the challenging MMVP-VLM benchmark which assesses fine-grained visual abilities to a large extent (e.g., 3-7%), and enhances the performance of MLLMs and vision models on multimodal understanding and segmentation tasks. Extensive evaluation on 29 image classification and retrieval benchmarks confirms that our framework preserves CLIP's strong zero-shot capabilities. The code is publicly available at https://github.com/baaivision/DIVA.
****
Yiming Xie,Chun-Han Yao,Vikram Voleti,Huaizu Jiang,Varun Jampani
SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency
We present Stable Video 4D (SV4D) — a latent video diffusion model for multi-frame and multi-view consistent dynamic 3D content generation. Unlike previous methods that rely on separately trained generative models for video generation and novel view synthesis, we design a unified diffusion model to generate novel view videos of dynamic 3D objects. Specifically, given a monocular reference video, SV4D generates novel views for each video frame that are temporally consistent. We then use the generated novel view videos to optimize an implicit 4D representation (dynamic NeRF) efficiently, without the need for cumbersome SDS-based optimization used in most prior works. To train our unified novel view video generation model, we curate a dynamic 3D object dataset from the existing Objaverse dataset. Extensive experimental results on multiple datasets and user studies demonstrate SV4D's state-of-the-art performance on novel-view video synthesis as well as 4D generation compared to prior works. Project page: https://sv4d.github.io.
****
Yuxian Gu,Hao Zhou,Fandong Meng,Jie Zhou,Minlie Huang
MiniPLM: Knowledge Distillation for Pre-training Language Models

Knowledge distillation (KD) is widely used to train small, high-performing stude
nt language models (LMs) using large teacher LMs.
While effective in fine-tuning, KD during pre-training faces efficiency, flexibi
lity, and effectiveness issues.
Existing methods either incur high computational costs due to online teacher inf
erence, require tokenization matching between teacher and student LMs, or risk l
osing the difficulty and diversity of the teacher-generated training data.
In this work, we propose **MiniPLM**, a KD framework for pre-training LMs by ref
ining the training data distribution with the teacher LM's knowledge.
For efficiency, MiniPLM performs offline teacher inference, allowing KD for mult
iple student LMs without adding training costs.
For flexibility, MiniPLM operates solely on the training corpus, enabling KD acr
oss model families.
For effectiveness, MiniPLM leverages the differences between large and small LMs
 to enhance the training data difficulty and diversity, helping student LMs acqu
ire versatile and sophisticated knowledge.
Extensive experiments demonstrate that MiniPLM boosts the student LMs' performan
ce on 9 common downstream tasks, improves language modeling capabilities, and re
duces pre-training computation.
The benefit of MiniPLM extends to larger training scales, evidenced by the scali
ng curve extrapolation.
Further analysis reveals that MiniPLM supports KD across model families and enha
nces the pre-training data utilization. Our code, data, and models can be found
at https://github.com/thu-coai/MiniPLM.
****

Chen Ma,Xinjie Xu,Shuyu Cheng,Qi Xuan
Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors
One of the most practical and challenging types of black-box adversarial attacks
 is the hard-label attack, where only top-1 predicted labels are available. One
effective approach is to search for the optimal ray direction from the benign im
age that minimizes the ℓ_p norm distance to the adversarial region. The uni
que advantage of this approach is that it transforms the hard-label attack into
a continuous optimization problem. The objective function value is the ray's rad
ius and can be obtained through a binary search with high query cost. Existing m
ethods use a "sign trick" in gradient estimation to reduce queries. In this pape
r, we theoretically analyze the quality of this gradient estimation, proposing a
 novel prior-guided approach to improve ray search efficiency, based on theoreti
cal and experimental analysis. Specifically, we utilize the transfer-based prior
s from surrogate models, and our gradient estimators appropriately integrate the
m by approximating the projection of the true gradient onto the subspace spanned
 by these priors and some random directions, in a query-efficient way. We theore
tically derive the expected cosine similarity between the obtained gradient esti
mators and the true gradient, and demonstrate the improvement brought by using p
riors. Extensive experiments on the ImageNet and CIFAR-10 datasets show that our
 approach significantly outperforms 11 state-of-the-art methods in terms of quer
y efficiency.
****

Beatrice Bevilacqua,Joshua Robinson,Jure Leskovec,Bruno Ribeiro
Holographic Node Representations: Pre-training Task-Agnostic Node Embeddings
Large general purpose pre-trained models have revolutionized computer vision and
 natural language understanding. However, the development of general purpose pre
-trained Graph Neural Networks (GNNs) lags behind other domains due to the lack
of suitable generalist node representations. Existing GNN architectures are ofte
n tailored to specific task orders, such as node-level, link-level, or higher-or
der tasks, because different tasks require distinct permutation symmetries, whic
h are difficult to reconcile within a single model. In this paper, we propose _h
olographic node representations_, a new blueprint for node representations capab
le of solving tasks of any order. Holographic node representations have two key
components: (1) a task-agnostic expansion map, which produces highly expressive,
 high-dimensional embeddings, free from node-permutation symmetries, to be fed i

nto (2) a reduction map that carefully reintroduces the relevant permutation sym
metries to produce low-dimensional, task-specific embeddings. We show that well-
constructed expansion maps enable simple and efficient reduction maps, which can
be adapted for any task order. Empirical results show that holographic node rep
resentations can be effectively pre-trained and reused across tasks of varying o
rders, yielding up to 100% relative performance improvement, including in cases
where prior methods fail entirely.
****
Wilson Yan,Volodymyr Mnih,Aleksandra Faust,Matei Zaharia,Pieter Abbeel,Hao Liu
ElasticTok: Adaptive Tokenization for Image and Video
Efficient video tokenization remains a key bottleneck in learning general purpos
e vision models that are capable of processing long video sequences. Prevailing
approaches are restricted to encoding videos to a fixed number of tokens, where
too few tokens will result in overly lossy encodings, and too many tokens will r
esult in prohibitively long sequence lengths. In this work, we introduce Elastic
Tok, a method that conditions on prior frames to adaptively encode a frame into
a variable number of tokens. To enable this in a computationally scalable way, w
e propose a masking technique that drops a random number of tokens at the end of
each frames's token encoding. During inference, ElasticTok can dynamically allo
cate tokens when needed -- more complex data can leverage more tokens, while sim
pler data only needs a few tokens. Our empirical evaluations on images and video
demonstrate the effectiveness of our approach in efficient token usage, paving
the way for future development of more powerful multimodal models, world models,
and agents. Video examples of using ElasticTok can be found on our website: htt
p://largeworldmodel.github.io/elastictok
****
Guanxing Lu,Ziwei Wang,Changliu Liu,Jiwen Lu,Yansong Tang
ThinkBot: Embodied Instruction Following with Thought Chain Reasoning
Embodied Instruction Following (EIF) requires agents to complete human instructi
on by interacting objects in complicated surrounding environments. Conventional
methods directly consider the sparse human instruction to generate action plans
for agents, which usually fail to achieve human goals because of the instruction
incoherence in action descriptions. On the contrary, we propose ThinkBot that r
easons the thought chain in human instruction to recover the missing action desc
riptions, so that the agent can successfully complete human goals by following t
he coherent instruction. Specifically, we first design an instruction completer
based on large language models to recover the missing actions with interacted ob
jects between consecutive human instruction, where the perceived surrounding env
ironments and the completed sub-goals are considered for instruction completion.
Based on the partially observed scene semantic maps, we present an object local
izer to infer the position of interacted objects and the related Bayesian uncert
ainty for close-loop planning. Extensive experiments in the simulated environmen
t show that our ThinkBot outperforms the state-of-the-art EIF methods by a sizab
le margin in both success rate and execution efficiency. Project page: https://g
uanxinglu.github.io/thinkbot/.
****
Samuel Garcin,Trevor McInroe,Pablo Samuel Castro,Christopher G. Lucas,David Abel
,Prakash Panangaden,Stefano V Albrecht
Studying the Interplay Between the Actor and Critic Representations in Reinforce
ment Learning
Extracting relevant information from a stream of high-dimensional observations i
s a central challenge for deep reinforcement learning agents. Actor-critic algor
ithms add further complexity to this challenge, as it is often unclear whether t
he same information will be relevant to both the actor and the critic. To this e
nd, we here explore the principles that underlie effective representations for a
n actor and for a critic. We focus our study on understanding whether an actor a
nd a critic will benefit from a decoupled, rather than shared, representation.
Our primary finding is that when decoupled, the representations for the actor an
d critic systematically specialise in extracting different types of information
from the environment---the actor's representation tends to focus on action-relev

ant information, while the critic's representation specialises in encoding value and dynamics information. Finally, we demonstrate how these insights help select representation learning objectives that play into the actor's and critic's respective knowledge specialisations, and improve performance in terms of agent returns.
****
Poojan Chetan Shah,Ragesh Jaiswal
Quantum (Inspired) D^2-sampling with Applications
D^2-sampling is a fundamental component of sampling-based clustering algorithms such as k-means++.
Given a dataset $V \subset \mathbb{R}^d$ with N points and a center set $C \subset \mathbb{R}^d$, D^2-sampling refers to picking a point from V where the sampling probability of a point is proportional to its squared distance from the nearest center in C.
The popular k-means++ algorithm is simply a k-round D^2-sampling process, which runs in $O(Nkd)$ time and gives $O(\log{k})$-approximation in expectation for the k-means problem.
In this work, we give a quantum algorithm for (approximate) D^2-sampling in the QRAM model that results in a quantum implementation of k-means++ with a running time $\tilde{O}(\zeta^2 k^2)$.
Here ζ is the aspect ratio (i.e., largest to smallest interpoint distance) and $\tilde{O}$ hides polylogarithmic factors in N, d, k.
It can be shown through a robust approximation analysis of k-means++ that the quantum version preserves its $O(\log{k})$ approximation guarantee.
Further, we show that our quantum algorithm for D^2-sampling can be dequantized using the sample-query access model of Tang (PhD Thesis, Ewin Tang, University of Washington, 2023). This results in a fast quantum-inspired classical implementation of k-means++, which we call QI-k-means++, with a running time $O(Nd) + \tilde{O}(\zeta^2k^2d)$, where the $O(Nd)$ term is for setting up the sample-query access data structure.
Experimental investigations show promising results for QI-k-means++ on large datasets with bounded aspect ratio.
Finally, we use our quantum D^2-sampling with the known $ D^2$-sampling-based classical approximation scheme
to obtain the first quantum approximation scheme for the k-means problem with polylogarithmic running time dependence on N.
****
Ni Zhang,Jingfeng Yang,Zhiguang Cao,Xu Chi
Adversarial Generative Flow Network for Solving Vehicle Routing Problems
Recent research into solving vehicle routing problems (VRPs) has gained significant traction, particularly through the application of deep (reinforcement) learning for end-to-end solution construction. However, many current construction-based neural solvers predominantly utilize Transformer architectures, which can face scalability challenges and struggle to produce diverse solutions. To address these limitations, we introduce a novel framework beyond Transformer-based approaches, i.e., Adversarial Generative Flow Networks (AGFN). This framework integrates the generative flow network (GFlowNet)—a probabilistic model inherently adept at generating diverse solutions (routes)—with a complementary model for discriminating (or evaluating) the solutions. These models are trained alternately in an adversarial manner to improve the overall solution quality, followed by a proposed hybrid decoding method to construct the solution. We apply the AGFN framework to solve the capacitated vehicle routing problem (CVRP) and travelling salesman problem (TSP), and our experimental results demonstrate that AGFN surpasses the popular construction-based neural solvers, showcasing strong generalization capabilities on synthetic and real-world benchmark instances.
****
Xirui Li,Charles Herrmann,Kelvin C.K. Chan,Yinxiao Li,Deqing Sun,Chao Ma,Ming-Hsuan Yang
A Simple Approach to Unifying Diffusion-based Conditional Generation
Recent progress in image generation has sparked research into controlling these

models through condition signals, with various methods addressing specific chall
enges in conditional generation. Instead of proposing another specialized techni
que, we introduce a simple, unified framework to handle diverse conditional gene
ration tasks involving a specific image-condition correlation. By learning a joi
nt distribution over a correlated image pair (e.g. image and depth) with a diffu
sion model, our approach enables versatile capabilities via different inference-
time sampling schemes, including controllable image generation (e.g. depth to im
age), estimation (e.g. image to depth), signal guidance, joint generation (image
 \& depth), and coarse control. Previous attempts at unification often introduce
 complexity through multi-stage training, architectural modification, or increas
ed parameter counts. In contrast, our simplified formulation requires a single,
computationally efficient training stage, maintains the standard model input, an
d adds minimal learned parameters (15% of the base model). Moreover, our model s
upports additional capabilities like non-spatially aligned and coarse conditioni
ng. Extensive results show that our single model can produce comparable results
with specialized methods and better results than prior unified methods. We also
demonstrate that multiple models can be effectively combined for multi-signal co
nditional generation.
****
Yuankai Luo,Hongkang Li,Qijiong Liu,Lei Shi,Xiao-Ming Wu
Node Identifiers: Compact, Discrete Representations for Efficient Graph Learning
We present a novel end-to-end framework that generates highly compact (typically
 6-15 dimensions), discrete (int4 type), and interpretable node representations—
termed node identifiers (node IDs)—to tackle inference challenges on large-scale
 graphs. By employing vector quantization, we compress continuous node embedding
s from multiple layers of a Graph Neural Network (GNN) into discrete codes, appl
icable under both self-supervised and supervised learning paradigms. These node
IDs capture high-level abstractions of graph data and offer interpretability tha
t traditional GNN embeddings lack. Extensive experiments on 34 datasets, encompa
ssing node classification, graph classification, link prediction, and attributed
 graph clustering tasks, demonstrate that the generated node IDs significantly e
nhance speed and memory efficiency while achieving competitive performance compa
red to current state-of-the-art methods. Our source code is available at https:/
/github.com/LUOyk1999/NodeID.
****
Dvir Samuel,Barak Meiri,Haggai Maron,Yoad Tewel,Nir Darshan,Shai Avidan,Gal Chec
hik,Rami Ben-Ari
Lightning-Fast Image Inversion and Editing for Text-to-Image Diffusion Models
Diffusion inversion is the problem of taking an image and a text prompt that des
cribes it and finding a noise latent that would generate the exact same image.
Most current deterministic inversion techniques operate by approximately solving
 an implicit equation and may converge slowly or yield poor reconstructed images
. We formulate the problem by finding the roots of an implicit equation and dev
lop a method to solve it efficiently. Our solution is based on Newton-Raphson (N
R), a well-known technique in numerical analysis. We show that a vanilla applica
tion of NR is computationally infeasible while naively transforming it to a comp
utationally tractable alternative tends to converge to out-of-distribution solut
ions, resulting in poor reconstruction and editing. We therefore derive an effic
ient guided formulation that fastly converges and provides high-quality reconstr
uctions and editing. We showcase our method on real image editing with three pop
ular open-sourced diffusion models: Stable Diffusion, SDXL-Turbo, and Flux with
different deterministic schedulers. Our solution, **Guided Newton-Raphson Invers
ion**, inverts an image within 0.4 sec (on an A100 GPU) for few-step models (SDX
L-Turbo and Flux.1),
opening the door for interactive image editing. We further show improved results
 in image interpolation and generation of rare objects.
****
Shengran Hu,Cong Lu,Jeff Clune
Automated Design of Agentic Systems
Researchers are investing substantial effort in developing powerful general-purp

ose agents, wherein Foundation Models are used as modules within agentic systems (e.g. Chain-of-Thought, Self-Reflection, Toolformer). However, the history of machine learning teaches us that hand-designed solutions are eventually replaced by learned solutions. We describe a newly forming research area, Automated Design of Agentic Systems (ADAS), which aims to automatically create powerful agentic system designs, including inventing novel building blocks and/or combining them in new ways. We further demonstrate that there is an unexplored yet promising approach within ADAS where agents can be defined in code and new agents can be automatically discovered by a meta agent programming ever better ones in code. Given that programming languages are Turing Complete, this approach theoretically enables the learning of any possible agentic system: including novel prompts, tool use, workflows, and combinations thereof. We present a simple yet effective algorithm named Meta Agent Search to demonstrate this idea, where a meta agent iteratively programs interesting new agents based on an ever-growing archive of previous discoveries. Through extensive experiments across multiple domains including coding, science, and math, we show that our algorithm can progressively invent agents with novel designs that greatly outperform state-of-the-art hand-designed agents. Importantly, we consistently observe the surprising result that agents invented by Meta Agent Search maintain superior performance even when transferred across domains and models, demonstrating their robustness and generality. Provided we develop it safely, our work illustrates the potential of an exciting new research direction toward automatically designing ever-more powerful agentic systems to benefit humanity.
****

Mengqi Zhang,Xiaotian Ye,Qiang Liu,Shu Wu,Pengjie Ren,Zhumin Chen
Uncovering Overfitting in Large Language Model Editing
Knowledge editing has been proposed as an effective method for updating and correcting the internal knowledge of Large Language Models (LLMs). However, existing editing methods often struggle with complex tasks, such as multi-hop reasoning. In this paper, we identify and investigate the phenomenon of Editing Overfit, where edited models assign disproportionately high probabilities to the edit target, hindering the generalization of new knowledge in complex scenarios. We attribute this issue to the current editing paradigm, which places excessive emphasis on the direct correspondence between the input prompt and the edit target for each edit sample. To further explore this issue, we introduce a new benchmark, EVOKE (EValuation of Editing Overfit in Knowledge Editing), along with fine-grained evaluation metrics. Through comprehensive experiments and analysis, we demonstrate that Editing Overfit is prevalent in current editing methods and that common overfitting mitigation strategies are ineffective in knowledge editing. To overcome this, inspired by LLMs’ knowledge recall mechanisms, we propose a new plug-and-play strategy called Learn the Inference (LTI), which introduce a Multi-stage Inference Constraint module to guide the edited models in recalling new knowledge similarly to how unedited LLMs leverage knowledge through in-context learning. Extensive experimental results across a wide range of tasks validate the effectiveness of LTI in mitigating Editing Overfit.
****

Hossein Mirzaei,Mojtaba Nafez,Jafar Habibi,Mohammad Sabokrou,Mohammad Hossein Rohban
Adversarially Robust Anomaly Detection through Spurious Negative Pair Mitigation
Despite significant progress in Anomaly Detection (AD), the robustness of existing detection methods against adversarial attacks remains a challenge, compromising their reliability in critical real-world applications such as autonomous driving. This issue primarily arises from the AD setup, which assumes that training data is limited to a group of unlabeled normal samples, making the detectors vulnerable to adversarial anomaly samples during testing. Additionally, implementing adversarial training as a safeguard encounters difficulties, such as formulating an effective objective function without access to labels. An ideal objective function for adversarial training in AD should promote strong perturbations both within and between the normal and anomaly groups to maximize margin between normal and anomaly distribution. To address these issues, we first propose crafting

a pseudo-anomaly group derived from normal group samples. Then, we demonstrate that adversarial training with contrastive loss could serve as an ideal objectiv e function, as it creates both inter- and intra-group perturbations. However, we notice that spurious negative pairs compromise the conventional contrastive los s for achieving robust AD. Spurious negative pairs are those that should be mapp ed closely but are erroneously separated. These pairs introduce noise and misgui de the direction of inter-group adversarial perturbations. To overcome the effec t of spurious negative pairs, we define opposite pairs and adversarially pull th em apart to strengthen inter-group perturbations. Experimental results demonstra te our superior performance in both clean and adversarial scenarios, with a 26.1 % improvement in robust detection across various challenging benchmark datasets.
****

Xuan Shen,Hangyu Zheng,Yifan Gong,Zhenglun Kong,Changdi Yang,Zheng Zhan,Yushu Wu ,Xue Lin,Yanzhi Wang,Pu Zhao,Wei Niu

Sparse Learning for State Space Models on Mobile

Transformer models have been widely investigated in different domains by providi ng long-range dependency handling and global contextual awareness, driving the d evelopment of popular AI applications such as ChatGPT, Gemini, and Alexa.

State Space Models (SSMs) have emerged as strong contenders in the field of sequ ential modeling, challenging the dominance of Transformers. SSMs incorporate a s elective mechanism that allows for dynamic parameter adjustment based on input d ata, enhancing their performance.

However, this mechanism also comes with increasing computational complexity and bandwidth demands, posing challenges for deployment on resource-constraint mobil e devices.

To address these challenges without sacrificing the accuracy of the selective me chanism, we propose a sparse learning framework that integrates architecture-awa re compiler optimizations. We introduce an end-to-end solution--$\mathbf{C}_4^n$ kernel sparsity, which prunes n elements from every four contiguous weights, and develop a compiler-based acceleration solution to ensure execution efficienc y for this sparsity on mobile devices.

Based on the kernel sparsity, our framework generates optimized sparse models ta rgeting specific sparsity or latency requirements for various model sizes. We fu rther leverage pruned weights to compensate for the remaining weights, enhancin g downstream task performance.

For practical hardware acceleration, we propose $\mathbf{C}_4^n$-specific optimi zations combined with a layout transformation elimination strategy.

This approach mitigates inefficiencies arising from fine-grained pruning in line ar layers and improves performance across other operations.

Experimental results demonstrate that our method achieves superior task performa nce compared to other semi-structured pruning methods and achieves up-to 7$\time s$ speedup compared to llama.cpp framework on mobile devices.
****

Yanzheng Chen,Jun Yu

Classic but Everlasting: Traditional Gradient-Based Algorithms Converges Fast Ev en in Time-Varying Multi-Player Games

Last-iterate convergence behaviours of well-known algorithms are intensively inv estigated in various games, such as two-player bilinear zero-sum games.

However, most known last-iterate convergence properties rely on strict settings where the underlying games must have time-invariant payoffs.

Besides, the limited known attempts on the games with time-varying payoffs are i n two-player bilinear time-varying zero-sum games and strictly monotone games. B y contrast, in other time-varying games, the last-iterate behaviours of two clas sic algorithms, i.e., extra gradient (EG) and optimistic gradient (OG) algorithm s, still lack research, especially the convergence rates in multi-player games.

In this paper, we investigate the last-iterate behaviours of EG and OG algorithm s for convergent perturbed games, which extend upon the usual model of time-inva riant games and incorporate external factors, such as vanishing noises.

Using the recently proposed notion of the tangent residual (or its modifications) as the potential function of games and the measure of proximity to the Nash eq

uilibrium, we prove that the last-iterate convergence rates of EG and OG algorit
hms for perturbed games on bounded convex closed sets are $O({1}/{\sqrt{T}})$ if
 such games converge to monotone games at rates fast enough and that such a resu
lt holds true for certain unconstrained perturbed games. With this result, we ad
dress an open question
asking for the last-iterate convergence rate of EG and OG algorithms in constrai
ned and time-varying settings. The above convergence rates are similar to known
tight results on corresponding time-invariant games.
****

Peng Jin,Bo Zhu,Li Yuan,Shuicheng YAN
MoE++: Accelerating Mixture-of-Experts Methods with Zero-Computation Experts
In this work, we aim to simultaneously enhance the effectiveness and efficiency
of Mixture-of-Experts (MoE) methods. To achieve this, we propose MoE++, a genera
l and heterogeneous MoE framework that integrates both Feed-Forward Network (FFN
) and zero-computation experts. Specifically, we introduce three types of zero-c
omputation experts: the zero expert, copy expert, and constant expert, which cor
respond to discard, skip, and replace operations, respectively. This design offe
rs three key advantages: (i) **Low Computing Overhead**: Unlike the uniform mixi
ng mechanism for all tokens within vanilla MoE, MoE++ allows each token to engag
e with a dynamic number of FFNs, be adjusted by constant vectors, or even skip t
he MoE layer entirely. (ii) **High Performance**: By enabling simple tokens to u
tilize fewer FFN experts, MoE++ allows more experts to focus on challenging toke
ns, thereby unlocking greater performance potential than vanilla MoE. (iii) **De
ployment Friendly**: Given that zero-computation experts have negligible paramet
ers, we can deploy all zero-computation experts on each GPU, eliminating the sig
nificant communication overhead and expert load imbalance associated with FFN ex
perts distributed across different GPUs. Moreover, we leverage gating residuals,
 enabling each token to consider the pathway taken in the previous layer when se
lecting the appropriate experts. Extensive experimental results demonstrate that
 MoE++ achieves better performance while delivering 1.1$\sim$2.1$\times$ expert
forward throughput compared to a vanilla MoE model of the same size, which lays
a solid foundation for developing advanced and efficient MoE-related models.
****

Yexiang Liu,Jie Cao,Zekun Li,Ran He,Tieniu Tan
Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate
Large Language Models (LLMs) have seen significant progress but continue to stru
ggle with persistent reasoning mistakes.
Previous methods of *self-reflection* have been proven limited due to the models
' inherent fixed thinking patterns.
While Multi-Agent Debate (MAD) attempts to mitigate this by incorporating multip
le agents, it often employs the same reasoning methods, even though assigning di
fferent personas to models. This leads to a "fixed mental set", where models rel
y on homogeneous thought processes without exploring alternative perspectives.
In this paper, we introduce Diverse Multi-Agent Debate (DMAD), a method that enc
ourages agents to think with distinct reasoning approaches. By leveraging divers
e problem-solving strategies, each agent can gain insights from different perspe
ctives, refining its responses through discussion and collectively arriving at t
he optimal solution. DMAD effectively breaks the limitations of fixed mental set
s. We evaluate DMAD against various prompting techniques, including *self-reflec
tion* and traditional MAD, across multiple benchmarks using both LLMs and Multim
odal LLMs. Our experiments show that DMAD consistently outperforms other methods
, delivering better results than MAD in fewer rounds. Code is available at https
://github.com/MraDonkey/DMAD.
****

Ruiyi Fang,Bingheng Li,zhao kang,Qiuhao Zeng,Nima Hosseini Dashtbayaz,Ruizhi Pu,
Charles Ling,Boyu Wang
On the Benefits of Attribute-Driven Graph Domain Adaptation
Graph Domain Adaptation (GDA) addresses a pressing challenge in cross-network le
arning, particularly pertinent due to the absence of labeled data in real-world
graph datasets. Recent studies attempted to learn domain invariant representatio

ns by eliminating structural shifts between graphs. In this work, we show that e
xisting methodologies have overlooked the significance of the graph node attribu
te, a pivotal factor for graph domain alignment.
Specifically, we first reveal the impact of node attributes for GDA by theoretic
ally proving that in addition to the graph structural divergence between the dom
ains, the node attribute discrepancy also plays a critical role in GDA. Moreover
, we also empirically show that the attribute shift is more substantial than the
 topology shift, which further underscore the importance of node attribute align
ment in GDA. Inspired by this finding, a novel cross-channel module is developed
 to fuse and align both views between the source and target graphs for GDA. Expe
rimental results on a variety of benchmark verify the effectiveness of our metho
d.
****
Yuena Lin,Yiyuan Wang,Gengyu Lyu,Yongjian Deng,Haichun Cai,Huibin Lin,Haobo Wang
,Zhen Yang
Enhance Multi-View Classification Through Multi-Scale Alignment and Expanded Bou
ndary
Multi-view classification aims at unifying the data from multiple views to compl
ementarily enhance the classification performance. Unfortunately, two major prob
lems in multi-view data are damaging model performance. The first is feature het
erogeneity, which makes it hard to fuse features from different views. Consideri
ng this, we introduce a multi-scale alignment module, including an instance-scal
e alignment module and a prototype-scale alignment module to mine the commonalit
y from an inter-view perspective and an inter-class perspective respectively, jo
intly alleviating feature heterogeneity. The second is information redundancy wh
ich easily incurs ambiguous data to blur class boundaries and impair model gener
alization. Therefore, we propose a novel expanded boundary by extending the orig
inal class boundary with fuzzy set theory, which adaptively adjusts the boundary
 to fit ambiguous data. By integrating the expanded boundary into the prototype-
scale alignment module, our model further tightens the produced representations
and reduces boundary ambiguity. Additionally, compared with the original class b
oundary, the expanded boundary preserves more margins for classifying unseen dat
a, which guarantees the model generalization. Extensive experiment results acros
s various real-world datasets demonstrate the superiority of the proposed model
against existing state-of-the-art methods.
****
Georgiana Dinu,Corey D Barrett,Yi Xiang,Miguel Romero Calvo,Anna Currey,Xing Niu
Effective post-training embedding compression via temperature control in contras
tive training
Fixed-size learned representations (dense representations, or embeddings) are wi
dely used in many machine learning applications across language, vision or speec
h modalities. This paper investigates the role of the temperature parameter in c
ontrastive training for text embeddings. We shed light on the impact this parame
ter has on the intrinsic dimensionality of the embedding spaces obtained, and sh
ow that lower intrinsic dimensionality is further correlated with effective comp
ression of embeddings. We still observe a trade-off between absolute performance
 and effective compression and we propose temperature aggregation methods which
reduce embedding size by an order of magnitude with minimal impact on quality.
****
Xiaosen Zheng,Tianyu Pang,Chao Du,Qian Liu,Jing Jiang,Min Lin
Cheating Automatic LLM Benchmarks: Null Models Achieve High Win Rates
Automatic LLM benchmarks, such as AlpacaEval 2.0, Arena-Hard-Auto, and MT-Bench,
 have become popular for evaluating language models due to their cost-effectiven
ess and scalability compared to human evaluation. Achieving high win rates on th
ese benchmarks can significantly boost the promotional impact of newly released
language models. This promotional benefit may motivate tricks, such as manipulat
ing model output length or style to game win rates, even though several mechanis
ms have been developed to control length and disentangle style to reduce gameabi
lity. Nonetheless, we show that even a **"null model"** that always outputs a **
constant** response (*irrelevant to input instructions*) can cheat automatic ben

chmarks and achieve top-ranked win rates: an $86.5\\%$ LC win rate on AlpacaEval 2.0; an 83.0 score on Arena-Hard-Auto; and a 9.55 score on MT-Bench. Moreover, the crafted cheating outputs are **transferable** because we assume that the instructions of these benchmarks (e.g., 805 samples of AlpacaEval 2.0) are *private* and cannot be accessed. While our experiments are primarily proof-of-concept, an adversary could use LLMs to generate more imperceptible cheating responses, unethically benefiting from high win rates and promotional impact. Our findings call for the development of anti-cheating mechanisms for reliable automatic benchmarks. The code is available at https://github.com/sail-sg/Cheating-LLM-Benchmarks.
****

Yuxiao Hu,Qian Li,Dongxiao Zhang,Jinyue Yan,Yuntian Chen
Context-Alignment: Activating and Enhancing LLMs Capabilities in Time Series
Recently, leveraging pre-trained Large Language Models (LLMs) for time series (TS) tasks has gained increasing attention, which involves activating and enhancing LLMs' capabilities. Many methods aim to activate LLMs' capabilities based on token-level alignment, but overlook LLMs' inherent strength in natural language processing — their deep understanding of linguistic logic and structure rather than superficial embedding processing. We propose Context-Alignment (CA), a new paradigm that aligns TS with a linguistic component in the language environments familiar to LLMs to enable LLMs to contextualize and comprehend TS data, thereby activating their capabilities. Specifically, such context-level alignment comprises structural alignment and logical alignment, which is achieved by Dual-Scale Context-Alignment GNNs (DSCA-GNNs) applied to TS-language multimodal inputs. Structural alignment utilizes dual-scale nodes to describe hierarchical structure in TS-language, enabling LLMs to treat long TS data as a whole linguistic component while preserving intrinsic token features. Logical alignment uses directed edges to guide logical relationships, ensuring coherence in the contextual semantics. Following the DSCA-GNNs framework, we propose an instantiation method of CA, termed Few-Shot prompting Context-Alignment (FSCA), to enhance the capabilities of pre-trained LLMs in handling TS tasks. FSCA can be flexibly and repeatedly integrated into various layers of pre-trained LLMs to improve awareness of logic and structure, thereby enhancing performance. Extensive experiments show the effectiveness of FSCA and the importance of Context-Alignment across tasks, particularly in few-shot and zero-shot forecasting, confirming that Context-Alignment provides powerful prior knowledge on context. The code is open-sourced at https://github.com/tokaka22/ICLR25-FSCA.
****

Song Mei
U-Nets as Belief Propagation: Efficient Classification, Denoising, and Diffusion in Generative Hierarchical Models
U-Nets are among the most widely used architectures in computer vision, renowned for their exceptional performance in applications such as image segmentation, denoising, and diffusion modeling. However, a theoretical explanation of the U-Net architecture design has not yet been fully established.

This paper introduces a novel interpretation of the U-Net architecture by studying certain generative hierarchical models, which are tree-structured graphical models extensively utilized in both language and image domains. With their encoder-decoder structure, long skip connections, and pooling and up-sampling layers, we demonstrate how U-Nets can naturally implement the belief propagation denoising algorithm in such generative hierarchical models, thereby efficiently approximating the denoising functions. This leads to an efficient sample complexity bound for learning the denoising function using U-Nets within these models. Additionally, we discuss the broader implications of these findings for diffusion models in generative hierarchical models. We also demonstrate that the conventional architecture of convolutional neural networks (ConvNets) is ideally suited for classification tasks within these models. This offers a unified view of the roles of ConvNets and U-Nets, highlighting the versatility of generative hierarchical models in modeling complex data distributions.

****
Peng Dai,Feitong Tan,Qiangeng Xu,David Futschik,Ruofei Du,Sean Fanello,XIAOJUAN QI,Yinda Zhang
SVG: 3D Stereoscopic Video Generation via Denoising Frame Matrix
Video generation models have demonstrated great capability of producing impressive monocular videos, however, the generation of 3D stereoscopic video remains under-explored. We propose a pose-free and training-free approach for generating 3D stereoscopic videos using an off-the-shelf monocular video generation model. Our method warps a generated monocular video into camera views on stereoscopic baseline using estimated video depth, and employs a novel frame matrix video inpainting framework. The framework leverages the video generation model to inpaint frames observed from different timestamps and views. This effective approach generates consistent and semantically coherent stereoscopic videos without scene optimization or model fine-tuning. Moreover, we develop a disocclusion boundary re-injection scheme that further improves the quality of video inpainting by alleviating the negative effects propagated from disoccluded areas in the latent space. We validate the efficacy of our proposed method by conducting experiments on videos from various generative models, including Sora [4], Lumiere [2], WALT [8], and Zeroscope [12]. The experiments demonstrate that our method has a significant improvement over previous methods. Project page at https://daipengwa.github.io/SVG_ProjectPage/
****
Juntong Shi,Minkai Xu,Harper Hua,Hengrui Zhang,Stefano Ermon,Jure Leskovec
TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation
Synthesizing high-quality tabular data is an important topic in many data science tasks, ranging from dataset augmentation to privacy protection. However, developing expressive generative models for tabular data is challenging due to its inherent heterogeneous data types, complex inter-correlations, and intricate column-wise distributions. In this paper, we introduce TabDiff, a joint diffusion framework that models all mixed-type distributions of tabular data in one model. Our key innovation is the development of a joint continuous-time diffusion process for numerical and categorical data, where we propose feature-wise learnable diffusion processes to counter the high disparity of different feature distributions. TabDiff is parameterized by a transformer handling different input types, and the entire framework can be efficiently optimized in an end-to-end fashion. We further introduce a mixed-type stochastic sampler to automatically correct the accumulated decoding error during sampling, and propose classifier-free guidance for conditional missing column value imputation. Comprehensive experiments on seven datasets demonstrate that TabDiff achieves superior average performance over existing competitive baselines across all eight metrics, with up to 22.5% improvement over the state-of-the-art model on pair-wise column correlation estimations. Code is available at https://github.com/MinkaiXu/TabDiff.
****
Zonglin Di,Sixie Yu,Yevgeniy Vorobeychik,Yang Liu
Adversarial Machine Unlearning
This paper focuses on the challenge of machine unlearning, aiming to remove the influence of specific training data on machine learning models. Traditionally, the development of unlearning algorithms runs parallel with that of membership inference attacks (MIA), a type of privacy threat to determine whether a data instance was used for training. However, the two strands are intimately connected: one can view machine unlearning through the lens of MIA success with respect to removed data. Recognizing this connection, we propose a game-theoretic framework that integrates MIAs into the design of unlearning algorithms. Specifically, we model the unlearning problem as a Stackelberg game in which an unlearner strives to unlearn specific training data from a model, while an auditor employs MIAs to detect the traces of the ostensibly removed data. Adopting this adversarial perspective allows the utilization of new attack advancements, facilitating the design of unlearning algorithms. Our framework stands out in two ways. First, it takes an adversarial approach and proactively incorporates the attacks into the design of unlearning algorithms. Secondly, it uses implicit differentiation to o

btain the gradients that limit the attacker's success, thus benefiting the proce
ss of unlearning. We present empirical results to demonstrate the effectiveness
of the proposed approach for machine unlearning.
****

Yulai Zhao,Masatoshi Uehara,Gabriele Scalia,Sunyuan Kung,Tommaso Biancalani,Serg
ey Levine,Ehsan Hajiramezanali
Adding Conditional Control to Diffusion Models with Reinforcement Learning
Diffusion models are powerful generative models that allow for precise control o
ver the characteristics of the generated samples. While these diffusion models t
rained on large datasets have achieved success, there is often a need to introdu
ce additional controls in downstream fine-tuning processes, treating these power
ful models as pre-trained diffusion models. This work presents a novel method ba
sed on reinforcement learning (RL) to add such controls using an offline dataset
 comprising inputs and labels. We formulate this task as an RL problem, with the
 classifier learned from the offline dataset and the KL divergence against pre-t
rained models serving as the reward functions. Our method, **CTRL** (**C**onditi
oning pre-**T**rained diffusion models with **R**einforcement **L**earning), pro
duces soft-optimal policies that maximize the abovementioned reward functions. W
e formally demonstrate that our method enables sampling from the conditional dis
tribution with additional controls during inference.
Our RL-based approach offers several advantages over existing methods. Compared
to classifier-free guidance,
it improves sample efficiency and can greatly simplify dataset construction by l
everaging conditional independence between the inputs and additional controls. A
dditionally, unlike classifier guidance, it eliminates the need to train classif
iers from intermediate states to additional controls.
The code is available at https://github.com/zhaoyl18/CTRL.
****

Ruizhong Qiu,Weiliang Will Zeng,James Ezick,Christopher Lott,Hanghang Tong
How efficient is LLM-generated code? A rigorous & high-standard benchmark
The emergence of large language models (LLMs) has significantly pushed the front
iers of program synthesis. Advancement of LLM-based program synthesis calls for
a thorough evaluation of LLM-generated code. Most evaluation frameworks focus on
 the (functional) correctness of generated code; efficiency, as an important mea
sure of code quality, has been overlooked in existing evaluations. In this work,
 we develop ENAMEL (EfficeNcy AutoMatic EvaLuator), a rigorous and high-standard
 benchmark for evaluating the capability of LLMs in generating efficient code. F
irstly, we propose a new efficiency metric called eff@k, which generalizes the p
ass@k metric from correctness to efficiency and appropriately handles right-cens
ored execution time. Furthermore, we derive an unbiased and variance-reduced est
imator of eff@k via Rao–Blackwellization; we also provide a numerically stable i
mplementation for the new estimator. Secondly, to set a high-standard for effici
ency evaluation, we employ a human expert to design best algorithms and implemen
tations as our reference solutions of efficiency, many of which are much more ef
ficient than existing canonical solutions in HumanEval and HumanEval+. Moreover,
 to ensure a rigorous evaluation, we employ a human expert to curate strong test
 case generators to filter out wrong code and differentiate suboptimal algorithm
s. An extensive study across 30 popular LLMs using our benchmark ENAMEL shows th
at LLMs still fall short of generating expert-level efficient code. Using two su
bsets of our problem set, we demonstrate that such deficiency is because current
 LLMs struggle in designing advanced algorithms and are barely aware of implemen
tation optimization.
****

Eric Mazumdar,Kishan Panaganti,Laixi Shi
Tractable Multi-Agent Reinforcement Learning through Behavioral Economics
A significant roadblock to the development of principled multi-agent reinforceme
nt learning (MARL) algorithms is the fact that desired solution concepts like Na
sh equilibria may be intractable to compute. We show how one can overcome this o
bstacle by introducing concepts from behavioral economics into MARL. To do so, w
e imbue agents with two key features of human decision-making: risk aversion and

bounded rationality. We show that introducing these two properties into games g
ives rise to a class of equilibria---risk-averse quantal response equilibria (RQ
E)---which are tractable to compute in \emph{all} n-player matrix and finite-h
orizon Markov games. In particular, we show that they emerge as the endpoint of
no-regret learning in suitably adjusted versions of the games. Crucially, the c
lass of computationally tractable RQE is independent of the underlying game stru
cture and only depends on agents' degrees of risk-aversion and bounded rationali
ty. To validate the expressivity of this class of solution concepts we show tha
t it captures peoples' patterns of play in a number of 2-player matrix games pre
viously studied in experimental economics. Furthermore, we give a first analysis
of the sample complexity of computing these equilibria in finite-horizon Markov
games when one has access to a generative model. We validate our findings on a
simple multi-agent reinforcement learning benchmark. Our results open the doors
for to the principled development of new decentralized multi-agent reinforcement
learning algorithms.
****

Jing He,Haodong LI,Wei Yin,Yixun Liang,Leheng Li,Kaiqiang Zhou,Hongbo Zhang,Bing
bing Liu,Ying-Cong Chen

Lotus: Diffusion-based Visual Foundation Model for High-quality Dense Prediction

Leveraging the visual priors of pre-trained text-to-image diffusion models offer
s a promising solution to enhance zero-shot generalization in dense prediction t
asks. However, existing methods often uncritically use the original diffusion fo
rmulation, which may not be optimal due to the fundamental differences between d
ense prediction and image generation. In this paper, we provide a systemic analy
sis of the diffusion formulation for the dense prediction, focusing on both qual
ity and efficiency. And we find that the original parameterization type for imag
e generation, which learns to predict noise, is harmful for dense prediction; th
e multi-step noising/denoising diffusion process is also unnecessary and challen
ging to optimize. Based on these insights, we introduce $\textbf{Lotus}$, a diff
usion-based visual foundation model with a simple yet effective adaptation proto
col for dense prediction. Specifically, Lotus is trained to directly predict ann
otations instead of noise, thereby avoiding harmful variance. We also reformulat
e the diffusion process into a single-step procedure, simplifying optimization a
nd significantly boosting inference speed. Additionally, we introduce a novel tu
ning strategy called detail preserver, which achieves more accurate and fine-gra
ined predictions. Without scaling up the training data or model capacity, Lotus
achieves SoTA performance in zero-shot depth and normal estimation across variou
s datasets. It also enhances efficiency, being significantly faster than most ex
isting diffusion-based methods. Lotus' superior quality and efficiency also enab
le a wide range of practical applications, such as joint estimation, single/mult
i-view 3D reconstruction, etc.
****

Sanjiban Choudhury,Paloma Sodhi

Better than Your Teacher: LLM Agents that learn from Privileged AI Feedback

While large language models (LLMs) show impressive decision-making abilities, cu
rrent methods lack a mechanism for automatic self-improvement from errors during
task execution. We propose LEAP, an iterative fine-tuning framework that contin
ually improves LLM agents using feedback from AI expert teachers. Our key insigh
t is to equip the expert teachers with a privileged state -- information availab
le during training but hidden at test time. This allows even weak experts to pro
vide precise guidance, significantly improving the student agent's performance w
ithout access to privileged information at test time.

We evaluate LEAP on multiple decision-making benchmarks, including text-based ga
mes (ALFWorld), web navigation (WebShop), and interactive coding (Intercode Bash
). Our experiments show that LEAP (1) outperforms behavior cloning and ReAct bas
elines (2) enables weak student models (e.g., Llama3-8B) to exceed performance o
f strong teacher models (GPT-4o), and (3) allows weak models to self-improve usi
ng privileged versions of themselves.

We provide a theoretical analysis showing that LEAP's success hinges on balancin
g privileged information with student’s realizability, which we empirically vali

date. Our code is available at \url{https://leap-llm.github.io}.
****
Zhiyuan Weng,Guikun Chen,Wenguan Wang
Do as We Do, Not as You Think: the Conformity of Large Language Models
Recent advancements in large language models (LLMs) revolutionize the field of intelligent agents, enabling collaborative multi-agent systems capable of tackling complex problems across various domains. However, the potential of conformity within these systems, analogous to phenomena like conformity bias and group-think in human group dynamics, remains largely unexplored, raising concerns about their collective problem-solving capabilities and possible ethical implications. This paper presents a comprehensive study on conformity in LLM-driven multi-agent systems, focusing on three aspects: the existence of conformity, the factors influencing conformity, and potential mitigation strategies. In particular, we introduce BenchForm, a new conformity-oriented benchmark, featuring reasoning-intensive tasks and five distinct interaction protocols designed to probe LLMs' behavior in collaborative scenarios. Several representative LLMs are evaluated on BenchForm, using metrics such as conformity rate and independence rate to quantify conformity's impact. Our analysis delves into factors influencing conformity, including interaction time and majority size, and examines how the subject agent rationalize its conforming behavior. Furthermore, we explore two strategies to mitigate conformity effects, i.e., developing enhanced persona and implementing a reflection mechanism. Several interesting findings regarding LLMs' conformity are derived from empirical results and case studies. We hope that these insights can pave the way for more robust and ethically-aligned collaborative AI systems. Our benchmark and code are available at BenchForm.
****
Giseung Park,Youngchul Sung
Reward Dimension Reduction for Scalable Multi-Objective Reinforcement Learning
In this paper, we introduce a simple yet effective reward dimension reduction method to tackle the scalability challenges of multi-objective reinforcement learning algorithms. While most existing approaches focus on optimizing two to four objectives, their abilities to scale to environments with more objectives remain uncertain. Our method uses a dimension reduction approach to enhance learning efficiency and policy performance in multi-objective settings. While most traditional dimension reduction methods are designed for static datasets, our approach is tailored for online learning and preserves Pareto-optimality after transformation. We propose a new training and evaluation framework for reward dimension reduction in multi-objective reinforcement learning and demonstrate the superiority of our method in environments including one with sixteen objectives, significantly outperforming existing online dimension reduction methods.
****
Shuang Liu,Yihan Wang,Yifan Zhu,Yibo Miao,Xiao-Shan Gao
Provable Robust Overfitting Mitigation in Wasserstein Distributionally Robust Optimization
Wasserstein distributionally robust optimization (WDRO) optimizes against worst-case distributional shifts within a specified uncertainty set, leading to enhanced generalization on unseen adversarial examples, compared to standard adversarial training which focuses on pointwise adversarial perturbations. However, WDRO still suffers fundamentally from the robust overfitting problem, as it does not consider statistical error. We address this gap by proposing a novel robust optimization framework under a new uncertainty set for adversarial noise via Wasserstein distance and statistical error via Kullback-Leibler divergence, called the Statistically Robust WDRO. We establish a robust generalization bound for the new optimization framework, implying that out-of-distribution adversarial performance is at least as good as the statistically robust training loss with high probability. Furthermore, we derive conditions under which Stackelberg and Nash equilibria exist between the learner and the adversary, giving an optimal robust model in certain sense.Finally, through extensive experiments, we demonstrate that our method significantly mitigates robust overfitting and enhances robustness within the framework of WDRO.

****
Caleb Dahlke,Jason Pacheco
Flow-based Variational Mutual Information: Fast and Flexible Approximations
Mutual Information (MI) is a fundamental measure of dependence between random variables, but its practical application is limited because it is difficult to calculate in many circumstances. Variational methods offer one approach by introducing an approximate distribution to create various bounds on MI, which in turn is an easier optimization problem to solve. In practice, the variational distribution chosen is often a Gaussian, which is convenient but lacks flexibility in modeling complicated distributions. In this paper, we introduce new classes of variational estimators based on Normalizing Flows that extend the previous Gaussian-based variational estimators. Our new estimators maintain many of the same theoretical guarantees while simultaneously enhancing the expressivity of the variational distribution. We experimentally verify that our new methods are effective on large MI problems where discriminative-based estimators, such as MINE and InfoNCE, are fundamentally limited. Furthermore, we compare against a diverse set of benchmarking tests to show that the flow-based estimators often perform as well, if not better, than the discriminative-based counterparts. Finally, we demonstrate how these estimators can be effectively utilized in the Bayesian Optimal Experimental Design setting for online sequential decision making.
****
Jerry Weihong Liu,Jessica Grogan,Owen M Dugan,Ashish Rao,Simran Arora,Atri Rudra,Christopher Re
Towards Learning High-Precision Least Squares Algorithms with Sequence Models
This paper investigates whether sequence models can learn to perform numerical algorithms, e.g. gradient descent, on the fundamental problem of least squares. Our goal is to inherit two properties of standard algorithms from numerical analysis: (1) machine precision, i.e. we want to obtain solutions that are accurate to near floating point error, and (2) numerical generality, i.e. we want them to apply broadly across problem instances. We find that prior approaches using Transformers fail to meet these criteria, and identify limitations present in existing architectures and training procedures. First, we show that softmax Transformers struggle to perform high-precision multiplications, which prevents them from precisely learning numerical algorithms. Second, we identify an alternate class of architectures, comprised entirely of polynomials, that can efficiently represent high-precision gradient descent iterates. Finally, we investigate precision bottlenecks during training and address them via a high-precision training recipe that reduces stochastic gradient noise. Our recipe enables us to train two polynomial architectures, gated convolutions and linear attention, to perform gradient descent iterates on least squares problems. For the first time, we demonstrate the ability to train to near machine precision. Applied iteratively, our models obtain $100,000\times$ lower MSE than standard Transformers trained end-to-end and they incur a $10,000\times$ smaller generalization gap on out-of-distribution problems. We make progress towards end-to-end learning of numerical algorithms for least squares.
****
Genta Indra Winata,David Anugraha,Lucky Susanto,Garry Kuwanto,Derry Tanti Wijaya
MetaMetrics: Calibrating Metrics for Generation Tasks Using Human Preferences
Understanding the quality of a performance evaluation metric is crucial for ensuring that model outputs align with human preferences. However, it remains unclear how well each metric captures the diverse aspects of these preferences, as metrics often excel in one particular area but not across all dimensions. To address this, it is essential to systematically calibrate metrics to specific aspects of human preference, catering to the unique characteristics of each aspect. We introduce MetaMetrics, a calibrated meta-metric designed to evaluate generation tasks across different modalities in a supervised manner. MetaMetrics optimizes the combination of existing metrics to enhance their alignment with human preferences. Our metric demonstrates flexibility and effectiveness in both language and vision downstream tasks, showing significant benefits across various multilingual and multi-domain scenarios. MetaMetrics aligns closely with human preferences

and is highly extendable and easily integrable into any application. This makes MetaMetrics a powerful tool for improving the evaluation of generation tasks, ensuring that metrics are more representative of human judgment across diverse contexts.
****
Haitz Sáez de Ocáriz Borde,Anastasis Kratsios,Marc T. Law,Xiaowen Dong,Michael M. Bronstein
Neural Spacetimes for DAG Representation Learning
We propose a class of trainable deep learning-based geometries called Neural SpaceTimes (NSTs), which can universally represent nodes in weighted Directed Acyclic Graphs (DAGs) as events in a spacetime manifold. While most works in the literature focus on undirected graph representation learning or causality embedding separately, our differentiable geometry can encode both graph edge weights in its spatial dimensions and causality in the form of edge directionality in its temporal dimensions. We use a product manifold that combines a quasi-metric (for space) and a partial order (for time). NSTs are implemented as three neural networks trained in an end-to-end manner: an embedding network, which learns to optimize the location of nodes as events in the spacetime manifold, and two other networks that optimize the space and time geometries in parallel, which we call a neural (quasi-)metric and a neural partial order, respectively. The latter two networks leverage recent ideas at the intersection of fractal geometry and deep learning to shape the geometry of the representation space in a data-driven fashion, unlike other works in the literature that use fixed spacetime manifolds such as Minkowski space or De Sitter space to embed DAGs. Our main theoretical guarantee is a universal embedding theorem, showing that any k-point DAG can be embedded into an NST with $1+\mathcal{O}(\log(k))$ distortion while exactly preserving its causal structure. The total number of parameters defining the NST is sub-cubic in k and linear in the width of the DAG. If the DAG has a planar Hasse diagram, this is improved to $\mathcal{O}(\log(k) + 2)$ spatial and 2 temporal dimensions. We validate our framework computationally with synthetic weighted DAGs and real-world network embeddings; in both cases, the NSTs achieve lower embedding distortions than their counterparts using fixed spacetime geometries.
****
Tong Wu,Shujian Zhang,Kaiqiang Song,Silei Xu,Sanqiang Zhao,Ravi Agrawal,Sathish Reddy Indurthi,Chong Xiang,Prateek Mittal,Wenxuan Zhou
Instructional Segment Embedding: Improving LLM Safety with Instruction Hierarchy
Large Language Models (LLMs) are susceptible to security and safety threats, such as prompt injection, prompt extraction, and harmful requests.
One major cause of these vulnerabilities is the lack of an instruction hierarchy.
Modern LLM architectures treat all inputs equally, failing to distinguish between and prioritize various types of instructions, such as system messages, user prompts, and data.
As a result, lower-priority user prompts may override more critical system instructions, including safety protocols.
Existing approaches to achieving instruction hierarchy, such as delimiters and instruction-based training, do not address this issue at the architectural level.
We introduce the $\textbf{I}$nstructional $\textbf{S}$egment $\textbf{E}$mbedding (ISE) technique, inspired by BERT, to modern large language models, which embeds instruction priority information directly into the model.
This approach enables models to explicitly differentiate and prioritize various instruction types, significantly improving safety against malicious prompts that attempt to override priority rules.
Our experiments on the Structured Query and Instruction Hierarchy benchmarks demonstrate an average robust accuracy increase of up to 15.75\% and 18.68\%, respectively.
Furthermore, we observe an improvement in the instruction-following capability of up to 4.1\% on AlpacaEval.
Overall, our approach offers a promising direction for enhancing the safety and effectiveness of LLM architectures.

****
Hao Yuan,Wenli Ouyang,Changwen Zhang,Yong Sun,Liming Gong,Junchi Yan
BTBS-LNS: Binarized-Tightening, Branch and Search on Learning LNS Policies for M
IP
Learning to solve large-scale Mixed Integer Program (MIP) problems is an emergin
g research topic, and policy learning-based Large Neighborhood Search (LNS) has
been a popular paradigm. However, the explored space of LNS policy is often limi
ted even in the training phase, making the learned policy sometimes wrongly fix
some potentially important variables early in the search, leading to local optim
um in some cases. Moreover, many methods only assume binary variables to deal wi
th. We present a practical approach, termed Binarized-Tightening Branch-and-Sear
ch for Large Neighborhood Search (BTBS-LNS). It comprises three key techniques:
1) the ``Binarized Tightening" technique for integer variables to handle their w
ide range by binary encoding and bound tightening; 2) an attention-based tripart
ite graph to capture global correlations among variables and constraints for an
MIP instance; 3) an extra branching network as a global view, to identify and op
timize wrongly-fixed backdoor variables at each search step. Experiments show it
s superior performance over the open-source solver SCIP and LNS baselines. Moreo
ver, it performs competitively with, and sometimes better than the commercial so
lver Gurobi (v9.5.0), especially on the MIPLIB2017 benchmark chosen by Hans Mitt
elmann, where our method can deliver 10\% better primal gaps compared with Gurob
i in a 300s cut-off time.
****
Huiang He,Minghui Hu,Chuanxia Zheng,Chaoyue Wang,Tat-Jen Cham
Semantix: An Energy-guided Sampler for Semantic Style Transfer
Recent advances in style and appearance transfer are impressive, but most method
s isolate global style and local appearance transfer, neglecting semantic corres
pondence. Additionally, image and video tasks are typically handled in isolation
, with little focus on integrating them for video transfer. To address these lim
itations, we introduce a novel task, *Semantic Style Transfer*, which involves t
ransferring style and appearance features from a reference image to a target vis
ual content based on semantic correspondence. We subsequently propose a training
-free method, *Semantix*, an energy-guided sampler designed for Semantic Style T
ransfer that simultaneously guides both style and appearance transfer based on s
emantic understanding capacity of pre-trained diffusion models. Additionally, as
 a sampler, *Semantix* can be seamlessly applied to both image and video models,
 enabling semantic style transfer to be generic across various visual media. Spe
cifically, once inverting both reference and context images or videos to noise s
pace by SDEs, *Semantix* utilizes a meticulously crafted energy function to guid
e the sampling process, including three key components: *Style Feature Guidance*
, *Spatial Feature Guidance* and *Semantic Distance* as a regularisation term. E
xperimental results demonstrate that *Semantix* not only effectively accomplishe
s the task of semantic style transfer across images and videos, but also surpass
es existing state-of-the-art solutions in both fields.
****
Yiming Liu,Kezhao Liu,Yao Xiao,ZiYi Dong,Xiaogang Xu,Pengxu Wei,Liang Lin
Towards Understanding the Robustness of Diffusion-Based Purification: A Stochast
ic Perspective
Diffusion-Based Purification (DBP) has emerged as an effective defense mechanism
 against adversarial attacks. The success of DBP is often attributed to the forw
ard diffusion process, which reduces the distribution gap between clean and adve
rsarial images by adding Gaussian noise. Although this explanation is theoretica
lly grounded, the precise contribution of this process to robustness remains unc
lear. In this paper, through a systematic investigation, we propose that the int
rinsic stochasticity in the DBP procedure is the primary factor driving robustne
ss. To explore this hypothesis, we introduce a novel Deterministic White-Box (DW
-box) evaluation protocol to assess robustness in the absence of stochasticity,
and analyze attack trajectories and loss landscapes. Our results suggest that DB
P models primarily leverage stochasticity to evade effective attack directions,
and that their ability to purify adversarial perturbations can be weak. To furth

er enhance the robustness of DBP models, we propose Adversarial Denoising Diffus
ion Training (ADDT), which incorporates classifier-guided adversarial perturbati
ons into diffusion training, thereby strengthening the models' ability to purify
 adversarial perturbations. Additionally, we propose Rank-Based Gaussian Mapping
 (RBGM) to improve the compatibility of perturbations with diffusion models. Exp
erimental results validate the effectiveness of ADDT. In conclusion, our study s
uggests that future research on DBP can benefit from the perspective of decoupli
ng stochasticity-based and purification-based robustness.
****

Yingyu Liang,Jiangxuan Long,Zhenmei Shi,Zhao Song,Yufa Zhou
Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix
Large Language Models (LLMs) have shown immense potential in enhancing various a
spects of our daily lives, from conversational AI to search and AI assistants. H
owever, their growing capabilities come at the cost of extremely large model siz
es, making deployment on edge devices challenging due to memory and computationa
l constraints. This paper introduces a novel approach to LLM weight pruning that
 directly optimizes for approximating the attention matrix, a core component of
transformer architectures. Unlike existing methods that focus on linear approxim
ations, our approach accounts for the non-linear nature of the Softmax attention
 mechanism. We provide theoretical guarantees for the convergence of our Gradien
t Descent-based optimization method to a near-optimal pruning mask solution. Our
 empirical results demonstrate the effectiveness of our non-linear pruning appro
ach in maintaining model performance while significantly reducing computational
costs, which is beyond the current state-of-the-art methods, i.e., SparseGPT and
 Wanda, by a large margin. This work establishes a new theoretical foundation fo
r pruning algorithm design in LLMs, potentially paving the way for more efficien
t LLM inference on resource-constrained devices.
****

Michael Tschannen,André Susano Pinto,Alexander Kolesnikov
JetFormer: An autoregressive generative model of raw images and text
Removing modeling constraints and unifying architectures across domains has been
 a key driver of the recent progress in training large multimodal models. Howeve
r, most of these models still rely on many separately trained components such as
 modality-specific encoders and decoders. In this work, we further streamline jo
int generative modeling of images and text. We propose an autoregressive decoder
-only transformer---JetFormer---which is trained to directly maximize the likeli
hood of raw data, without relying on any separately pretrained components, and c
an understand and generate both text and images. Specifically, we leverage a nor
malizing flow model to obtain a soft-token image representation that is jointly
trained with an autoregressive multimodal transformer. The normalizing flow mode
l serves as both an image encoder for perception tasks and an image decoder for
image generation tasks during inference. JetFormer achieves text-to-image genera
tion quality competitive with recent VQVAE- and VAE-based baselines. These basel
ines rely on pretrained image autoencoders, which are trained with a complex mix
ture of losses, including perceptual ones. At the same time, JetFormer demonstra
tes robust image understanding capabilities. To the best of our knowledge, JetFo
rmer is the first model that is capable of generating high-fidelity images and p
roducing strong log-likelihood bounds.
****

Shihao Shao,Haoran Geng,Zun Wang,Qinghua Cui
FreeCG: Free the Design Space of Clebsch-Gordan Transform for Machine Learning F
orce Fields
Machine Learning Force Fields (MLFFs) are of great importance for chemistry, phy
sics, materials science, and many other related fields. The Clebsch–Gordan trans
form (CG transform) effectively encodes many-body interactions and is thus an im
portant building block for many models of MLFFs. However, the permutation-equiva
riance requirement of MLFFs limits the design space of CG transform, that is, in
tensive CG transform has to be conducted for each neighboring edge and the opera
tions should be performed in the same manner for all edges. Freeing up the desig
n space can greatly improve the model's expressiveness while simultaneously decr

easing computational demands. To reach this goal, we utilize a mathematical proposition, invariance transitivity, to show that implementing the CG transform layer on the permutation-invariant abstract edges allows complete freedom in the design of the layer without compromising the overall permutation equivariance. Developing on this free design space, we further propose group CG transform with sparse path, abstract edges shuffling, and attention enhancer to form a powerful and efficient CG transform layer. Our method, known as FreeCG, achieves state-of-the-art (SOTA) results in force prediction for MD17, rMD17, MD22, and is well extended to property prediction in QM9 datasets with several improvements greater than 15% and the maximum beyond 20%. The extensive real-world applications showcase high practicality. FreeCG introduces a novel paradigm for carrying out efficient and expressive CG transform in future geometric network designs. To demonstrate this, the recent SOTA, QuinNet, is also enhanced under our paradigm. Code: https://github.com/ShihaoShao-GH/FreeCG.
****
Kun-Peng Ning,Shuo Yang,Yuyang Liu,Jia-Yu Yao,Zhenhui Liu,Yonghong Tian,Yibing Song,Li Yuan
PiCO: Peer Review in LLMs based on Consistency Optimization
Existing large language models (LLMs) evaluation methods typically focus on testing the performance on some closed-environment and domain-specific benchmarks with human annotations. In this paper, we explore a novel unsupervised evaluation direction, utilizing peer-review mechanisms to measure LLMs automatically without any human feedback. In this setting, both open-source and closed-source LLMs lie in the same environment, capable of answering unlabeled questions and evaluating each other, where each LLM’s response score is jointly determined by other anonymous ones. During this process, we found that those answers that are more recognized by other ``reviewers'' (models) usually come from LLMs with stronger abilities, while these models can also evaluate others' answers more accurately. We formalize it as a consistency assumption, i.e., the ability and score of the model usually have consistency. We exploit this to optimize each model's confidence, thereby re-ranking the LLMs to be closer to human rankings. We perform experiments on multiple datasets with standard rank-based metrics, validating the effectiveness of the proposed approach.
****
Toni J.B. Liu,Nicolas Boulle,Raphaël Sarfati,Christopher Earls
Density estimation with LLMs: a geometric investigation of in-context learning trajectories
Large language models (LLMs) demonstrate remarkable emergent abilities to perform in-context learning across various tasks, including time series forecasting. This work investigates LLMs' ability to estimate probability density functions (PDFs) from data observed in-context;
such density estimation (DE) is a fundamental task underlying many probabilistic modeling problems.
We leverage the Intensive Principal Component Analysis (InPCA) to visualize and analyze the in-context learning dynamics of LLaMA-2 models.
Our main finding is that these LLMs all follow similar learning trajectories in a low-dimensional InPCA space, which are distinct from those of traditional density estimation methods like histograms and Gaussian kernel density estimation (KDE).
We interpret the LLaMA in-context DE process as a KDE with an adaptive kernel width and shape.
This custom kernel model captures a significant portion of LLaMA's behavior despite having only two parameters.
We further speculate on why LLaMA's kernel width and shape differs from classical algorithms, providing insights into the mechanism of in-context probabilistic reasoning in LLMs.
Our codebase, along with a 3D visualization of an LLM's in-context learning trajectory, is publicly available at https://github.com/AntonioLiu97/LLMICL_inPCA.
****
Ilya Loshchilov,Cheng-Ping Hsieh,Simeng Sun,Boris Ginsburg

nGPT: Normalized Transformer with Representation Learning on the Hypersphere
We propose a novel neural network architecture, the normalized Transformer (nGPT) with representation learning on the hypersphere. In nGPT, all vectors forming the embeddings, MLP, attention matrices and hidden states are unit norm normalized. The input stream of tokens travels on the surface of a hypersphere, with each layer contributing a displacement towards the target output predictions. These displacements are defined by the MLP and attention blocks, whose vector components also reside on the same hypersphere. Experiments show that nGPT learns much faster, reducing the number of training steps required to achieve the same accuracy by a factor of 4 to 20, depending on the sequence length.
****
Hualin Zhang,Haozhen Zhang,Zhekai Liu,Bin Gu,Yi Chang
Collaborative Discrete-Continuous Black-Box Prompt Learning for Language Models
Large Scale Pre-Trained Language Models (PTMs) have demonstrated unprecedented capabilities across diverse natural language processing tasks.
Adapting such models to downstream tasks is computationally intensive and time-consuming, particularly in black-box scenarios common in Language-Model-as-a-Service (LMaaS) environments, where model parameters and gradients are inaccessible. Recently, black-box prompt learning using zeroth-order gradients has emerged as a promising approach to address these challenges by optimizing learnable continuous prompts in embedding spaces, starting with \textit{randomly initialized discrete text prompts}. However, its reliance on randomly initialized discrete prompts limits adaptability to diverse downstream tasks or models. To address this limitation,
this paper introduces ZO-PoG, a novel framework that optimizes prompts through a collaborative approach, combining Policy Gradient optimization for initial discrete text prompts and Zeroth-Order optimization for continuous prompts in embedding space. By optimizing collaboratively between discrete and continuous prompts, ZO-PoG maximizes adaptability to downstream tasks, achieving superior results without direct access to the model’s internal structures.
Importantly, we establish the sub-linear convergence of ZO-PoG under mild assumptions.
The experiments on different datasets demonstrate significant improvements in various tasks compared to the baselines.
****
Shentong Mo,Xufang Luo,Dongsheng Li
pMoE: Prompting Diverse Experts Together Wins More in Visual Adaptation
Parameter-efficient fine-tuning has demonstrated promising results across various visual adaptation tasks, such as classification and segmentation. Typically, prompt tuning techniques have harnessed knowledge from a single pre-trained model, whether from a general or a specialized medical domain. However, this approach typically overlooks the potential synergies that could arise from integrating diverse domain knowledge within the same tuning process. In this work, we propose a novel Mixture-of-Experts prompt tuning method called pMoE, which leverages the strengths of multiple expert domains through expert-specialized prompt tokens and the learnable dispatcher, effectively combining their expertise in a unified model framework. Our pMoE introduces expert-specific prompt tokens and utilizes a dynamic token dispatching mechanism at various prompt layers to optimize the contribution of each domain expert during the adaptation phase. By incorporating both domain knowledge from diverse experts, the proposed pMoE significantly enhances the model's versatility and applicability to a broad spectrum of tasks. We conduct extensive experiments across 47 adaptation tasks, including both classification and segmentation in general and medical domains. The results demonstrate that our pMoE not only achieves superior performance with a large margin of improvements but also offers an optimal trade-off between computational efficiency and adaptation effectiveness compared to existing methods.
****
Matan Rusanovsky,Or Hirschorn,Shai Avidan
CapeX: Category-Agnostic Pose Estimation from Textual Point Explanation
Conventional 2D pose estimation models are constrained by their design to specif

ic object categories. This limits their applicability to predefined objects. To overcome these limitations, category-agnostic pose estimation (CAPE) emerged as a solution. CAPE aims to facilitate keypoint localization for diverse object categories using a unified model, which can generalize from minimal annotated support images.
Recent CAPE works have produced object poses based on arbitrary keypoint definitions annotated on a user-provided support image. Our work departs from conventional CAPE methods, which require a support image, by adopting a text-based approach instead of the support image.
Specifically, we use a pose-graph, where nodes represent keypoints that are described with text. This representation takes advantage of the abstraction of text descriptions and the structure imposed by the graph.
Our approach effectively breaks symmetry, preserves structure, and improves occlusion handling.
We validate our novel approach using the MP-100 benchmark, a comprehensive dataset covering over 100 categories and 18,000 images. MP-100 is structured so that the evaluation categories are unseen during training, making it especially suited for CAPE. Under a 1-shot setting, our solution achieves a notable performance boost of 1.26\%, establishing a new state-of-the-art for CAPE. Additionally, we enhance the dataset by providing text description annotations for both training and testing. We also include alternative text annotations specifically for testing the model's ability to generalize across different textual descriptions, further increasing its value for future research. Our code and dataset are publicly available at https://github.com/matanr/capex.
****
Jacob Bamberger,Federico Barbero,Xiaowen Dong,Michael M. Bronstein
Bundle Neural Network for message diffusion on graphs
The dominant paradigm for learning on graphs is message passing. Despite being a strong inductive bias, the local message passing mechanism faces challenges such as over-smoothing, over-squashing, and limited expressivity. To address these issues, we introduce Bundle Neural Networks (BuNNs), a novel graph neural network architecture that operates via *message diffusion* on *flat vector bundles* — geometrically inspired structures that assign to each node a vector space and an orthogonal map. A BuNN layer evolves node features through a diffusion-type partial differential equation, where its discrete form acts as a special case of the recently introduced Sheaf Neural Network (SNN), effectively alleviating over-smoothing. The continuous nature of message diffusion enables BuNNs to operate at larger scales, reducing over-squashing. We establish the universality of BuNNs in approximating feature transformations on infinite families of graphs with injective positional encodings, marking the first positive expressivity result of its kind. We support our claims with formal analysis and synthetic experiments. Empirically, BuNNs perform strongly on heterophilic and long-range tasks, which demonstrates their robustness on a diverse range of challenging real-world tasks.
****
Sayan Banerjee,Krishna Balasubramanian,PROMIT GHOSAL
Improved Finite-Particle Convergence Rates for Stein Variational Gradient Descent
We provide finite-particle convergence rates for the Stein Variational Gradient Descent (SVGD) algorithm in the Kernelized Stein Discrepancy ($\KSD$) and Wasserstein-2 metrics. Our key insight is that the time derivative of the relative entropy between the joint density of N particle locations and the N-fold product target measure, starting from a regular initial distribution, splits into a dominant 'negative part' proportional to N times the expected $\KSD^2$ and a smaller 'positive part'. This observation leads to $\KSD$ rates of order $1/\sqrt{N}$, in both continuous and discrete time, providing a near optimal (in the sense of matching the corresponding i.i.d. rates) double exponential improvement over the recent result by~\cite{shi2024finite}. Under mild assumptions on the kernel and potential, these bounds also grow polynomially in the dimension d. By adding a bilinear component to the kernel, the above approach is used to further obtain Wasserstein-2 convergence in continuous time. For the case of `bilinear + M

at\'ern' kernels, we derive Wasserstein-2 rates that exhibit a curse-of-dimensio
nality similar to the i.i.d. setting. We also obtain marginal convergence and lo
ng-time propagation of chaos results for the time-averaged particle laws.
****

Wenzhuo Liu,Fei Zhu,Longhui Wei,Qi Tian
C-CLIP: Multimodal Continual Learning for Vision-Language Model
Multimodal pre-trained models like CLIP need large image-text pairs for training
 but often struggle with domain-specific tasks. Since retraining with specialize
d and historical data incurs significant memory and time costs, it is important
to continually learn new domains in the open world while preserving original per
formance. However, current continual learning research mainly focuses on single-
modal scenarios, and the evaluation criteria are insufficient without considerin
g image-text matching performance and the forgetting of zero-shot performance. T
his work introduces image-caption datasets from various domains and establishes
a multimodal vision-language continual learning benchmark. Then, a novel framewo
rk named C-CLIP is proposed, which not only prevents forgetting but also enhance
s new task learning impressively. Comprehensive experiments demonstrate that our
 method has strong continual learning ability across different domain image-text
 datasets, and has little forgetting of the original capabilities of zero-shot p
rediction, significantly outperforming existing methods.
****

Rong-Xi Tan,Ke Xue,Shen-Huan Lyu,Haopu Shang,yaowang,Yaoyuan Wang,Fu Sheng,Chao
Qian
Offline Model-Based Optimization by Learning to Rank
Offline model-based optimization (MBO) aims to identify a design that maximizes
a black-box function using only a fixed, pre-collected dataset of designs and th
eir corresponding scores. This problem has garnered significant attention from b
oth scientific and industrial domains. A common approach in offline MBO is to tr
ain a regression-based surrogate model by minimizing mean squared error (MSE) an
d then find the best design within this surrogate model by different optimizers
(e.g., gradient ascent). However, a critical challenge is the risk of out-of-dis
tribution errors, i.e., the surrogate model may typically overestimate the score
s and mislead the optimizers into suboptimal regions. Prior works have attempted
 to address this issue in various ways, such as using regularization techniques
and ensemble learning to enhance the robustness of the model, but it still remai
ns. In this paper, we argue that regression models trained with MSE are not well
-aligned with the primary goal of offline MBO, which is to \textit{select} promi
sing designs rather than to predict their scores precisely. Notably, if a surrog
ate model can maintain the order of candidate designs based on their relative sc
ore relationships, it can produce the best designs even without precise predicti
ons. To validate it, we conduct experiments to compare the relationship between
the quality of the final designs and MSE, finding that the correlation is really
 very weak. In contrast, a metric that measures order-maintaining quality shows
a significantly stronger correlation. Based on this observation, we propose lear
ning a ranking-based model that leverages learning to rank techniques to priorit
ize promising designs based on their relative scores. We show that the generaliz
ation error on ranking loss can be well bounded. Empirical results across divers
e tasks demonstrate the superior performance of our proposed ranking-based metho
d than twenty existing methods. Our implementation is available at \url{https://
github.com/lamda-bbo/Offline-RaM}.
****

Luca Pinchetti,Chang Qi,Oleh Lokshyn,Cornelius Emde,Amine M'Charrak,Mufeng Tang,
Simon Frieder,Bayar Menzat,Gaspard Oliviers,Rafal Bogacz,Thomas Lukasiewicz,Tomm
aso Salvatori
Benchmarking Predictive Coding Networks -- Made Simple
In this work, we tackle the problems of efficiency and scalability for predictiv
e coding networks (PCNs) in machine learning. To do so, we propose a library th
at focuses on performance and simplicity, and use it to implement a large set of
 standard benchmarks for the community to use for their experiments. As most wor
ks in the field propose their own tasks and architectures, do not compare one ag

ainst each other, and focus on small-scale tasks, a simple and fast open-source library, and a comprehensive set of benchmarks, would address all of these concerns. Then, we perform extensive tests on such benchmarks using both existing algorithms for PCNs, as well as adaptations of other methods popular in the bio-plausible deep learning community. All of this has allowed us to (i) test architectures much larger than commonly used in the literature, on more complex datasets; (ii) reach new state-of-the-art results in all of the tasks and dataset provided; (iii) clearly highlight what the current limitations of PCNs are, allowing us to state important future research directions. With the hope of galvanizing community efforts towards one of the main open problems in the field, scalability, we will release the code, tests, and benchmarks.
****

Paolo Pellizzoni,Till Hendrik Schulz,Karsten Borgwardt
Graph Neural Networks Can (Often) Count Substructures
Message passing graph neural networks (GNNs) are known to have limited expressive power in their ability to distinguish some non-isomorphic graphs.
Because of this, it is well known that they are unable to detect or count arbitrary graph substructures (i.e., solving the subgraph isomorphism problem), a task that is of great importance for several types of graph-structured data.
However, we observe that GNNs are in fact able to count graph patterns quite accurately across several real-world graph datasets.
Motivated by this observation, we provide an analysis of the subgraph-counting capabilities of GNNs beyond the worst case, deriving several sufficient conditions for GNNs to be able to count subgraphs and, more importantly, to be able to sample-efficiently learn to count subgraphs.
Moreover, we develop novel dynamic programming algorithms for solving the subgraph isomorphism problem on restricted classes of pattern and target graphs, and show that message-passing GNNs can efficiently simulate these dynamic programs.
Finally, we empirically validate that our sufficient conditions for GNNs to count subgraphs hold on many real-world datasets, providing a theoretically-grounded explanation to our motivating observations.
****

Fusheng Liu,Qianxiao Li
Autocorrelation Matters: Understanding the Role of Initialization Schemes for State Space Models
Current methods for initializing state space model (SSM) parameters primarily rely on the HiPPO framework \citep{gu2023how}, which is based on online function approximation with the SSM kernel basis.
However, the HiPPO framework does not explicitly account for the effects of the temporal structures of input sequences on the optimization of SSMs.
In this paper, we take a further step to investigate the roles of SSM initialization schemes by considering the autocorrelation of input sequences.
Specifically, we: (1) rigorously characterize the dependency of the SSM timescale on sequence length based on sequence autocorrelation; (2) find that with a proper timescale, allowing a zero real part for the eigenvalues of the SSM state matrix mitigates the curse of memory while still maintaining stability at initialization; (3) show that the imaginary part of the eigenvalues of the SSM state matrix determines the conditioning of SSM optimization problems, and uncover an approximation-estimation tradeoff when training SSMs with a specific class of target functions.
****

Mayee F Chen,Michael Y. Hu,Nicholas Lourie,Kyunghyun Cho,Christopher Re
Aioli: A Unified Optimization Framework for Language Model Data Mixing
Language model performance depends on identifying the optimal mixture of data groups to train on (e.g., law, code, math). Prior work has proposed a diverse set of methods to efficiently learn mixture proportions, ranging from fitting regression models over training runs to dynamically updating proportions throughout training. Surprisingly, we find that no existing method consistently outperforms a simple stratified sampling baseline in terms of average test perplexity. To understand this inconsistency, we unify existing methods into a standard framework,

showing they are equivalent to solving a common optimization problem: minimize average loss subject to a method-specific mixing law---an implicit assumption on the relationship between loss and mixture proportions. This framework suggests that measuring the fidelity of a method's mixing law can offer insights into its performance. Empirically, we find that existing methods set their mixing law parameters inaccurately, resulting in the inconsistent mixing performance we observe. Using this insight, we derive a new online method named Aioli, which directly estimates the mixing law parameters throughout training and uses them to dynamically adjust proportions. Empirically, Aioli outperforms stratified sampling on 6 out of 6 datasets by an average of 0.27 test perplexity points, whereas existing methods fail to consistently beat stratified sampling, doing up to 6.9 points worse. Moreover, in a practical setting where proportions are learned on shorter runs due to computational constraints, Aioli can dynamically adjust these proportions over the full training run, consistently improving performance over existing methods by up to 12.012 test perplexity points.
****

Rui Ye,Jingyi Chai,Xiangrui Liu,Yaodong Yang,Yanfeng Wang,Siheng Chen
Emerging Safety Attack and Defense in Federated Instruction Tuning of Large Language Models
Federated learning (FL) enables multiple parties to collaboratively fine-tune an large language model (LLM) without the need of direct data sharing. Ideally, by training on decentralized data that is aligned with human preferences and safety principles, federated instruction tuning (FedIT) can result in an LLM that could behave helpfully and safely. In this paper, we for the first time reveal the vulnerability of safety alignment in FedIT by proposing a simple, stealthy, yet effective safety attack method. Specifically, the malicious clients could automatically generate attack data without involving manual efforts and attack the FedIT system by training their local LLMs on such attack data. Unfortunately, this proposed safety attack not only can compromise the safety alignment of LLM trained via FedIT, but also can not be effectively defended against by many existing FL defense methods. Targeting this, we further propose a post-hoc defense method, which could rely on a fully automated pipeline: generation of defense data and further fine-tuning of the LLM. Extensive experiments show that our safety attack method can significantly compromise the LLM's safety alignment (e.g., reduce safety rate by 70\%), which can not be effectively defended by existing defense methods (at most 4\% absolute improvement), while our safety defense method can significantly enhance the attacked LLM's safety alignment (at most 69\% absolute improvement). Code is available at https://github.com/19dx/FedLLM-Attack.
****

Lu Sang,Zehranaz Canfes,Dongliang Cao,Florian Bernard,Daniel Cremers
Implicit Neural Surface Deformation with Explicit Velocity Fields
In this work, we introduce the first unsupervised method that simultaneously predicts time-varying neural implicit surfaces and deformations between pairs of point clouds. We propose to model the point movement using an explicit velocity field and directly deform a time-varying implicit field using the modified level-set equation. This equation utilizes an iso-surface evolution with Eikonal constraints in a compact formulation, ensuring the integrity of the signed distance field. By applying a smooth, volume-preserving constraint to the velocity field, our method successfully recovers physically plausible intermediate shapes. Our method is able to handle both rigid and non-rigid deformations without any intermediate shape supervision. Our experimental results demonstrate that our method significantly outperforms existing works, delivering superior results in both quality and efficiency.
****

Bing Cao,Quanhao Lu,Jiekang Feng,Qilong Wang,Pengfei Zhu,Qinghua Hu
Efficient Masked AutoEncoder for Video Object Counting and A Large-Scale Benchmark
The dynamic imbalance of the fore-background is a major challenge in video object counting, which is usually caused by the sparsity of target objects. This remains understudied in existing works and often leads to severe under-/over-predict

ion errors. To tackle this issue in video object counting, we propose a density-embedded Efficient Masked Autoencoder Counting (E-MAC) framework in this paper. To empower the model’s representation ability on density regression, we develop a new Density-Embedded Masked mOdeling (DEMO) method, which first takes the density map as an auxiliary modality to perform multimodal self-representation learning for image and density map. Although DEMO contributes to effective cross-modal regression guidance, it also brings in redundant background information, making it difficult to focus on the foreground regions. To handle this dilemma, we propose an efficient spatial adaptive masking derived from density maps to boost efficiency. Meanwhile, we employ an optical flow-based temporal collaborative fusion strategy to effectively capture the dynamic variations across frames, aligning features to derive multi-frame density residuals. The counting accuracy of the current frame is boosted by harnessing the information from adjacent frames. In addition, considering that most existing datasets are limited to human-centric scenarios, we propose a large video bird counting dataset, DroneBird, in natural scenarios for migratory bird protection. Extensive experiments on three crowd datasets and our DroneBird validate our superiority against the counterparts. The code and dataset are available.
****

Fengyu Gao,Ruida Zhou,Tianhao Wang,Cong Shen,Jing Yang
Data-adaptive Differentially Private Prompt Synthesis for In-Context Learning
Large Language Models (LLMs) rely on the contextual information embedded in examples/demonstrations to perform in-context learning (ICL). To mitigate the risk of LLMs potentially leaking private information contained in examples in the prompt, we introduce a novel data-adaptive differentially private algorithm called **AdaDPSyn** to generate synthetic examples from the private dataset and then use these synthetic examples to perform ICL. The objective of AdaDPSyn is to adaptively adjust the noise level in the data synthesis mechanism according to the inherent statistical properties of the data, thereby preserving high ICL accuracy while maintaining formal differential privacy guarantees. A key innovation in AdaDPSyn is the *Precision-Focused Iterative Radius Reduction* technique, which dynamically refines the aggregation radius - the scope of data grouping for noise addition - based on patterns observed in data clustering, thereby minimizing the amount of additive noise. We conduct extensive experiments on standard benchmarks and compare AdaDPSyn with DP few-shot generation algorithm (Tang et al., 2023). The experiments demonstrate that AdaDPSyn not only outperforms DP few-shot generation, but also maintains high accuracy levels close to those of non-private baselines, providing an effective solution for ICL with privacy protection.
****

Zhongpai Gao,Benjamin Planche,Meng Zheng,Anwesa Choudhuri,Terrence Chen,Ziyan Wu
6DGS: Enhanced Direction-Aware Gaussian Splatting for Volumetric Rendering
Novel view synthesis has advanced significantly with the development of neural radiance fields (NeRF) and 3D Gaussian splatting (3DGS). However, achieving high quality without compromising real-time rendering remains challenging, particularly for physically-based rendering using ray/path tracing with view-dependent effects. Recently, N-dimensional Gaussians (N-DG) introduced a 6D spatial-angular representation to better incorporate view-dependent effects, but the Gaussian representation and control scheme are sub-optimal. In this paper, we revisit 6D Gaussians and introduce 6D Gaussian Splatting (6DGS), which enhances color and opacity representations and leverages the additional directional information in the 6D space for optimized Gaussian control. Our approach is fully compatible with the 3DGS framework and significantly improves real-time radiance field rendering by better modeling view-dependent effects and fine details. Experiments demonstrate that 6DGS significantly outperforms 3DGS and N-DG, achieving up to a 15.73 dB improvement in PSNR with a reduction of 66.5\% Gaussian points compared to 3DGS. The project page is: https://gaozhongpai.github.io/6dgs/.
****

Linbao Li,Yannan Liu,Daojing He,YU LI
One Model Transfer to All: On Robust Jailbreak Prompts Generation against LLMs
Safety alignment in large language models (LLMs) is increasingly compromised by

jailbreak attacks, which can manipulate these models to generate harmful or unin
tended content. Investigating these attacks is crucial for uncovering model vuln
erabilities. However, many existing jailbreak strategies fail to keep pace with
the rapid development of defense mechanisms, such as defensive suffixes, renderi
ng them ineffective against defended models. To tackle this issue, we introduce
a novel attack method called ArrAttack, specifically designed to target defended
 LLMs. ArrAttack automatically generates robust jailbreak prompts capable of byp
assing various defense measures. This capability is supported by a universal rob
ustness judgment model that, once trained, can perform robustness evaluation for
 any target model with a wide variety of defenses. By leveraging this model, we
can rapidly develop a robust jailbreak prompt generator that efficiently convert
s malicious input prompts into effective attacks. Extensive evaluations reveal t
hat ArrAttack significantly outperforms existing attack strategies, demonstratin
g strong transferability across both white-box and black-box models, including G
PT-4 and Claude-3. Our work bridges the gap between jailbreak attacks and defens
es, providing a fresh perspective on generating robust jailbreak prompts.
****

Yaxi Lu,Shenzhi Yang,Cheng Qian,Guirong Chen,Qinyu Luo,Yesai Wu,Huadong Wang,Xin
 Cong,Zhong Zhang,Yankai Lin,Weiwen Liu,Yasheng Wang,Zhiyuan Liu,Fangming Liu,Ma
osong Sun
Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistanc
e
Agents powered by large language models have shown remarkable abilities in solvi
ng complex tasks. However, most agent systems remain reactive, limiting their ef
fectiveness in scenarios requiring foresight and autonomous decision-making. In
this paper, we tackle the challenge of developing proactive agents capable of an
ticipating and initiating tasks without explicit human instructions. We propose
a novel data-driven approach for this problem. Firstly, we collect real-world hu
man activities to generate proactive task predictions. These predictions are the
n labeled by human annotators as either accepted or rejected. The labeled data i
s used to train a reward model that simulates human judgment and serves as an au
tomatic evaluator of the proactiveness of LLM agents. Building on this, we devel
op a comprehensive data generation pipeline to create a diverse dataset, Proacti
veBench, containing 6,790 events. Finally, we demonstrate that fine-tuning model
s with the proposed ProactiveBench can significantly elicit the proactiveness of
 LLM agents. Experimental results show that our fine-tuned model achieves an F1-
Score of 66.47% in proactively offering assistance, outperforming all open-sourc
e and close-source models. These results highlight the potential of our method i
n creating more proactive and effective agent systems, paving the way for future
 advancements in human-agent collaboration.
****

Yury Demidovich,Grigory Malinovsky,Egor Shulgin,Peter Richtárik
MAST: model-agnostic sparsified training
We introduce a novel optimization problem formulation that departs from the conv
entional way of minimizing machine learning model loss as a black-box function.
Unlike traditional formulations, the proposed approach explicitly incorporates a
n initially pre-trained model and random sketch operators, allowing for sparsifi
cation of both the model and gradient during training. We establish insightful p
roperties of the proposed objective function and highlight its connections to th
e standard formulation. Furthermore, we present several variants of the Stochast
ic Gradient Descent (SGD) method adapted to the new problem formulation, includi
ng SGD with general sampling, a distributed version, and SGD with variance reduc
tion techniques. We achieve tighter convergence rates and relax assumptions, bri
dging the gap between theoretical principles and practical applications, coverin
g several important techniques such as Dropout and Sparse training. This work pr
esents promising opportunities to enhance the theoretical understanding of model
 training through a sparsification-aware optimization approach.
****

Md Ashiqur Rahman,Raymond A. Yeh
Group Downsampling with Equivariant Anti-aliasing

Downsampling layers are crucial building blocks in CNN architectures, which help to increase the receptive field for learning high-level features and reduce the amount of memory/computation in the model. In this work, we study the generalization of the uniform downsampling layer for group equivariant architectures, e.g., G-CNNs. That is, we aim to downsample signals (feature maps) on general finite groups *with* anti-aliasing. This involves the following: **(a)** Given a finite group and a downsampling rate, we present an algorithm to form a suitable choice of subgroup. **(b)** Given a group and a subgroup, we study the notion of bandlimited-ness and propose how to perform anti-aliasing. Notably, our method generalizes the notion of downsampling based on classical sampling theory. When the signal is on a cyclic group, i.e., periodic, our method recovers the standard downsampling of an ideal low-pass filter followed by a subsampling operation. Finally, we conducted experiments on image classification tasks demonstrating that the proposed downsampling operation improves accuracy, better preserves equivariance, and reduces model size when incorporated into G-equivariant networks
****
Jingyang Ou,Shen Nie,Kaiwen Xue,Fengqi Zhu,Jiacheng Sun,Zhenguo Li,Chongxuan Li
Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data
Discrete diffusion models with absorbing processes have shown promise in language modeling. The key quantities to be estimated are the ratios between the marginal probabilities of two transitive states at all timesteps, called the concrete score. In this paper, we reveal that the concrete score in absorbing diffusion can be expressed as conditional probabilities of clean data, multiplied by a time-dependent scalar in an analytic form. Motivated by this finding, we propose reparameterized absorbing discrete diffusion (RADD), a dedicated diffusion model without time-condition that characterizes the time-independent conditional probabilities. Besides its simplicity, RADD can reduce the number of function evaluations (NFEs) by caching the output of the time-independent network when the noisy sample remains unchanged in a sampling interval, which enables sampling acceleration. Built upon the new perspective of conditional distributions, we further unify absorbing discrete diffusion and any-order autoregressive models (AO-ARMs), showing that the upper bound on the negative log-likelihood for the diffusion model can be interpreted as an expected negative log-likelihood for AO-ARMs. Further, our RADD models achieve SOTA performance among diffusion models on 5 zero-shot language modeling benchmarks (measured by perplexity) at the GPT-2 scale. Our code is available at \url{https://github.com/ML-GSAI/RADD}.
****
Boye Niu,Yiliao Song,Kai Lian,Yifan Shen,Yu Yao,Kun Zhang,Tongliang Liu
Flow: Modularized Agentic Workflow Automation
Multi-agent frameworks powered by large language models (LLMs) have demonstrated great success in automated planning and task execution. However, the effective adjustment of agentic workflows during execution has not been well studied. An effective workflow adjustment is crucial in real-world scenarios, as the initial plan must adjust to unforeseen challenges and changing conditions in real time to ensure the efficient execution of complex tasks. In this paper, we define workflows as an activity-on-vertex (AOV) graph, which allows continuous workflow refinement by LLM agents through dynamic subtask allocation adjustment based on historical performance and previous AOVs. To further enhance framework performance, we emphasize modularity in workflow design based on evaluating parallelism and dependency complexity. With this design, our proposed multi-agent framework achieves efficient concurrent execution of subtasks, effective goal achievement, and enhanced error tolerance. Empirical results across various practical tasks demonstrate significant improvements in the efficiency of multi-agent frameworks through dynamic workflow refinement and modularization.
****
Minkai Xu,Tomas Geffner,Karsten Kreis,Weili Nie,Yilun Xu,Jure Leskovec,Stefano Ermon,Arash Vahdat
Energy-Based Diffusion Language Models for Text Generation
Despite remarkable progress in autoregressive language models, alternative gener

ative paradigms beyond left-to-right generation are still being actively explore
d. Discrete diffusion models, with the capacity for parallel generation, have re
cently emerged as a promising alternative. Unfortunately, these models still und
erperform the autoregressive counterparts, with the performance gap increasing w
hen reducing the number of sampling steps. Our analysis reveals that this degrad
ation is a consequence of an imperfect approximation used by diffusion models. I
n this work, we propose Energy-based Diffusion Language Model (EDLM), an energy-
based model operating at the full sequence level for each diffusion step, introd
uced to improve the underlying approximation used by diffusion models. More spec
ifically, we introduce an EBM in a residual form, and show that its parameters c
an be obtained by leveraging a pretrained autoregressive model or by finetuning
a bidirectional transformer via noise contrastive estimation. We also propose an
 efficient generation algorithm via parallel important sampling. Comprehensive e
xperiments on language modeling benchmarks show that our model can consistently
outperform state-of-the-art diffusion models by a significant margin, and approa
ches autoregressive models' perplexity. We further show that, without any genera
tion performance drop, our framework offers a 1.3x sampling speedup over existin
g diffusion models. Reproduced code is available at https://github.com/MinkaiXu/
Energy-Diffusion-LLM.
****
Colin White,Samuel Dooley,Manley Roberts,Arka Pal,Benjamin Feuer,Siddhartha Jain
,Ravid Shwartz-Ziv,Neel Jain,Khalid Saifullah,Sreemanti Dey,Shubh-Agrawal,Sandee
p Singh Sandha,Siddartha Venkat Naidu,Chinmay Hegde,Yann LeCun,Tom Goldstein,Wil
lie Neiswanger,Micah Goldblum
LiveBench: A Challenging, Contamination-Limited LLM Benchmark
Test set contamination, wherein test data from a benchmark ends up in a newer mo
del's training set, is a well-documented obstacle for fair LLM evaluation and ca
n quickly render benchmarks obsolete. To mitigate this, many recent benchmarks c
rowdsource new prompts and evaluations from human or LLM judges; however, these
can introduce significant biases, and break down when scoring hard questions. In
 this work, we introduce a new benchmark for LLMs designed to be resistant to bo
th test set contamination and the pitfalls of LLM judging and human crowdsourcin
g. We release LiveBench, the first benchmark that (1) contains frequently-update
d questions from recent information sources, (2) scores answers automatically ac
cording to objective ground-truth values, and (3) contains a wide variety of cha
llenging tasks, spanning math, coding, reasoning, language, instruction followin
g, and data analysis. To achieve this, LiveBench contains questions that are bas
ed on recently-released math competitions, arXiv papers, news articles, and data
sets, and it contains harder, contamination-limited versions of tasks from previ
ous benchmarks such as Big-Bench Hard, AMPS, and IFEval. We evaluate many promin
ent closed-source models, as well as dozens of open-source models ranging from 0
.5B to 405B in size. LiveBench is difficult, with top models achieving below 70%
 accuracy. We release all questions, code, and model answers. Questions are adde
d and updated on a monthly basis, and we release new tasks and harder versions o
f tasks over time so that LiveBench can distinguish between the capabilities of
LLMs as they improve in the future. We welcome community engagement and collabor
ation for expanding the benchmark tasks and models.
****
Jeremy Cohen,Alex Damian,Ameet Talwalkar,J Zico Kolter,Jason D. Lee
Understanding Optimization in Deep Learning with Central Flows
Optimization in deep learning remains poorly understood. A key difficulty is th
at optimizers exhibit complex oscillatory dynamics, referred to as "edge of stab
ility," which cannot be captured by traditional optimization theory. In this pa
per, we show that the path taken by an oscillatory optimizer can often be captur
ed by a _central flow_: a differential equation which directly models the time-
averaged (i.e. smoothed) optimization trajectory. We empirically show that these
 central flows can predict long-term optimization trajectories for generic neura
l networks with a high degree of numerical accuracy. By interpreting these flow
s, we are able to understand how gradient descent makes progress even as the lo
ss sometimes goes up; how adaptive optimizers ``adapt'' to the local loss landsc

ape; and how adaptive optimizers implicitly seek out regions of weight space where they can take larger steps. These insights (and others) are not apparent from the optimizers' update rules, but are revealed by the central flows. Therefore, we believe that central flows constitute a promising tool for reasoning about optimization in deep learning.
****
Lei Li,Yuanxin Liu,Linli Yao,Peiyuan Zhang,Chenxin An,Lean Wang,Xu Sun,Lingpeng Kong,Qi Liu
Temporal Reasoning Transfer from Text to Video
Video Large Language Models (Video LLMs) have shown promising capabilities in video comprehension, yet they struggle with tracking temporal changes and reasoning about temporal relationships.
While previous research attributed this limitation to the ineffective temporal encoding of visual inputs, our diagnostic study reveals that video representations contain sufficient information for even small probing classifiers to achieve perfect accuracy.
Surprisingly, we find that the key bottleneck in Video LLMs' temporal reasoning capability stems from the underlying LLM's inherent difficulty with temporal concepts, as evidenced by poor performance on textual temporal question-answering tasks.
Building on this discovery, we introduce the Textual Temporal reasoning Transfer (T3).
T3 synthesizes diverse temporal reasoning tasks in pure text format from existing image-text datasets, addressing the scarcity of video samples with complex temporal scenarios.
Remarkably, without using any video data, T3 enhances LongVA-7B's temporal understanding, yielding a 5.3 absolute accuracy improvement on the challenging TempCompass benchmark, which enables our model to outperform ShareGPT4Video-8B trained on 28,000 video samples.
Additionally, the enhanced LongVA-7B model achieves competitive performance on comprehensive video benchmarks. For example, it achieves a 49.7 accuracy on the Temporal Reasoning task of Video-MME, surpassing powerful large-scale models such as InternVL-Chat-V1.5-20B and VILA1.5-40B.
Further analysis reveals a strong correlation between textual and video temporal task performance, validating the efficacy of transferring temporal reasoning abilities from text to video domains.
****
Leitian Tao,Yixuan Li
Your Weak LLM is Secretly a Strong Teacher for Alignment
The burgeoning capabilities of large language models (LLMs) have underscored the need for alignment to ensure these models act in accordance with human values and intentions. Existing alignment frameworks present constraints either in the form of expensive human effort or high computational costs. This paper explores a promising middle ground, where we employ a weak LLM that is significantly less resource-intensive than top-tier models, yet offers more automation than purely human feedback. We present a systematic study to evaluate and understand weak LLM's ability to generate feedback for alignment. Our empirical findings demonstrate that weak LLMs can provide feedback that rivals or even exceeds that of fully human-annotated data. Our study indicates a minimized impact of model size on feedback efficacy, shedding light on a scalable and sustainable alignment strategy. To deepen our understanding of alignment under weak LLM feedback, we conduct a series of qualitative and quantitative analyses, offering novel insights into the quality discrepancies between human feedback vs. weak LLM feedback. Code is publicly available at https://github.com/deeplearning-wisc/weak_llm_teacher.
****
Cheng Yang,Chufan Shi,Yaxin Liu,Bo Shui,Junjie Wang,Mohan Jing,Linran XU,Xinyu Zhu,Siheng Li,Yuxiang Zhang,Gongye Liu,Xiaomei Nie,Deng Cai,Yujiu Yang
ChartMimic: Evaluating LMM's Cross-Modal Reasoning Capability via Chart-to-Code Generation
We introduce a new benchmark, ChartMimic, aimed at assessing the visually-ground

ed code generation capabilities of large multimodal models (LMMs). ChartMimic ut
ilizes information-intensive visual charts and textual instructions as inputs, r
equiring LMMs to generate the corresponding code for chart rendering.
ChartMimic includes $4,800$ human-curated (figure, instruction, code) triplets,
which represent the authentic chart use cases found in scientific papers across
various domains (e.g., Physics, Computer Science, Economics, etc). These charts
span 18 regular types and 4 advanced types, diversifying into 201 subcateg
ories.
Furthermore, we propose multi-level evaluation metrics to provide an automatic a
nd thorough assessment of the output code and the rendered charts.
Unlike existing code generation benchmarks, ChartMimic places emphasis on evalua
ting LMMs' capacity to harmonize a blend of cognitive capabilities, encompassing
 visual understanding, code generation, and cross-modal reasoning. The evaluatio
n of 3 proprietary models and 14 open-weight models highlights the substanti
al challenges posed by ChartMimic. Even the advanced GPT-4o, InternVL2-Llama3-76
B only achieved an average score across Direct Mimic and Customized Mimic tasks
of 82.2 and 61.6, respectively, indicating significant room for improvement.

We anticipate that ChartMimic will inspire the development of LMMs, advancing th
e pursuit of artificial general intelligence.
****
Wonjeong Choi,Do-Yeon Kim,Jungwuk Park,Jungmoon Lee,Younghyun Park,Dong-Jun Han,
Jaekyun Moon
Adaptive Energy Alignment for Accelerating Test-Time Adaptation
In response to the increasing demand for tackling out-of-domain (OOD) scenarios,
 test-time adaptation (TTA) has garnered significant research attention in recen
t years. To adapt a source pre-trained model to target samples without getting a
ccess to their labels, existing approaches have typically employed entropy minim
ization (EM) loss as a primary objective function. In this paper, we propose an
adaptive energy alignment (AEA) solution that achieves fast online TTA. We start
 from the re-interpretation of the EM loss by decomposing it into two energy-bas
ed terms with conflicting roles, showing that the EM loss can potentially hinder
 the assertive model adaptation. Our AEA addresses this challenge by strategical
ly reducing the energy gap between the source and target domains during TTA, aim
ing to effectively align the target domain with the source domains and thus to
accelerate adaptation. We specifically propose two novel strategies, each contri
buting a necessary component for TTA: (i) aligning the energy level of each targ
et sample with the energy zone of the source domain that the pre-trained model i
s already familiar with, and (ii) precisely guiding the direction of the energy
alignment by matching the class-wise correlations between the source and target
domains. Our approach demonstrates its effectiveness on various domain shift dat
asets including CIFAR10-C, CIFAR100-C, and TinyImageNet-C.
****
Yikun Bai,Rocio Diaz Martin,Abihith Kothapalli,Hengrong Du,Xinran Liu,Soheil Kol
ouri
Partial Gromov-Wasserstein Metric
The Gromov-Wasserstein (GW) distance has gained increasing interest in the machi
ne learning community in recent years, as it allows for the comparison of measur
es in different metric spaces. To overcome the limitations imposed by the equal
mass requirements of the classical GW problem, researchers have begun exploring
its application in unbalanced settings. However, Unbalanced GW (UGW) can only be
 regarded as a discrepancy rather than a rigorous metric/distance between two me
tric measure spaces (mm-spaces). In this paper, we propose a particular case of
the UGW problem, termed Partial Gromov-Wasserstein (PGW). We establish that PGW
is a well-defined metric between mm-spaces and discuss its theoretical propertie
s, including the existence of a minimizer for the PGW problem and the relationsh
ip between PGW and GW, among others. We then propose two variants of the Frank-W
olfe algorithm for solving the PGW problem and show that they are mathematically
 and computationally equivalent. Moreover, based on our PGW metric, we introduce
 the analogous concept of barycenters for mm-spaces. Finally, we validate the ef

fectiveness of our PGW metric and related solvers in applications such as shape matching, shape retrieval, and shape interpolation, comparing them against existing baselines. Our code is available at https://github.com/mint-vu/PGW_Metric.
****

Renjie Pi,Jianshu Zhang,Tianyang Han,Jipeng Zhang,Rui Pan,Tong Zhang
Personalized Visual Instruction Tuning
Recent advancements in multimodal large language models (MLLMs) have demonstrated significant progress; however, these models exhibit a notable limitation, which we refer to as "face blindness." Specifically, they can engage in general conversations but fail to conduct personalized dialogues targeting at specific individuals. This deficiency hinders the application of MLLMs in personalized settings, such as tailored visual assistants on mobile devices, or domestic robots that need to recognize members of the family. In this paper, we introduce Personalized Visual Instruction Tuning (PVIT), a novel data curation and training framework designed to enable MLLMs to identify target individuals within an image and engage in personalized and coherent dialogues. Our approach involves the development of a sophisticated pipeline that autonomously generates training data containing personalized conversations. This pipeline leverages the capabilities of various visual experts, image generation models, and (multi-modal) large language models. To evaluate the personalized potential of MLLMs, we present a benchmark called P-Bench, which encompasses various question types with different levels of difficulty. The experiments demonstrate a substantial personalized performance enhancement after fine-tuning with our curated dataset.
****

Chaochen Gao,Xing W,Qi Fu,Songlin Hu
Quest: Query-centric Data Synthesis Approach for Long-context Scaling of Large Language Model
Recent advancements in large language models (LLMs) have highlighted the importance of extending context lengths for handling complex tasks. While traditional methods for training on long contexts often use filtered long documents, these approaches lead to domain imbalances, limiting model performance. To address this, techniques like random document concatenation (Standard) and similarity-based methods (KNN, ICLM) have been developed. However, they either sacrifice semantic coherence or diversity. To balance both aspects, we introduce Quest, a query-centric data synthesis method aggregating semantically relevant yet diverse documents. Quest uses a generative model to predict potential queries for each document, grouping documents with similar queries and keywords. Extensive experiments demonstrate Quest's superior performance on long-context tasks, achieving remarkable results with context lengths of up to 1M tokens and confirming its scalability across various model sizes.
****

Xia Jiang,Yaoxin Wu,Chenhao Zhang,Yingqian Zhang
DRoC: Elevating Large Language Models for Complex Vehicle Routing via Decomposed Retrieval of Constraints
This paper proposes Decomposed Retrieval of Constraints (DRoC), a novel framework aimed at enhancing large language models (LLMs) in exploiting solvers to tackle vehicle routing problems (VRPs) with intricate constraints. While LLMs have shown promise in solving simple VRPs, their potential in addressing complex VRP variants is still suppressed, due to the limited embedded internal knowledge that is required to accurately reflect diverse VRP constraints. Our approach mitigates the issue by integrating external knowledge via a novel retrieval-augmented generation (RAG) approach. More specifically, the DRoC decomposes VRP constraints, externally retrieves information relevant to each constraint, and synergistically combines internal and external knowledge to benefit the program generation for solving VRPs. The DRoC also allows LLMs to dynamically select between RAG and self-debugging mechanisms, thereby optimizing program generation without the need for additional training. Experiments across 48 VRP variants exhibit the superiority of DRoC, with significant improvements in the accuracy rate and runtime error rate delivered by the generated programs. The DRoC framework has the potential to elevate LLM performance in complex optimization tasks, fostering the appli

cability of LLMs in industries such as transportation and logistics.
****
Chien-yu Huang,Wei-Chih Chen,Shu-wen Yang,Andy T. Liu,Chen-An Li,Yu-Xiang Lin,Wei-Cheng Tseng,Anuj Diwan,Yi-Jen Shih,Jiatong Shi,William Chen,Chih-Kai Yang,Xuanjun Chen,Chi-Yuan Hsiao,Puyuan Peng,Shih-Heng Wang,Chun-Yi Kuan,Ke-Han Lu,Kai-Wei Chang,Fabian Alejandro Ritter Gutierrez,Huang Kuan-Po,Siddhant Arora,You-Kuan Lin,CHUANG Ming To,Eunjung Yeo,Kalvin Chang,Chung-Ming Chien,Kwanghee Choi,Cheng-Hsiu Hsieh,Yi-Cheng Lin,Chee-En Yu,I-Hsiang Chiu,Heitor Guimarães,Jionghao Han,Tzu-Quan Lin,Tzu-Yuan Lin,Homu Chang,Ting-Wu Chang,Chun Wei Chen,Shou-Jen Chen,Yu-Hua Chen,Hsi-Chun Cheng,Kunal Dhawan,Jia-Lin Fang,Shi-Xin Fang,KUAN YU FANG CHIANG,Chi An Fu,Hsien-Fu Hsiao,Ching Yu Hsu,Shao-Syuan Huang,Lee Chen Wei,Hsi-Che Lin,Hsuan-Hao Lin,Hsuan-Ting Lin,Jian-Ren Lin,Ting-Chun Liu,Li-Chun Lu,Tsung-Min Pai,Ankita Pasad,Shih-Yun Shan Kuan,Suwon Shon,Yuxun Tang,Yun-Shao Tsai,Wei Jui Chiang,Tzu-Chieh Wei,Chengxi Wu,Dien-Ruei Wu,Chao-Han Huck Yang,Chieh-Chi Yang,Jia Qi Yip,Shao-Xiang Yuan,Haibin Wu,Karen Livescu,David Harwath,Shinji Watanabe,Hung-yi Lee
Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks
Multimodal foundation models, such as Gemini and ChatGPT, have revolutionized human-machine interactions by seamlessly integrating various forms of data. Developing a universal spoken language model that comprehends a wide range of natural language instructions is critical for bridging communication gaps and facilitating more intuitive interactions. However, the absence of a comprehensive evaluation benchmark poses a significant challenge. We present Dynamic-SUPERB Phase-2, an open and evolving benchmark for the comprehensive evaluation of instruction-based universal speech models. Building upon the first generation, this second version incorporates 125 new tasks contributed collaboratively by the global research community, expanding the benchmark to a total of 180 tasks, making it the largest benchmark for speech and audio evaluation. While the first generation of Dynamic-SUPERB was limited to classification tasks, Dynamic-SUPERB Phase-2 broadens its evaluation capabilities by introducing a wide array of novel and diverse tasks, including regression and sequence generation, across speech, music, and environmental audio. Evaluation results show that no model performed well universally. SALMONN-13B excelled in English ASR and Qwen2-Audio-7B-Instruct showed high accuracy in emotion recognition, but current models still require further innovations to handle a broader range of tasks. We open-source all task data and the evaluation pipeline at https://github.com/dynamic-superb/dynamic-superb.
****
Jun Zhang,Jue WANG,Huan Li,Lidan Shou,Ke Chen,Yang You,Guiming Xie,Xuejian Gong,Kunlong Zhou
Train Small, Infer Large: Memory-Efficient LoRA Training for Large Language Models
Large Language Models (LLMs) have significantly advanced natural language processing with exceptional task generalization capabilities. Low-Rank Adaption (LoRA) offers a cost-effective fine-tuning solution, freezing the original model parameters and training only lightweight, low-rank adapter matrices. However, the memory footprint of LoRA is largely dominated by the original model parameters. To mitigate this, we propose LoRAM, a memory-efficient LoRA training scheme founded on the intuition that many neurons in over-parameterized LLMs have low training utility but are essential for inference. LoRAM presents a unique twist: it trains on a pruned (small) model to obtain pruned low-rank matrices, which are then recovered and utilized with the original (large) model for inference. Additionally, minimal-cost continual pre-training, performed by the model publishers in advance, aligns the knowledge discrepancy between pruned and original models. Our extensive experiments demonstrate the efficacy of LoRAM across various pruning strategies and downstream tasks. For a model with 70 billion parameters, LoRAM enables training on a GPU with only 20G HBM, replacing an A100-80G GPU for LoRA training and 15 GPUs for full fine-tuning. Specifically, QLoRAM implemented by structured pruning combined with 4-bit quantization, for LLaMA-3.1-70B (LLaMA-2-70B), reduces the parameter storage cost that dominates the memory usage in low-ran

k matrix training by 15.81× (16.95×), while achieving dominant performance gains over both the original LLaMA-3.1-70B (LLaMA-2-70B) and LoRA-trained LLaMA-3.1-8B (LLaMA-2-13B). Code is available at https://github.com/junzhang-zj/LoRAM.
****

Lei Yu,Virginie Do,Karen Hambardzumyan,Nicola Cancedda
Robust LLM safeguarding via refusal feature adversarial training
Large language models (LLMs) are vulnerable to adversarial attacks that can elicit harmful responses. Defending against such attacks remains challenging due to the opacity of jailbreaking mechanisms and the high computational cost of training LLMs robustly. We demonstrate that adversarial attacks share a universal mechanism for circumventing LLM safeguards that works by ablating a dimension in the residual stream embedding space called the refusal feature. We further show that the operation of refusal feature ablation (RFA) approximates the worst-case perturbation of offsetting model safety. Based on these findings, we propose Refusal Feature Adversarial Training (ReFAT), a novel algorithm that efficiently performs LLM adversarial training by simulating the effect of input-level attacks via RFA. Experiment results show that ReFAT significantly improves the robustness of three popular LLMs against a wide range of adversarial attacks, with considerably less computational overhead compared to existing adversarial training methods.
****

Peng Xia,Kangyu Zhu,Haoran Li,Tianze Wang,Weijia Shi,Sheng Wang,Linjun Zhang,James Zou,Huaxiu Yao
MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models
Artificial Intelligence (AI) has demonstrated significant potential in healthcare, particularly in disease diagnosis and treatment planning. Recent progress in Medical Large Vision-Language Models (Med-LVLMs) has opened up new possibilities for interactive diagnostic tools. However, these models often suffer from factual hallucination, which can lead to incorrect diagnoses. Fine-tuning and retrieval-augmented generation (RAG) have emerged as methods to address these issues. However, the amount of high-quality data and distribution shifts between training data and deployment data limit the application of fine-tuning methods. Although RAG is lightweight and effective, existing RAG-based approaches are not sufficiently general to different medical domains and can potentially cause misalignment issues, both between modalities and between the model and the ground truth. In this paper, we propose a versatile multimodal RAG system, MMed-RAG, designed to enhance the factuality of Med-LVLMs. Our approach introduces a domain-aware retrieval mechanism, an adaptive retrieved contexts selection, and a provable RAG-based preference fine-tuning strategy. These innovations make the RAG process sufficiently general and reliable, significantly improving alignment when introducing retrieved contexts. Experimental results across five medical datasets (involving radiology, ophthalmology, pathology) on medical VQA and report generation demonstrate that MMed-RAG can achieve an average improvement of 43.8% in factual accuracy in the factual accuracy of Med-LVLMs.
****

Yinan Huang,Haoyu Peter Wang,Pan Li
What Are Good Positional Encodings for Directed Graphs?
Positional encodings (PEs) are essential for building powerful and expressive graph neural networks and graph transformers, as they effectively capture the relative spatial relationships between nodes. Although extensive research has been devoted to PEs in undirected graphs, PEs for directed graphs remain relatively unexplored. This work seeks to address this gap. We first introduce the notion of *Walk Profile*, a generalization of walk-counting sequences for directed graphs. A walk profile encompasses numerous structural features crucial for directed graph-relevant applications, such as program analysis and circuit performance prediction. We identify the limitations of existing PE methods in representing walk profiles and propose a novel *Multi-q Magnetic Laplacian PE*, which extends the Magnetic Laplacian eigenvector-based PE by incorporating multiple potential factors. The new PE can provably express walk profiles. Furthermore, we generalize prior basis-invariant neural networks to enable the stable use of the new PE in t

he complex domain. Our numerical experiments validate the expressiveness of the proposed PEs and demonstrate their effectiveness in solving sorting network satisfiability and performing well on general circuit benchmarks. Our code is available at https://github.com/Graph-COM/Multi-q-Maglap.
****

Xin Gao,Jian Pu
Deep Incomplete Multi-view Learning via Cyclic Permutation of VAEs
Multi-View Representation Learning (MVRL) aims to derive a unified representation from multi-view data by leveraging shared and complementary information across views. However, when views are irregularly missing, the incomplete data can lead to representations that lack sufficiency and consistency. To address this, we propose Multi-View Permutation of Variational Auto-Encoders (MVP), which excavates invariant relationships between views in incomplete data. MVP establishes inter-view correspondences in the latent space of Variational Auto-Encoders, enabling the inference of missing views and the aggregation of more sufficient information. To derive a valid Evidence Lower Bound (ELBO) for learning, we apply permutations to randomly reorder variables for cross-view generation and then partition them by views to maintain invariant meanings under permutations. Additionally, we enhance consistency by introducing an informational prior with cyclic permutations of posteriors, which turns the regularization term into a similarity measure across distributions. We demonstrate the effectiveness of our approach on seven diverse datasets with varying missing ratios, achieving superior performance in multi-view clustering and generation tasks.
****

Yingshan Chang,Yonatan Bisk
Language Models Need Inductive Biases to Count Inductively
Counting constitutes a core skill underlying a wide range of tasks, such as formal language recognition, multi-hop reasoning and simulating algorithms. Generali z- ing counting inductively is central to task success on out-of-distribution (OOD) instances where testing inputs are longer than those seen in training. While there is a large body of literature reporting poor length generalization in language models, few papers have tried to distill the “reasoning” failure to the simplest case of count- ing failure. We aim to provide a broader picture on whether various language model architectures can a) learn to count, and b) generalize counting inductively. This work provides extensive empirical results on architectures ranging from RNNs, Transformers, State-Space Models and RWKV. We present carefully-designed task formats, auxiliary tasks and positional embeddings to avoid limitations in general- ization with OOD-position and OOD-vocabulary. We find that while traditional RNNs trivially achieve inductive counting, Transformers have to rely on positional embeddings (PEs) to count OOD. Further analyses on interpreting the learned solution reveal that different PEs encode different inductive biases that facilitate counting in different task formats. As counting is the basis for many arguments concerning the expressivity of Transformers, our finding calls for the community to reexamine the application scope of primitive functions defined in formal charac- terizations. Finally, modern RNNs also largely underperform traditional RNNs in generalizing counting inductively, hinting at the tradeoff modern RNNs struggle to balance between parallelized training and maintaining their recurrent nature.
****

Long Le,Jason Xie,William Liang,Hung-Ju Wang,Yue Yang,Yecheng Jason Ma,Kyle Vedder,Arjun Krishna,Dinesh Jayaraman,Eric Eaton
Articulate-Anything: Automatic Modeling of Articulated Objects via a Vision-Language Foundation Model
Interactive 3D simulated objects are crucial in AR/VR, animations, and robotics, driving immersive experiences and advanced automation.
However, creating these articulated objects requires extensive human effort and expertise, limiting their broader applications. To overcome this challenge, we present Articulate-Anything, a system that automates the articulation of diverse, complex objects from many input modalities, including text, images, and videos. Articulate-Anything leverages vision-language models (VLMs) to generate code th

at can be compiled into an interactable digital twin for use in standard 3D simu
lators. Our system exploits existing 3D asset datasets via a mesh retrieval mech
anism, along with an actor-critic system that iteratively proposes, evaluates, a
nd refines solutions for articulating the objects, self-correcting errors to ach
ieve a robust out- come. Qualitative evaluations demonstrate Articulate-Anything
's capability to articulate complex and even ambiguous object affordances by lev
eraging rich grounded inputs. In extensive quantitative experiments on the stand
ard PartNet-Mobility dataset, Articulate-Anything substantially outperforms prio
r work, increasing the success rate from 8.7-11.6\% to 75\% and setting a new ba
r for state-of-art performance. We further showcase the utility of our generate
d assets by using them to train robotic policies for fine-grained manipulation t
asks that go beyond basic pick and place.
****
Guobin Shen,Dongcheng Zhao,Yiting Dong,Xiang He,Yi Zeng
Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adj
ustment in Large Language Models
As large language models (LLMs) become integral to various applications, ensurin
g both their safety and utility is paramount. Jailbreak attacks, which manipulat
e LLMs into generating harmful content, pose significant challenges to this bala
nce. Existing defenses, such as prompt engineering and safety fine-tuning, often
 introduce computational overhead, increase inference latency, and lack runtime
flexibility. Moreover, overly restrictive safety measures can degrade model util
ity by causing refusals of benign queries. In this paper, we introduce *Jailbrea
k Antidote*, a method that enables real-time adjustment of LLM safety preference
s by manipulating a sparse subset of the model's internal states during inferenc
e. By shifting the model's hidden representations along a safety direction with
varying strengths, we achieve flexible control over the safety-utility balance w
ithout additional token overhead or inference delays. Our analysis reveals that
safety-related information in LLMs is sparsely distributed; adjusting approximat
ely *5\%* of the internal state is as effective as modifying the entire state. E
xtensive experiments on nine LLMs (ranging from 2 billion to 72 billion paramete
rs), evaluated against ten jailbreak attack methods and compared with six defens
e strategies, validate the effectiveness and efficiency of our approach. By dire
ctly manipulating internal states during reasoning, *Jailbreak Antidote* offers
a lightweight, scalable solution that enhances LLM safety while preserving utili
ty, opening new possibilities for real-time safety mechanisms in widely-deployed
 AI systems.
****
Yongyi Yang,Core Francisco Park,Ekdeep Singh Lubana,Maya Okawa,Wei Hu,Hidenori T
anaka
Swing-by Dynamics in Concept Learning and Compositional Generalization
Prior work has shown that text-conditioned diffusion models can learn to identif
y and manipulate primitive concepts underlying a compositional data-generating p
rocess, enabling generalization to entirely novel, out-of-distribution compositi
ons.
Beyond performance evaluations, these studies develop a rich empirical phenomeno
logy of learning dynamics, showing that models generalize sequentially, respecti
ng the compositional hierarchy of the data-generating process.
Moreover, concept-centric structures within the data significantly influence a m
odel's speed of learning the ability to manipulate a concept.
In this paper, we aim to better characterize these empirical results from a theo
retical standpoint.
Specifically, we propose an abstraction of prior work's compositional generaliza
tion problem by introducing a structured identity mapping (SIM) task, where a mo
del is trained to learn the identity mapping on a Gaussian mixture with structur
ally organized centroids.
We mathematically analyze the learning dynamics of neural networks trained on th
is SIM task and show that, despite its simplicity, SIM's learning dynamics captu
re and help explain key empirical observations on compositional generalization w
ith diffusion models identified in prior work.

Our theory also offers several new insights---e.g., we find a novel mechanism fo
r non-monotonic learning dynamics of test loss in early phases of training.
We validate our new predictions by training a text-conditioned diffusion model,
bridging our simplified framework and complex generative models.
Overall, this work establishes the SIM task as a meaningful theoretical abstract
ion of concept learning dynamics in modern generative models.
****

Edoardo Cetin,Qi Sun,Tianyu Zhao,Yujin Tang
An Evolved Universal Transformer Memory
Prior methods propose to offset the escalating costs of modern foundation models
 by dropping specific parts of their contexts with hand-designed rules, while at
tempting to preserve their original performance. We overcome this trade-off with
 Neural Attention Memory Models (NAMMs), introducing a learned network for memor
y management that improves both the performance and efficiency of transformers.
We evolve NAMMs atop pre-trained transformers to provide different latent contex
ts focusing on the most relevant information for individual layers and attention
 heads. NAMMs are universally applicable to any model using self-attention as th
ey condition exclusively on the values in the produced attention matrices. Learn
ing NAMMs on a small set of problems, we achieve substantial performance improve
ments across multiple long-context benchmarks while cutting the model's input co
ntexts up to a fraction of the original sizes. We show the generality of our con
ditioning enables zero-shot transfer of NAMMs trained only on language to entire
ly new transformer architectures even across input modalities, with their benefi
ts carrying over to vision and reinforcement learning.
****

Bokun Wang,Yunwen Lei,Yiming Ying,Tianbao Yang
On Discriminative Probabilistic Modeling for Self-Supervised Representation Lear
ning
We study the discriminative probabilistic modeling on a continuous domain for th
e data prediction task of (multimodal) self-supervised representation learning.
To address the challenge of computing the integral in the partition function for
 each anchor data, we leverage the multiple importance sampling (MIS) technique
for robust Monte Carlo integration, which can recover InfoNCE-based contrastive
loss as a special case. Within this probabilistic modeling framework, we conduct
 generalization error analysis to reveal the limitation of current InfoNCE-based
 contrastive loss for self-supervised representation learning and derive insight
s for developing better approaches by reducing the error of Monte Carlo integrat
ion. To this end, we propose a novel non-parametric method for approximating the
 sum of conditional probability densities required by MIS through convex optimiz
ation, yielding a new contrastive objective for self-supervised representation l
earning. Moreover, we design an efficient algorithm for solving the proposed obj
ective. We empirically compare our algorithm to representative baselines on the
contrastive image-language pretraining task. Experimental results on the CC3M an
d CC12M datasets demonstrate the superior overall performance of our algorithm.
Our code is available at https://github.com/bokun-wang/NUCLR.
****

Tianyu Zhang,Suyuchen Wang,Lu Li,Ge Zhang,Perouz Taslakian,Sai Rajeswar,Jie Fu,B
ang Liu,Yoshua Bengio
VCR: A Task for Pixel-Level Complex Reasoning in Vision Language Models via Rest
oring Occluded Text
We introduce Visual Caption Restoration (VCR), a novel vision-language task that
 challenges models to accurately restore partially obscured texts using pixel-le
vel hints within images through complex reasoning. This task stems from the obse
rvation that text embedded in images intrinsically differs from common visual el
ements and text due to the need to align the modalities of vision, text, and tex
t embedded in images. While many works incorporate text into images for visual q
uestion answering, they mostly rely on OCR or masked language modeling, reducing
 the task to text-based processing. However, text-based processing becomes ineff
ective in VCR as accurate text restoration depends on the combined information f
rom provided images, context, and subtle cues from the tiny, exposed areas of ma

sked texts. We develop a pipeline to generate synthetic images for the VCR task using image-caption pairs, with adjustable caption visibility to control the tas k difficulty. With this pipeline, we construct VCR-WIKI for VCR using Wikipedia images with captions, including 2.11M English and 346K Chinese training entities , plus 5K validation and 5K test entities in both languages, each in easy and ha rd configurations. We also make a hidden test set, VCR-HIDDEN, to avoid potentia l overfitting on VCR-WIKI. Our results reveal that current vision-language model s significantly lag behind human performance in the VCR task, and merely fine-tu ning the models on our dataset does not lead to notable improvements. We release VCR-WIKI and the data construction code to facilitate future research.
****

Mohan Xu,Kai Li,Guo Chen,Xiaolin Hu

TIGER: Time-frequency Interleaved Gain Extraction and Reconstruction for Efficie nt Speech Separation

In recent years, much speech separation research has focused primarily on improv ing model performance. However, for low-latency speech processing systems, high efficiency is equally important. Therefore, we propose a speech separation model with significantly reduced parameters and computational costs: Time-frequency I nterleaved Gain Extraction and Reconstruction network (TIGER). TIGER leverages p rior knowledge to divide frequency bands and compresses frequency information. W e employ a multi-scale selective attention module to extract contextual features , while introducing a full-frequency-frame attention module to capture both temp oral and frequency contextual information. Additionally, to more realistically e valuate the performance of speech separation models in complex acoustic environm ents, we introduce a dataset called EchoSet. This dataset includes noise and mor e realistic reverberation (e.g., considering object occlusions and material prop erties), with speech from two speakers overlapping at random proportions. Experi mental results showed that models trained on EchoSet had better generalization a bility than those trained on other datasets to the data collected in the physica l world, which validated the practical value of the EchoSet. On EchoSet and real -world data, TIGER significantly reduces the number of parameters by 94.3% and t he MACs by 95.3% while achieving performance surpassing state-of-the-art (SOTA) model TF-GridNet.
****

Shikun Sun,Min Zhou,Zixuan Wang,Xubin Li,Tiezheng Ge,Zijie Ye,Xiaoyu Qin,Junlian g Xing,Bo Zheng,Jia Jia

Minimal Impact ControlNet: Advancing Multi-ControlNet Integration

With the advancement of diffusion models, there is a growing demand for high-qua lity, controllable image generation, particularly through methods that utilize o ne or multiple control signals based on ControlNet. However, in current ControlN et training, each control is designed to influence all areas of an image, which can lead to conflicts when different control signals are expected to manage diff erent parts of the image in practical applications. This issue is especially pro nounced with edge-type control conditions, where regions lacking boundary inform ation often represent low-frequency signals, referred to as silent control signa ls. When combining multiple ControlNets, these silent control signals can suppre ss the generation of textures in related areas, resulting in suboptimal outcomes . To address this problem, we propose Minimal Impact ControlNet. Our approach mi tigates conflicts through three key strategies: constructing a balanced dataset, combining and injecting feature signals in a balanced manner, and addressing th e asymmetry in the score function’s Jacobian matrix induced by ControlNet. These improvements enhance the compatibility of control signals, allowing for freer a nd more harmonious generation in areas with silent control signals.
****

Haixin Zhong,Haoyu Wang,Wei P Dai,Yuchao Huang,Mingyi Huang,Rubin Wang,Anna Wang Roe,yuguo yu

Emergent Orientation Maps —— Mechanisms, Coding Efficiency and Robustness

Extensive experimental studies have shown that in lower mammals, neuronal orient ation preference in the primary visual cortex is organized in disordered "salt-a nd-pepper" organizations. In contrast, higher-order mammals display a continuous

variation in orientation preference, forming pinwheel-like structures. Despite these observations, the spiking mechanisms underlying the emergence of these distinct topological structures and their functional roles in visual processing remain poorly understood. To address this, we developed a self-evolving spiking neural network model with Hebbian plasticity, trained using physiological parameters characteristic of rodents, cats, and primates, including retinotopy, neuronal morphology, and connectivity patterns. Our results identify critical factors, such as the degree of input visual field overlap, neuronal connection range, and the balance between localized connectivity and long-range competition, that determine the emergence of either salt-and-pepper or pinwheel-like topologies. Furthermore, we demonstrate that pinwheel structures exhibit lower wiring costs and enhanced sparse coding capabilities compared to salt-and-pepper organizations. They also maintain greater coding robustness against noise in naturalistic visual stimuli. These findings suggest that such topological structures confer significant computational advantages in visual processing and highlight their potential application in the design of brain-inspired deep learning networks and algorithms.

****

Qijun Gan,Song Wang,Shengtao Wu,Jianke Zhu

PianoMotion10M: Dataset and Benchmark for Hand Motion Generation in Piano Performance

Recently, artificial intelligence techniques for education have been received increasing attentions, while it still remains an open problem to design the effective music instrument instructing systems. Although key presses can be directly derived from sheet music, the transitional movements among key presses require more extensive guidance in piano performance. In this work, we construct a piano-hand motion generation benchmark to guide hand movements and fingerings for piano playing. To this end, we collect an annotated dataset, PianoMotion10M, consisting of 116 hours of piano playing videos from a bird's-eye view with 10 million annotated hand poses. We also introduce a powerful baseline model that generates hand motions from piano audios through a position predictor and a position-guided gesture generator. Furthermore, a series of evaluation metrics are designed to assess the performance of the baseline model, including motion similarity, smoothness, positional accuracy of left and right hands, and overall fidelity of movement distribution. Despite that piano key presses with respect to music scores or audios are already accessible, PianoMotion10M aims to provide guidance on piano fingering for instruction purposes. The source code and dataset can be accessed at https://github.com/agnJason/PianoMotion10M.

****

Mihaela C. Stoian,Eleonora Giunchiglia

Beyond the convexity assumption: Realistic tabular data generation under quantifier-free real linear constraints

Synthetic tabular data generation has traditionally been a challenging problem due to the high complexity of the underlying distributions that characterise this type of data. Despite recent advances in deep generative models (DGMs), existing methods often fail to produce realistic datapoints that are well-aligned with available background knowledge.

In this paper, we address this limitation by introducing Disjunctive Refinement Layer (DRL), a novel layer designed

to enforce the alignment of generated data with the background knowledge specified in user-defined constraints.

DRL is the first method able to automatically make deep learning models inherently compliant with constraints as expressive as quantifier-free linear formulas, which can define non-convex and even disconnected spaces.

Our experimental analysis shows that DRL not only guarantees constraint satisfaction but also improves efficacy in downstream tasks. Notably, when applied to DGMs that frequently violate constraints, DRL eliminates violations entirely. Further, it improves performance metrics by up to 21.4\% in F1-score and 20.9\% in Area Under the ROC Curve, thus demonstrating its practical impact on data generation.

****
Junsol Kim,James Evans,Aaron Schein
Linear Representations of Political Perspective Emerge in Large Language Models
Large language models (LLMs) have demonstrated the ability to generate text that realistically reflects a range of different subjective human perspectives. This paper studies how LLMs are seemingly able to reflect more liberal versus more conservative viewpoints among other political perspectives in American politics. We show that LLMs possess linear representations of political perspectives within activation space, wherein more similar perspectives are represented closer together. To do so, we probe the attention heads across the layers of three open transformer-based LLMs (\texttt{Llama-2-7b-chat}, \texttt{Mistral-7b-instruct}, \texttt{Vicuna-7b}). We first prompt models to generate text from the perspectives of different U.S.~lawmakers. We then identify sets of attention heads whose activations linearly predict those lawmakers' DW-NOMINATE scores, a widely-used and validated measure of political ideology. We find that highly predictive heads are primarily located in the middle layers, often speculated to encode high-level concepts and tasks. Using probes only trained to predict lawmakers' ideology, we then show that the same probes can predict measures of news outlets' slant from the activations of models prompted to simulate text from those news outlets. These linear probes allow us to visualize, interpret, and monitor ideological stances implicitly adopted by an LLM as it generates open-ended responses. Finally, we demonstrate that by applying linear interventions to these attention heads, we can steer the model outputs toward a more liberal or conservative stance. Overall, our research suggests that LLMs possess a high-level linear representation of American political ideology and that by leveraging recent advances in mechanistic interpretability, we can identify, monitor, and steer the subjective perspective underlying generated text.
****
Yalun Dai,Lingao Xiao,Ivor Tsang,Yang He
Training-Free Dataset Pruning for Instance Segmentation
Existing dataset pruning techniques primarily focus on classification tasks, limiting their applicability to more complex and practical tasks like instance segmentation. Instance segmentation presents three key challenges: pixel-level annotations, instance area variations, and class imbalances, which significantly complicate dataset pruning efforts. Directly adapting existing classification-based pruning methods proves ineffective due to their reliance on time-consuming model training process. To address this, we propose a novel **T**raining-**F**ree **D**ataset **P**runing (**TFDP**) method for instance segmentation. Specifically, we leverage shape and class information from image annotations to design a Shape Complexity Score (SCS), refining it into a Scale-Invariant (SI-SCS) and Class-Balanced (CB-SCS) versions to address instance area variations and class imbalances, all without requiring model training. We achieve state-of-the-art results on VOC 2012, Cityscapes, and MS COCO datasets, generalizing well across CNN and Transformer architectures. Remarkably, our approach accelerates the pruning process by an average of **1349$\times$** on COCO compared to the adapted baselines.
****
Julian Dörfler,Benito van der Zander,Markus Bläser,Maciej Liskiewicz
From Probability to Counterfactuals: the Increasing Complexity of Satisfiability in Pearl's Causal Hierarchy
The framework of Pearl's Causal Hierarchy (PCH) formalizes three types of reasoning: probabilistic (i.e. purely observational), interventional, and counterfactual, that reflect the progressive sophistication of human thought regarding causation. We investigate the computational complexity aspects of reasoning in this framework focusing mainly on satisfiability problems expressed in probabilistic and causal languages across the PCH. That is, given a system of formulas in the standard probabilistic and causal languages, does there exist a model satisfying the formulas?

Our main contribution is to prove the exact computational complexities showing that languages allowing addition and marginalization (via the summation operator)

yield NP^{PP}-, PSPACE-, and NEXP-complete satisfiability problems, depending on the level of the PCH. These are the first results to demonstrate a strictly increasing complexity across the PCH: from probabilistic to causal and counterfactual reasoning. On the other hand, in the case of full languages, i.e.~allowing addition, marginalization, and multiplication, we show that the satisfiability for the counterfactual level remains the same as for the probabilistic and causal levels, solving an open problem in the field.
****
Zhao Song,Mingquan Ye,Junze Yin,Lichen Zhang
Efficient Alternating Minimization with Applications to Weighted Low Rank Approximation
Weighted low rank approximation is a fundamental problem in numerical linear algebra, and it has many applications in machine learning. Given a matrix $M \in \mathbb{R}^{n \times n}$, a non-negative weight matrix $W \in \mathbb{R}_{\geq 0}^{n \times n}$, a parameter k, the goal is to output two matrices $X,Y\in \mathbb{R}^{n \times k}$ such that $\\| W \circ (M - X Y^\top) \\|_F$ is minimized, where $\circ$ denotes the Hadamard product. It naturally generalizes the well-studied low rank matrix completion problem. Such a problem is known to be NP-hard and even hard to approximate assuming the Exponential Time Hypothesis. Meanwhile, alternating minimization is a good heuristic solution for weighted low rank approximation. In particular, [Li, Liang and Risteski, ICML'16] shows that, under mild assumptions, alternating minimization does provide provable guarantees. In this work, we develop an efficient and robust framework for alternating minimization that allows the alternating updates to be computed approximately. For weighted low rank approximation, this improves the runtime of [Li, Liang and Risteski, ICML'16] from $\\|W\\|_0k^2$ to $\\|W\\|_0 k$ where $\\|W\\|_0$ denotes the number of nonzero entries of the weight matrix. At the heart of our framework is a high-accuracy multiple response regression solver together with a robust analysis of alternating minimization.
****
Markel Zubia,Thiago D. Simão,Nils Jansen
Robust Transfer of Safety-Constrained Reinforcement Learning Agents
Reinforcement learning (RL) often relies on trial and error, which may cause undesirable outcomes. As a result, standard RL is inappropriate for safety-critical applications. To address this issue, one may train a safe agent in a controlled environment (where safety violations are allowed) and then transfer it to the real world (where safety violations may have disastrous consequences). Prior work has made this transfer safe as long as the new environment preserves the safety-related dynamics. However, in most practical applications, differences or shifts in dynamics between the two environments are inevitable, potentially leading to safety violations after the transfer. This work aims to guarantee safety even when the new environment has different (safety-related) dynamics. In other words, we aim to make the process of safe transfer robust. Our methodology (1) robustifies an agent in the controlled environment and (2) provably provides---under mild assumption---a safe transfer to new environments. The empirical evaluation shows that this method yields policies that are robust against changes in dynamics, demonstrating safety after transfer to a new environment.
****
Xu Liu,Haobo Fu,Stefano V Albrecht,QIANG FU,Shuai Li
Online-to-Offline RL for Agent Alignment
Reinforcement learning (RL) has shown remarkable success in training agents to achieve high-performing policies, particularly in domains like Game AI where simulation environments enable efficient interactions. However, despite their success in maximizing these returns, such online-trained policies often fail to align with human preferences concerning actions, styles, and values. The challenge lies in efficiently adapting these online-trained policies to align with human preferences, given the scarcity and high cost of collecting human behavior data. In this work, we formalize the problem as *online-to-offline* RL and propose ALIGNment of Game AI to Preferences (ALIGN-GAP), an innovative approach for the alignment of well-trained game agents to human preferences. Our method features a care

fully designed reward model that encodes human preferences from limited offline data and incorporates curriculum-based preference learning to align RL agents with targeted human preferences. Experiments across diverse environments and preference types demonstrate the performance of ALIGN-GAP, achieving effective alignment with human preferences.
****
Fushuo Huo,Wenchao Xu,Zhong Zhang,Haozhao Wang,Zhicheng Chen,Peilin Zhao
Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models
Hallucination remains a significant challenge in Large Vision-Language Models (LVLMs). To alleviate this issue, some methods, known as contrastive decoding, induce hallucinations by manually disturbing the raw vision or instruction inputs and then mitigate them by contrasting the outputs of the original and disturbed LVLMs. However, these holistic input disturbances sometimes induce potential noise and also double the inference cost. To tackle these issues, we propose a simple yet effective method named $\textit{Self-Introspective Decoding}$ (SID). Our empirical investigations reveal that pre-trained LVLMs can introspectively assess the importance of vision tokens based on preceding vision and text (both instruction and generated) tokens. Leveraging this insight, we develop the Context and Text-aware Token Selection (CT^{2}S) strategy, which preserves only the least important vision tokens after the early decoder layers, thereby adaptively amplify vision-and-text association hallucinations during auto-regressive decoding. This strategy ensures that multimodal knowledge absorbed in the early decoder layers induces multimodal contextual rather than aimless hallucinations, and significantly reduces computation burdens. Subsequently, the original token logits subtract the amplified fine-grained hallucinations, effectively alleviating hallucinations without compromising the LVLMs' general ability. Extensive experiments illustrate SID generates less-hallucination and higher-quality texts across various metrics, without much additional computation cost.
****
Aram Davtyan,Leello Tadesse Dadi,Volkan Cevher,Paolo Favaro
Faster Inference of Flow-Based Generative Models via Improved Data-Noise Coupling
Conditional Flow Matching (CFM), a simulation-free method for training continuous normalizing flows, provides an efficient alternative to diffusion models for key tasks like image and video generation. The performance of CFM in solving these tasks depends on the way data is coupled with noise. A recent approach uses minibatch optimal transport (OT) to reassign noise-data pairs in each training step to streamline sampling trajectories and thus accelerate inference. However, its optimization is restricted to individual minibatches, limiting its effectiveness on large datasets. To address this shortcoming, we introduce LOOM-CFM (Looking Out Of Minibatch-CFM), a novel method to extend the scope of minibatch OT by preserving and optimizing these assignments across minibatches over training time. Our approach demonstrates consistent improvements in the sampling speed-quality trade-off across multiple datasets. LOOM-CFM also enhances distillation initialization and supports high-resolution synthesis in latent space training.
****
Peiwen Yuan,Shaoxiong Feng,Yiwei Li,Xinglin Wang,Yueqi Zhang,Jiayi Shi,Chuyi Tan,Boyuan Pan,Yao Hu,Kan Li
UniCBE: An Uniformity-driven Comparing Based Evaluation Framework with Unified Multi-Objective Optimization
Human preference plays a significant role in measuring large language models and guiding them to align with human values. Unfortunately, current comparing-based evaluation (CBE) methods typically focus on a single optimization objective, failing to effectively utilize scarce yet valuable preference signals. To address this, we delve into key factors that can enhance the accuracy, convergence, and scalability of CBE: suppressing sampling bias, balancing descending process of uncertainty, and mitigating updating uncertainty.
Following the derived guidelines, we propose UniCBE, a unified uniformity-driven CBE framework which simultaneously optimize these core objectives by constructi

ng and integrating three decoupled sampling probability matrices, each designed to ensure uniformity in specific aspects. We further ablate the optimal tuple sampling and preference aggregation strategies to achieve efficient CBE.
On the AlpacaEval benchmark, UniCBE saves over 17% of evaluation budgets while achieving a Pearson correlation with ground truth exceeding 0.995, demonstrating excellent accuracy and convergence. In scenarios where new models are continuously introduced, UniCBE can even save over 50% of evaluation costs, highlighting its improved scalability.
****
Hojae Han,seung-won hwang,Rajhans Samdani,Yuxiong He
ConvCodeWorld: Benchmarking Conversational Code Generation in Reproducible Feedback Environments
Large language models (LLMs) have proven invaluable for code generation, particularly in interactive settings. However, existing code generation benchmarks fail to capture the diverse feedback encountered in multi-turn interactions, limiting our ability to evaluate LLMs in these contexts. To address this gap, we present a set of novel benchmarks that explicitly model the quality of feedback provided to code generation LLMs. Our contributions are threefold: **First**, we introduce CONVCODEWORLD, a novel and reproducible environment for benchmarking interactive code generation. CONVCODEWORLD simulates 9 distinct interactive code generation scenarios while systematically combining three types of feedback: (a) compilation feedback; (b) execution feedback with varying test coverage; (c) verbal feedback generated by GPT-4o with different levels of expertise. **Second**, we introduce CONVCODEBENCH, a fast, static version of benchmark that uses pre-generated feedback logs, eliminating the need for costly dynamic verbal feedback generation while maintaining
strong Spearman’s rank correlations (0.82 to 0.99) with CONVCODEWORLD. **Third**, extensive evaluations of both closed-source and open-source LLMs including R1-Distill on CONVCODEWORLD reveal key insights: (a) LLM performance varies significantly based on the feedback provided; (b) Weaker LLMs, with sufficient feedback, can outperform single-turn results of state-of-the-art LLMs without feedback; (c) Training on a specific feedback combination can limit an LLM’s ability to utilize unseen combinations; (d) LLMs solve problems in fewer turns (high MRR) may not solve as many problems overall (high Recall), and vice versa. All implementations and benchmarks will be made publicly available at https://huggingface.co/spaces/ConvCodeWorld/ConvCodeWorld
****
Shunyu Yao,Noah Shinn,Pedram Razavi,Karthik R Narasimhan
{τ}-bench: A Benchmark for \underline{T}ool-\underline{A}gent-\underline{User} Interaction in Real-World Domains
Existing benchmarks for language agents do not set them up to interact with human users or follow domain-specific rules, both of which are vital to safe and realistic deployment. We propose τ-bench, a benchmark with two domains (retail and airline) emulating dynamic conversations between a user (simulated by language models) and a customer service agent provided with domain-specific API tools and policy guidelines. We employ a efficient and faithful evaluation process that compares the database state at the end of a conversation with the annotated goal state, and propose a new metric (pass^k) to evaluate the reliability of agent behavior over multiple trials. Our experiments show that even state-of-the-art function calling agents (gpt-4o) succeed on $<50\%$ of the tasks, and are terribly inconsistent (pass^8 < 25\% in retail). Our findings point to the need for methods that can improve the ability of agents to act consistently and reliably follow rules.
****
Yuliang Xu,Jian Kang
Bayesian Image Regression with Soft-thresholded Conditional Autoregressive Prior
In the analysis of brain functional MRI (fMRI) data using regression models, Bayesian methods are highly valued for their flexibility and ability to quantify uncertainty. However, these methods face computational challenges in high-dimensional settings typical of brain imaging, and the often pre-specified correlations

tructures may not accurately capture the true spatial relationships within the b
rain. To address these issues, we develop a general prior specifically designed
for regression models with large-scale imaging data. We introduce the Soft-Thres
holded Conditional AutoRegressive (ST-CAR) prior, which reduces instability to p
re-fixed correlation structures and provides inclusion probabilities to account
for the uncertainty in choosing active voxels in the brain. We apply the ST-CAR
prior to scalar-on-image (SonI) and image-on-scalar (IonS) regression models—bot
h critical in brain imaging studies—and develop efficient computational algorith
ms using variational inference (VI) and stochastic subsampling techniques. Simul
ation studies demonstrate that the ST-CAR prior outperforms existing methods in
identifying active brain regions with complex correlation patterns, while our VI
 algorithms offer superior computational performance. We further validate our ap
proach by applying the ST-CAR to working memory fMRI data from the Adolescent Br
ain Cognitive Development (ABCD) study, highlighting its effectiveness in practi
cal brain imaging applications.
****

Manohar Kaul,Aditya Saibewar,Sadbhavana Babar
Beyond Mere Token Analysis: A Hypergraph Metric Space Framework for Defending Ag
ainst Socially Engineered LLM Attacks
Recent jailbreak attempts on Large Language Models (LLMs) have shifted from algo
rithm-focused to human-like social engineering attacks, with persuasion-based te
chniques emerging as a particularly effective subset. These attacks evolve rapid
ly, demonstrate high creativity, and boast superior attack success rates. To com
bat such threats, we propose a promising approach to enhancing LLM safety by lev
eraging the underlying geometry of input prompt token embeddings using hypergrap
hs. This approach allows us to model the differences in information flow between
 benign and malicious LLM prompts.

In our approach, each LLM prompt is represented as a metric hypergraph, forming
a compact metric space. We then construct a higher-order metric space over these
 compact metric hypergraphs using the Gromov-Hausdorff distance as a generalized
 metric. Within this space of metric hypergraph spaces, our safety filter learns
 to classify between harmful and benign prompts. Our study presents theoretical
guarantees on the classifier's generalization error for novel and unseen LLM inp
ut prompts. Extensive empirical evaluations demonstrate that our method signific
antly outperforms both existing state-of-the-art generic defense mechanisms and
naive baselines. Notably, our approach also achieves comparable performance to s
pecialized defenses against algorithm-focused attacks.
****

Junda Wu,Xintong Li,Ruoyu Wang,Yu Xia,Yuxin Xiong,Jianing Wang,Tong Yu,Xiang Che
n,Branislav Kveton,Lina Yao,Jingbo Shang,Julian McAuley
OCEAN: Offline Chain-of-thought Evaluation and Alignment in Large Language Model
s
Offline evaluation of LLMs is crucial in understanding their capacities, though
current methods remain underexplored in existing research. In this work, we focu
s on the offline evaluation of the chain-of-thought capabilities and show how to
 optimize LLMs based on the proposed evaluation method. To enable offline feedba
ck with rich knowledge and reasoning paths, we use knowledge graphs (KGs) (e.g.,
 Wikidata5M) to provide feedback on the generated chain of thoughts. Due to the
heterogeneity between LLM reasoning and KG structures, direct interaction and fe
edback from knowledge graphs on LLM behavior are challenging, as they require ac
curate entity linking and grounding of LLM-generated chains of thought in the KG
. To address the above challenge, we propose an offline chain-of-thought evaluat
ion framework, OCEAN, which models chain-of-thought reasoning in LLMs as a Marko
v Decision Process (MDP), and evaluate the policy’s alignment with KG preference
 modeling. To overcome the reasoning heterogeneity and grounding problems, we le
verage on-policy KG exploration and reinforcement learning to model a KG policy
that generates token-level likelihood distributions for LLM-generated chain-of-t
hought reasoning paths, simulating KG reasoning preference. Then we incorporate
the knowledge-graph feedback on the validity and alignment of the generated reas

oning paths into inverse propensity scores and propose KG-IPS estimator. Theoret
ically, we prove the unbiasedness of the proposed KG-IPS estimator and provide a
 lower bound on its variance. With the off-policy evaluated value function, we c
an directly enable off-policy optimization to further enhance chain-of-thought a
lignment. Our empirical study shows that OCEAN can be efficiently optimized for
generating chain-of-thought reasoning paths with higher estimated values without
 affecting LLMs’ general abilities in downstream tasks or their internal knowled
ge.
****
Jiahao Cui,Hui Li,Yao Yao,Hao Zhu,Hanlin Shang,Kaihui Cheng,Hang Zhou,Siyu Zhu,J
ingdong Wang
Hallo2: Long-Duration and High-Resolution Audio-Driven Portrait Image Animation
Recent advances in latent diffusion-based generative models for portrait image a
nimation, such as Hallo, have achieved impressive results in short-duration vide
o synthesis. In this paper, we present updates to Hallo, introducing several des
ign enhancements to extend its capabilities.First, we extend the method to produ
ce long-duration videos. To address substantial challenges such as appearance dr
ift and temporal artifacts, we investigate augmentation strategies within the im
age space of conditional motion frames. Specifically, we introduce a patch-drop
technique augmented with Gaussian noise to enhance visual consistency and tempor
al coherence over long duration.Second, we achieve 4K resolution portrait video
generation. To accomplish this, we implement vector quantization of latent codes
 and apply temporal alignment techniques to maintain coherence across the tempor
al dimension. By integrating a high-quality decoder, we realize visual synthesis
 at 4K resolution.Third, we incorporate adjustable semantic textual labels for p
ortrait expressions as conditional inputs. This extends beyond traditional audio
 cues to improve controllability and increase the diversity of the generated con
tent. To the best of our knowledge, Hallo2, proposed in this paper, is the first
 method to achieve 4K resolution and generate hour-long, audio-driven portrait i
mage animations enhanced with textual prompts. We have conducted extensive exper
iments to evaluate our method on publicly available datasets, including HDTF, Ce
lebV, and our introduced ''Wild'' dataset. The experimental results demonstrate
that our approach achieves state-of-the-art performance in long-duration portrai
t video animation, successfully generating rich and controllable content at 4K r
esolution for duration extending up to tens of minutes.
****
Xikun ZHANG,Dongjin Song,Yushan Jiang,Yixin Chen,Dacheng Tao
Learning system dynamics without forgetting
Observation-based trajectory prediction for systems with unknown dynamics is ess
ential in fields such as physics and biology. Most existing approaches are limit
ed to learning within a single system with fixed dynamics patterns. However, man
y real-world applications require learning across systems with evolving dynamics
 patterns, a challenge that has been largely overlooked. To address this, we sys
tematically investigate the problem of Continual Dynamics Learning (CDL), examin
ing task configurations and evaluating the applicability of existing techniques,
 while identifying key challenges. In response, we propose the Mode-switching Gr
aph ODE (MS-GODE) model, which integrates the strengths LG-ODE and sub-network l
earning with a mode-switching module, enabling efficient learning over varying d
ynamics. Moreover, we construct a novel benchmark of biological dynamic systems
for CDL, Bio-CDL, featuring diverse systems with disparate dynamics and signific
antly enriching the research field of machine learning for dynamic systems. Our
code available at \url{https://github.com/QueuQ/MS-GODE}.
****
Xiao Lin,Yuge Huang,Jianqing Xu,Yuxi Mi,Shuigeng Zhou,Shouhong Ding
UIFace: Unleashing Inherent Model Capabilities to Enhance Intra-Class Diversity
in Synthetic Face Recognition
Face recognition (FR) stands as one of the most crucial applications in computer
 vision. The accuracy of FR models has significantly improved in recent years du
e to the availability of large-scale human face datasets. However, directly usin
g these datasets can inevitably lead to privacy and legal problems. Generating s

ynthetic data to train FR models is a feasible solution to circumvent these issu
es. While existing synthetic-based face recognition methods have made significan
t progress in generating identity-preserving images, they are severely plagued b
y context overfitting, resulting in a lack of intra-class diversity of generated
 images and poor face recognition performance. In this paper, we propose a frame
work to $\textbf{U}$nleash model $\textbf{I}$nherent capabilities to enhance int
ra-class diversity for synthetic face recognition, shorted as $\textbf{UIFace}$.
 Our framework first train a diffusion model that can perform denoising conditio
ned on either identity contexts or a learnable empty context. The former generat
es identity-preserving images but lacks variations, while the latter exploits th
e model's intrinsic ability to synthesize intra-class-diversified images but wit
h random identities. Then we adopt a novel two-stage denoising strategy to fully
 leverage the strengths of both type of contexts, resulting in images that are d
iverse as well as identity-preserving. Moreover, an attention injection module i
s introduced to further augment the intra-class variations by utilizing attentio
n maps from the empty context to guide the denoising process in ID-conditioned g
eneration. Experiments show that our method significantly surpasses previous app
roaches with even less training data and half the size of synthetic dataset. Mor
e surprisingly, the proposed $\textbf{UIFace}$ even achieves comparable performa
nce of FR models trained on real datasets when we increase the number of synthet
ic identities.
****
John Yang,Carlos E Jimenez,Alex L Zhang,Kilian Lieret,Joyce Yang,Xindi Wu,Ori Pr
ess,Niklas Muennighoff,Gabriel Synnaeve,Karthik R Narasimhan,Diyi Yang,Sida Wang
,Ofir Press
SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains?
Autonomous systems for software engineering are now capable of fixing bugs and d
eveloping features. These systems are commonly evaluated on SWE-bench (Jimenez e
t al., 2024a), which assesses their ability to solve software issues from GitHub
 repositories. However, SWE-bench uses only Python repositories, with problem st
atements presented predominantly as text and lacking visual elements such as ima
ges. This limited coverage motivates our inquiry into how existing systems might
 perform on unrepresented software engineering domains
(e.g., front-end, game development, DevOps), which use different programming lan
guages and paradigms. Therefore, we propose SWE-bench Multimodal (SWE-bench M),
to evaluate systems on their ability to fix bugs in visual, user-facing JavaScri
pt software. SWE-bench M features 617 task instances collected from 17 JavaScrip
t libraries used for web interface design, diagramming, data visualization, synt
ax highlighting, and interactive mapping. Each SWE-bench M task instance contain
s at least one image in its problem statement or unit tests. Our analysis finds
that top-performing SWE-bench systems struggle with SWE-bench M, revealing limit
ations in visual problem-solving and cross-language generalization. Lastly, we s
how that SWE-agent’s flexible language-agnostic features enable it to substantia
lly outperform alternatives on SWE-bench M, resolving 12% of task
instances compared to 6% for the next best system.
****
Francesco Mori,Stefano Sarao Mannelli,Francesca Mignacco
Optimal Protocols for Continual Learning via Statistical Physics and Control The
ory
Artificial neural networks often struggle with _catastrophic forgetting_ when le
arning multiple tasks sequentially, as training on new tasks degrades the perfor
mance on previously learned tasks. Recent theoretical work has addressed this is
sue by analysing learning curves in synthetic frameworks under predefined traini
ng protocols. However, these protocols relied on heuristics and lacked a solid t
heoretical foundation assessing their optimality. In this paper, we fill this ga
p by combining exact equations for training dynamics, derived using statistical
physics techniques, with optimal control methods. We apply this approach to teac
her-student models for continual learning and multi-task problems, obtaining a t
heory for task-selection protocols maximising performance while minimising forge
tting. Our theoretical analysis offers non-trivial yet interpretable strategies

for mitigating catastrophic forgetting, shedding light on how optimal learning p
rotocols modulate established effects, such as the influence of task similarity
on forgetting. Finally, we validate our theoretical findings with experiments on
 real-world data.
****

Qingxuan Wu,Zhiyang Dou,Sirui Xu,Soshi Shimada,Chen Wang,Zhengming Yu,Yuan Liu,C
heng Lin,Zeyu Cao,Taku Komura,Vladislav Golyanik,Christian Theobalt,Wenping Wang
,Lingjie Liu
DICE: End-to-end Deformation Capture of Hand-Face Interactions from a Single Ima
ge
Reconstructing 3D hand-face interactions with deformations from a single image i
s a challenging yet crucial task with broad applications in AR, VR, and gaming.
The challenges stem from self-occlusions during single-view hand-face interactio
ns, diverse spatial relationships between hands and face, complex deformations,
and the ambiguity of the single-view setting. The previous state-of-the-art, Dec
af, employs a global fitting optimization guided by contact and deformation esti
mation networks trained on studio-collected data with 3D annotations. However, D
ecaf suffers from a time-consuming optimization process and limited generalizati
on capability due to its reliance on 3D annotations of hand-face interaction dat
a. To address these issues, we present DICE, the first end-to-end method for Def
ormation-aware hand-face Interaction reCovEry from a single image. DICE estimate
s the poses of hands and faces, contacts, and deformations simultaneously using
a Transformer-based architecture. It features disentangling the regression of lo
cal deformation fields and global mesh vertex locations into two network branche
s, enhancing deformation and contact estimation for precise and robust hand-face
 mesh recovery. To improve generalizability, we propose a weakly-supervised trai
ning approach that augments the training set using in-the-wild images without 3D
 ground-truth annotations, employing the depths of 2D keypoints estimated by off
-the-shelf models and adversarial priors of poses for supervision. Our experimen
ts demonstrate that DICE achieves state-of-the-art performance on a standard ben
chmark and in-the- wild data in terms of accuracy and physical plausibility. Add
itionally, our method operates at an interactive rate (20 fps) on an Nvidia 4090
 GPU, whereas Decaf requires more than 15 seconds for a single image. The code w
ill be available at: https://github.com/Qingxuan-Wu/DICE.
****

Jiachen Qian,Hongye Yang,Shuang Wu,Jingxi Xu,Feihu Zhang
High-quality Text-to-3D Character Generation with SparseCubes and Sparse Transfo
rmers.
Current state-of-the-art text-to-3D generation methods struggle to produce 3D mo
dels with fine details and delicate structures due to limitations in differentia
ble mesh representation techniques. This limitation is particularly pronounced i
n anime character generation, where intricate features such as fingers, hair, an
d facial details are crucial for capturing the essence of the characters.

In this paper, we introduce a novel, efficient, sparse differentiable mesh repre
sentation method, termed SparseCubes, alongside a sparse transformer network des
igned to generate high-quality 3D models. Our method significantly reduces compu
tational requirements by over 95% and storage memory by 50%, enabling the creati
on of higher resolution meshes with enhanced details and delicate structures. We
 validate the effectiveness of our approach through its application to text-to-3
D anime character generation, demonstrating its capability to accurately render
subtle details and thin structures (e.g. individual fingers) in both meshes and
textures.
****

Hao Sun,Yunyi Shen,Jean-Francois Ton
Rethinking Reward Modeling in Preference-based Large Language Model Alignment
The Bradley-Terry (BT) model is a common and successful practice in reward model
ing for Large Language Model (LLM) alignment. However, it remains unclear *why*
this model --- originally developed for multi-player stochastic game matching --
- can be adopted to convert pairwise response comparisons to reward values and m

ake predictions. Especially given the fact that only a limited number of prompt-response pairs are sparsely compared with others.
In this paper, we first establish the convergence rate of BT reward models based on deep neural networks using embeddings, providing a theoretical foundation for their use.
Despite theoretically sound, we argue that the BT model is not a necessary choice from the perspective of downstream optimization, this is because a reward model only needs to preserve the correct ranking predictions through a monotonic transformation of the true reward.
We highlight the critical concept of *order consistency* in reward modeling and demonstrate that the BT model possesses this property.
Moreover, we propose a simple and straightforward upper-bound algorithm, compatible with off-the-shelf binary classifiers, as an alternative order-consistent reward modeling objective.
To offer practical insights, we empirically evaluate the performance of these different reward modeling approaches across more than 12,000 experimental setups, using 6 base LLMs, 2 datasets, and diverse annotation designs that vary in quantity, quality, and pairing choices in preference annotations.
****
Yuchen Liang,Peizhong Ju,Yingbin Liang,Ness Shroff
Broadening Target Distributions for Accelerated Diffusion Models via a Novel Analysis Approach
Accelerated diffusion models hold the potential to significantly enhance the efficiency of standard diffusion processes. Theoretically, these models have been shown to achieve faster convergence rates than the standard $\mathcal O(1/\epsilon^2)$ rate of vanilla diffusion models, where ϵ denotes the target accuracy. However, current theoretical studies have established the acceleration advantage only for restrictive target distribution classes, such as those with smoothness conditions imposed along the entire sampling path or with bounded support. In this work, we significantly broaden the target distribution classes with a new accelerated stochastic DDPM sampler. In particular, we show that it achieves accelerated performance for three broad distribution classes not considered before. Our first class relies on the smoothness condition posed only to the target density q_0, which is far more relaxed than the existing smoothness conditions posed to all q_t along the entire sampling path. Our second class requires only a finite second moment condition, allowing for a much wider class of target distributions than the existing finite-support condition. Our third class is Gaussian mixture, for which our result establishes the first acceleration guarantee. Moreover, among accelerated DDPM type samplers, our results specialized for bounded-support distributions show an improved dependency on the data dimension d. Our analysis introduces a novel technique for establishing performance guarantees via constructing a tilting factor representation of the convergence error and utilizing Tweedie's formula to handle Taylor expansion terms. This new analytical framework may be of independent interest.
****
Jingchu Gai,Yiheng Du,Bohang Zhang,Haggai Maron,Liwei Wang
Homomorphism Expressivity of Spectral Invariant Graph Neural Networks
Graph spectra are an important class of structural features on graphs that have shown promising results in enhancing Graph Neural Networks (GNNs). Despite their widespread practical use, the theoretical understanding of the power of spectral invariants --- particularly their contribution to GNNs --- remains incomplete. In this paper, we address this fundamental question through the lens of homomorphism expressivity, providing a comprehensive and quantitative analysis of the expressive power of spectral invariants. Specifically, we prove that spectral invariant GNNs can homomorphism-count exactly a class of specific tree-like graphs which we refer to as \emph{parallel trees}. We highlight the significance of this result in various contexts, including establishing a quantitative expressiveness hierarchy across different architectural variants, offering insights into the impact of GNN depth, and understanding the subgraph counting capabilities of spectral invariant GNNs. In particular, our results significantly extend \citet{ar

vind2024hierarchy} and settle their open questions. Finally, we generalize our a
nalysis to higher-order GNNs and answer an open question raised by \citet{zhang2
024expressive}.
****

Hao-Chien Hsueh,Wen-Hsiao Peng,Ching-Chun Huang
Warm Diffusion: Recipe for Blur-Noise Mixture Diffusion Models
Diffusion probabilistic models have achieved remarkable success in generative ta
sks across diverse data types. While recent studies have explored alternative de
gradation processes beyond Gaussian noise, this paper bridges two key diffusion
paradigms: hot diffusion, which relies entirely on noise, and cold diffusion, wh
ich uses only blurring without noise. We argue that hot diffusion fails to explo
it the strong correlation between high-frequency image detail and low-frequency
structures, leading to random behaviors in the early steps of generation. Conver
sely, while cold diffusion leverages image correlations for prediction, it negle
cts the role of noise (randomness) in shaping the data manifold, resulting in ou
t-of-manifold issues and partially explaining its performance drop. To integrate
 both strengths, we propose Warm Diffusion, a unified Blur-Noise Mixture Diffusi
on Model (BNMD), to control blurring and noise jointly. Our divide-and-conquer s
trategy exploits the spectral dependency in images, simplifying score model esti
mation by disentangling the denoising and deblurring processes. We further analy
ze the Blur-to-Noise Ratio (BNR) using spectral analysis to investigate the trad
e-off between model learning dynamics and changes in the data manifold. Extensiv
e experiments across benchmarks validate the effectiveness of our approach for i
mage generation.
****

Hansi Yang,James Kwok
Joint Gradient Balancing for Data Ordering in Finite-Sum Multi-Objective Optimiz
ation
In finite-sum optimization problems, the sample orders for parameter updates can
 significantly influence the convergence rate of optimization algorithms. While
numerous sample ordering techniques have been proposed in the context of single-
objective optimization, the problem of sample ordering in finite-sum multi-objec
tive optimization has not been thoroughly explored. To address this gap, we prop
ose a sample ordering method called JoGBa, which finds the sample orders for mul
tiple objectives by jointly performing online vector balancing on the gradients
of all objectives. Our theoretical analysis demonstrates that this approach outp
erforms the standard baseline of random ordering and accelerates the convergence
 rate for the MGDA algorithm. Empirical evaluation across various datasets with
different multi-objective optimization algorithms further demonstrates that JoGB
a can achieve faster convergence and superior final performance than other data
ordering strategies.
****

Wenji Fang,Shang Liu,Jing Wang,Zhiyao Xie
CircuitFusion: Multimodal Circuit Representation Learning for Agile Chip Design
The rapid advancements of AI rely on the support of integrated circuits (ICs). H
owever, the growing complexity of digital ICs makes the traditional IC design pr
ocess costly and time-consuming. In recent years, AI-assisted IC design methods
have demonstrated great potential, but most methods are task-specific or focus s
olely on the circuit structure in graph format, overlooking other circuit modali
ties with rich functional information. In this paper, we introduce CircuitFusion
, the first multimodal and implementation-aware circuit encoder. It encodes circ
uits into general representations that support different downstream circuit desi
gn tasks. To learn from circuits, we propose to fuse three circuit modalities: h
ardware code, structural graph, and functionality summary. More importantly, we
identify four unique properties of circuits: parallel execution, functional equi
valent transformation, multiple design stages, and circuit reusability. Based on
 these properties, we propose new strategies for both the development and applic
ation of CircuitFusion: 1) During circuit preprocessing, utilizing the parallel
nature of circuits, we split each circuit into multiple sub-circuits based on se
quential-element boundaries, each sub-circuit in three modalities. It enables fi

ne-grained encoding at the sub-circuit level. 2) During CircuitFusion pre-training, we introduce three self-supervised tasks that utilize equivalent transformations both within and across modalities. We further utilize the multi-stage property of circuits to align representation with ultimate circuit implementation. 3) When applying CircuitFusion to downstream tasks, we propose a new retrieval-augmented inference method, which retrieves similar known circuits as a reference for predictions. It improves fine-tuning performance and even enables zero-shot inference. Evaluated on five different circuit design tasks, CircuitFusion consistently outperforms the state-of-the-art supervised method specifically developed for every single task, demonstrating its generalizability and ability to learn circuits' inherent properties.
****
Zihan Zhang,Xiangyang Ji,Yuan Zhou
Almost Optimal Batch-Regret Tradeoff for Batch Linear Contextual Bandits
We study the optimal batch-regret tradeoff for batch linear contextual bandits. For this problem, we design batch learning algorithms and prove that they achieve the optimal regret bounds (up to logarithmic factors) for any batch number M, number of actions K, time horizon T, and dimension d. Therefore, we establish the \emph{full-parameter-range} (almost) optimal batch-regret tradeoff for the batch linear contextual bandit problem.

Along our analysis, we also prove a new matrix concentration inequality with dependence on their dynamic upper bounds, which, to the best of our knowledge, is the first of its kind in literature and maybe of independent interest.
****
Jeremy Vonderfecht,Feng Liu
Lossy Compression with Pretrained Diffusion Models
We apply Theis et al. (2022)'s DiffC algorithm to Stable Diffusion 1.5, 2.1, XL, and and Flux-dev, and demonstrate that these pretrained models are remarkably capable lossy image compressors. A principled algorithm for compression using pretrained diffusion models has been understood since at least 2020 (Ho et al.), but challenges in reverse-channel coding have prevented such algorithms from ever being fully implemented. We introduce simple workarounds that lead to the first complete implementation of DiffC, which is capable of compressing and decompressing images using Stable Diffusion in under 10 seconds. Despite requiring no additional training, our method is competitive with other state-of-the-art generative compression methods at low ultra-low bitrates.
****
Rachel Teo,Tan Minh Nguyen
MoLEx: Mixture of Layer Experts for Fine-tuning with Sparse Upcycling
Large-scale pre-training of deep models, followed by fine-tuning them to adapt to downstream tasks, has become the cornerstone of natural language processing (NLP). The prevalence of vast corpses of data coupled with computational resources has led to large models with a considerable number of parameters. While the massive size of these models has led to remarkable success in many NLP tasks, a detriment is the expense required to retrain all the base model's parameters for the adaptation to each task or domain. Parameter Efficient Fine-Tuning (PEFT) provides a highly effective solution for this challenge by minimizing the number of parameters required to be trained in adjusting to the new task while maintaining the quality of the model. While existing methods have achieved impressive results, they mainly focus on adapting a subset of parameters using adapters, weight reparameterization, and prompt engineering. In this paper, we study layers as extractors of different types of linguistic information that are valuable when used in conjunction with each other. We then propose the Mixture of Layer Experts (MoLEx), a novel Sparse Mixture of Experts (SMoE) whose experts are layers in the pre-trained model. In particular, MoLEx is applied at each layer of the pre-trained model. It performs a conditional computation of a mixture of layers during fine-tuning to provide the model with more structural knowledge about the data. By providing an avenue for information exchange between layers, MoLEx enables the model to make a more well-informed prediction for the downstream task, leading

to better fine-tuning results with the same number of effective parameters. As experts can be processed in parallel, MoLEx introduces minimal additional computational overhead. We empirically corroborate the advantages of MoLEx when combined with popular PEFT baseline methods on a variety of downstream fine-tuning tasks, including the popular GLUE benchmark for natural language understanding (NLU) as well as the natural language generation (NLG) End-to-End Challenge (E2E).
****

Eric Lei,Hsiang Hsu,Chun-Fu Chen

PaLD: Detection of Text Partially Written by Large Language Models

Advances in large language models (LLM) have produced text that appears increasingly human-like and difficult to detect with the human eye. In order to mitigate the impact of misusing LLM-generated texts, e.g., copyright infringement, fair student assessment, fraud, and other societally harmful LLM usage, a line of work on detecting human and LLM-written text has been explored. While recent work has focused on classifying entire text samples (e.g., paragraphs) as human or LLM-written, this paper investigates a more realistic setting of mixed-text, where the text's individual segments (e.g., sentences) could each be written by either a human or an LLM. A text encountered in practical usage cannot generally be assumed to be fully human or fully LLM-written; simply predicting whether it is human or LLM-written is insufficient as it does not provide the user with full context on its origins, such as the amount of LLM-written text, or locating the LLM-written parts. Therefore, we study two relevant problems in the mixed-text setting: (i) estimating the percentage of a text that was LLM-written, and (ii) determining which segments were LLM-written. To this end, we propose Partial-LLM Detector (PaLD), a black-box method that leverages the scores of text classifiers. Experimentally, we demonstrate the effectiveness of PaLD compared to baseline methods that build on existing LLM text detectors.
****

Xiang Cheng,Lawrence Carin,Suvrit Sra

Graph Transformers Dream of Electric Flow

We show theoretically and empirically that the linear Transformer, when applied to graph data, can implement algorithms that solve canonical problems such as electric flow and eigenvector decomposition. The Transformer has access to information on the input graph only via the graph's incidence matrix. We present explicit weight configurations for implementing each algorithm, and we bound the constructed Transformers' errors by the errors of the underlying algorithms. Our theoretical findings are corroborated by experiments on synthetic data. Additionally, on a real-world molecular regression task, we observe that the linear Transformer is capable of learning a more effective positional encoding than the default one based on Laplacian eigenvectors. Our work is an initial step towards elucidating the inner-workings of the Transformer for graph data. Code is available at https://github.com/chengxiang/LinearGraphTransformer
****

Senmao Li,Kai Wang,Joost van de Weijer,Fahad Shahbaz Khan,Chun-Le Guo,Shiqi Yang,Yaxing Wang,jian Yang,Ming-Ming Cheng

$InterLCM$: Low-Quality Images as Intermediate States of Latent Consistency Models for Effective Blind Face Restoration

Diffusion priors have been used for blind face restoration (BFR) by fine-tuning diffusion models (DMs) on restoration datasets to recover low-quality images. However, the naive application of DMs presents several key limitations.
(i) The diffusion prior has inferior semantic consistency (e.g., ID, structure and color.), increasing the difficulty of optimizing the BFR model;
(ii) reliance on hundreds of denoising iterations, preventing the effective cooperation with perceptual losses, which is crucial for faithful restoration.
Observing that the latent consistency model (LCM) learns consistency noise-to-data mappings on the ODE-trajectory and therefore shows more semantic consistency in the subject identity, structural information and color preservation,
we propose $\textit{InterLCM}$ to leverage the LCM for its superior semantic consistency and efficiency to counter the above issues.
Treating low-quality images as the intermediate state of LCM, $\textit{InterLCM}

$ achieves a balance between fidelity and quality by starting from earlier LCM s teps.
LCM also allows the integration of perceptual loss during training, leading to i mproved restoration quality, particularly in real-world scenarios.
To mitigate structural and semantic uncertainties, $\textit{InterLCM}$ incorpora tes a Visual Module to extract visual features and a Spatial Encoder to capture spatial details, enhancing the fidelity of restored images.
Extensive experiments demonstrate that $\textit{InterLCM}$ outperforms existing approaches in both synthetic and real-world datasets while also achieving faster inference speed. Code and models will be publicly available.
****

Xiangyu Wang,Donglin Yang,Ziqin Wang,Hohin Kwan,Jinyu Chen,Wenjun Wu,Hongsheng L i,Yue Liao,Si Liu
Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Metho dology
Developing agents capable of navigating to a target location based on language i nstructions and visual information, known as vision-language navigation (VLN), h as attracted widespread interest. Most research has focused on ground-based agen ts, while UAV-based VLN remains relatively underexplored. Recent efforts in UAV vision-language navigation predominantly adopt ground-based VLN settings, relyin g on predefined discrete action spaces and neglecting the inherent disparities i n agent movement dynamics and the complexity of navigation tasks between ground and aerial environments. To address these disparities and challenges, we propose solutions from three perspectives: platform, benchmark, and methodology. To ena ble realistic UAV trajectory simulation in VLN tasks, we propose the OpenUAV pla tform, which features diverse environments, realistic flight control, and exten sive algorithmic support. We further construct a target-oriented VLN dataset con sisting of approximately 12k trajectories on this platform, serving as the first dataset specifically designed for realistic UAV VLN tasks. To tackle the challe nges posed by complex aerial environments, we propose an assistant-guided UAV ob ject search benchmark called UAV-Need-Help, which provides varying levels of gui dance information to help UAVs better accomplish realistic VLN tasks. We also pr opose a UAV navigation LLM that, given multi-view images, task descriptions, and assistant instructions, leverages the multimodal understanding capabilities of the MLLM to jointly process visual and textual information, and performs hierarc hical trajectory generation. The evaluation results of our method significantly outperform the baseline models, while there remains a considerable gap between o ur results and those achieved by human operators, underscoring the challenge pre sented by the UAV-Need-Help task.
****

Tiberiu Mu■at
Mechanism and emergence of stacked attention heads in multi-layer transformers
In this paper, I introduce the retrieval problem, a simple yet common reasoning task that can be solved only by transformers with a minimum number of layers, wh ich grows logarithmically with the input size. I empirically show that large lan guage models can solve the task under different prompting formulations without a ny fine-tuning. To understand how transformers solve the retrieval problem, I tr ain several transformers on a minimal formulation. Successful learning occurs on ly under the presence of an implicit curriculum. I uncover the learned mechanism s by studying the attention maps in the trained transformers. I also study the t raining process, uncovering that attention heads always emerge in a specific seq uence guided by the implicit curriculum.
****

Nicolas Yax,Pierre-Yves Oudeyer,Stefano Palminteri
PhyloLM: Inferring the Phylogeny of Large Language Models and Predicting their P erformances in Benchmarks
This paper introduces PhyloLM, a method adapting phylogenetic algorithms to Larg e Language Models (LLMs) to explore whether and how they relate to each other an d to predict their performance characteristics. Our method calculates a phylogen etic distance metric based on the similarity of LLMs' output. The resulting metr

ic is then used to construct dendrograms, which satisfactorily capture known re lationships across a set of 111 open-source and 45 closed models. Furthermore, o ur phylogenetic distance predicts performance in standard benchmarks, thus demon strating its functional validity and paving the way for a time and cost-effecti ve estimation of LLM capabilities. To sum up, by translating population genetic concepts to machine learning, we propose and validate a tool to evaluate LLM dev elopment, relationships and capabilities, even in the absence of transparent tr aining information.
****

Song Wang,Junhong Lin,Xiaojie Guo,Julian Shun,Jundong Li,Yada Zhu
Reasoning of Large Language Models over Knowledge Graphs with Super-Relations
While large language models (LLMs) have made significant progress in processing and reasoning over knowledge graphs, current methods suffer from a high non-retr ieval rate. This limitation reduces the accuracy of answering questions based on these graphs. Our analysis reveals that the combination of greedy search and fo rward reasoning is a major contributor to this issue. To overcome these challeng es, we introduce the concept of super-relations, which enables both forward and backward reasoning by summarizing and connecting various relational paths within the graph. This holistic approach not only expands the search space, but also s ignificantly improves retrieval efficiency. In this paper, we propose the ReKnoS framework, which aims to Reason over Knowledge Graphs with Super-Relations. Our framework’s key advantages include the inclusion of multiple relation paths thr ough super-relations, enhanced forward and backward reasoning capabilities, and increased efficiency in querying LLMs. These enhancements collectively lead to a substantial improvement in the successful retrieval rate and overall reasoning performance. We conduct extensive experiments on a variety of datasets to evalua te ReKnoS, and the results demonstrate the superior performance of ReKnoS over e xisting state-of-the-art baselines, with an average accuracy gain of 2.92% acros s nine real-world datasets.
****

Ciyuan Peng,Yuelong Huang,Qichao Dong,Shuo Yu,Feng Xia,Chengqi Zhang,Yaochu Jin
Biologically Plausible Brain Graph Transformer
State-of-the-art brain graph analysis methods fail to fully encode the small-wor ld architecture of brain graphs (accompanied by the presence of hubs and functio nal modules), and therefore lack biological plausibility to some extent. This li mitation hinders their ability to accurately represent the brain's structural an d functional properties, thereby restricting the effectiveness of machine learni ng models in tasks such as brain disorder detection. In this work, we propose a novel Biologically Plausible Brain Graph Transformer (BioBGT) that encodes the s mall-world architecture inherent in brain graphs. Specifically, we present a net work entanglement-based node importance encoding technique that captures the str uctural importance of nodes in global information propagation during brain graph communication, highlighting the biological properties of the brain structure. F urthermore, we introduce a functional module-aware self-attention to preserve th e functional segregation and integration characteristics of brain graphs in the learned representations. Experimental results on three benchmark datasets demons trate that BioBGT outperforms state-of-the-art models, enhancing biologically pl ausible brain graph representations for various brain graph analytical tasks
****

Gang Liu,Michael Sun,Wojciech Matusik,Meng Jiang,Jie Chen
Multimodal Large Language Models for Inverse Molecular Design with Retrosyntheti c Planning
While large language models (LLMs) have integrated images, adapting them to grap hs remains challenging, limiting their applications in materials and drug design . This difficulty stems from the need for coherent autoregressive generation acr oss texts and graphs. To address this, we introduce Llamole, the first multimoda l LLM capable of interleaved text and graph generation, enabling molecular inver se design with retrosynthetic planning. Llamole integrates a base LLM with the G raph Diffusion Transformer and Graph Neural Networks for multi-conditional molec ular generation and reaction inference within texts, while the LLM, with enhance

d molecular understanding, flexibly controls activation among the different grap
h modules. Additionally, Llamole integrates A* search with LLM-based cost functi
ons for efficient retrosynthetic planning. We create benchmarking datasets and c
onduct extensive experiments to evaluate Llamole against in-context learning and
 supervised fine-tuning. Llamole significantly outperforms 14 adapted LLMs acros
s 12 metrics for controllable molecular design and retrosynthetic planning. Code
 and model at https://github.com/liugangcode/Llamole.
****

Jacob Lear,Lu Zhang

A Causal Lens for Learning Long-term Fair Policies

Fairness-aware learning studies the development of algorithms that avoid discrim
inatory decision outcomes despite biased training data. While most studies have
concentrated on immediate bias in static contexts, this paper highlights the imp
ortance of investigating long-term fairness in dynamic decision-making systems w
hile simultaneously considering instantaneous fairness requirements. In the cont
ext of reinforcement learning, we propose a general framework where long-term fa
irness is measured by the difference in the average expected qualification gain
that individuals from different groups could obtain. Then, through a causal lens
, we decompose this metric into three components that represent the direct impac
t, the delayed impact, as well as the spurious effect the policy has on the qual
ification gain. We analyze the intrinsic connection between these components and
 an emerging fairness notion called benefit fairness that aims to control the eq
uity of outcomes in decision-making. Finally, we develop a simple yet effective
approach for balancing various fairness notions.
****

Zirun Guo,Tao Jin

Smoothing the Shift: Towards Stable Test-Time Adaptation under Complex Multimoda
l Noises

Test-Time Adaptation (TTA) aims to tackle distribution shifts using unlabeled te
st data without access to the source data. In the context of multimodal data, th
ere are more complex noise patterns than unimodal data such as simultaneous corr
uptions for multiple modalities and missing modalities. Besides, in real-world a
pplications, corruptions from different distribution shifts are always mixed. Ex
isting TTA methods always fail in such multimodal scenario because the abrupt di
stribution shifts will destroy the prior knowledge from the source model, thus l
eading to performance degradation. To this end, we reveal a new challenge named
multimodal wild TTA. To address this challenging problem, we propose two novel
 strategies: sample identification with interquartile range **S**moothing and **
u**nimodal assistance, and **M**utual **i**nformation sharing (SuMi). SuMi smoot
hs the adaptation process by interquartile range which avoids the abrupt distrib
ution shifts. Then, SuMi fully utilizes the unimodal features to select low-entr
opy samples with rich multimodal information for optimization. Furthermore, mutu
al information sharing is introduced to align the information, reduce the discre
pancies and enhance the information utilization across different modalities. Ext
ensive experiments show the effectiveness and superiority over existing methods
under the complex noise patterns in multimodal data. Code is available at https:
//github.com/zrguo/SuMi.
****

Ziqi Pang,Xin Xu,Yu-Xiong Wang

Aligning Generative Denoising with Discriminative Objectives Unleashes Diffusion
 for Visual Perception

With success in image generation, generative diffusion models are increasingly a
dopted for discriminative scenarios because generating pixels is a unified and n
atural perception interface. Although directly re-purposing their generative den
oising process has established promising progress in specialist (e.g., depth est
imation) and generalist models, the inherent gaps between a generative process a
nd discriminative objectives are rarely investigated. For instance, generative m
odels can tolerate deviations at intermediate sampling steps as long as the fina
l distribution is reasonable, while discriminative tasks with rigorous ground tr
uth for evaluation are sensitive to such errors. Without mitigating such gaps, d

iffusion for perception still struggles on tasks represented by multi-modal unde
rstanding (e.g., referring image segmentation). Motivated by these challenges, w
e analyze and improve the alignment between the generative diffusion process and
 perception objectives centering around the key observation: how perception qual
ity evolves with the denoising process. (1) Notably, earlier denoising steps con
tribute more than later steps, necessitating a tailored learning objective for t
raining: loss functions should reflect varied contributions of timesteps for eac
h perception task. (2) Perception quality drops unexpectedly at later denoising
steps, revealing the sensitiveness of perception to training-denoising distribut
ion shift. We introduce diffusion-tailored data augmentation to simulate such dr
ift in the training data. (3) We suggest a novel perspective to the long-standin
g question: why should a generative process be useful for discriminative tasks -
 interactivity. The denoising process can be leveraged as a controllable user in
terface adapting to users' correctional prompts and conducting multi-round inter
action in an agentic workflow. Collectively, our insights enhance multiple gener
ative diffusion-based perception models without architectural changes: state-of-
the-art diffusion-based depth estimator, previously underplayed referring image
segmentation models, and perception generalists. Our code is available at https:
//github.com/ziqipang/ADDP.
****
Liping Wang,Xujia Li,Jingshu Peng,Yue Wang,Chen Zhang,Yan Zhou,Lei Chen
N-ForGOT: Towards Not-forgetting and Generalization of Open Temporal Graph Learn
ing
Temporal Graph Neural Networks (TGNNs) lay emphasis on capturing node interactio
ns over time but often overlook evolution in node classes and dynamic data distr
ibutions triggered by the continuous emergence of new class labels, known as the
 open-set problem. This problem poses challenges for existing TGNNs in preservin
g learned classes while rapidly adapting to new, unseen classes. To address this
, this paper identifies two primary factors affecting model performance on the o
pen temporal graph, backed by a theoretical guarantee: (1) the forgetting of pr
ior knowledge and (2) distribution discrepancies between successive tasks. Build
ing on these insights, we propose N-ForGOT, which incorporates two plug-in modul
es into TGNNs to preserve prior knowledge and enhance model generalizability for
 new classes simultaneously. The first module preserves previously established i
nter-class connectivity and decision boundaries during the training of new class
es to mitigate the forgetting caused by temporal evolutions of class characteris
tics. The second module introduces an efficient method for measuring distributio
n discrepancies with designed temporal Weisfeiler-Lehman subtree patterns, effec
tively addressing both structural and temporal shifts while reducing time comple
xity. Experimental results on four public datasets demonstrate that our method s
ignificantly outperforms state-of-the-art approaches in prediction accuracy, pre
vention of forgetting, and generalizability.
****
Yongqi An,Xu Zhao,Tao Yu,Ming Tang,Jinqiao Wang
Systematic Outliers in Large Language Models
Outliers have been widely observed in Large Language Models (LLMs), significantl
y impacting model performance and posing challenges for model compression. Under
standing the functionality and formation mechanisms of these outliers is critica
lly important. Existing works, however, largely focus on reducing the impact of
outliers from an algorithmic perspective, lacking an in-depth investigation into
 their causes and roles. In this work, we provide a detailed analysis of the for
mation process, underlying causes, and functions of outliers in LLMs. We define
and categorize three types of outliers—activation outliers, weight outliers, and
 attention outliers—and analyze their distributions across different dimensions,
 uncovering inherent connections between their occurrences and their ultimate in
fluence on the attention mechanism. Based on these observations, we hypothesize
and explore the mechanisms by which these outliers arise and function, demonstra
ting through theoretical derivations and experiments that they emerge due to the
 self-attention mechanism's softmax operation. These outliers act as implicit co
ntext-aware scaling factors within the attention mechanism. As these outliers st

em from systematic influences, we term them systematic outliers. Our study not o
nly enhances the understanding of Transformer-based LLMs but also shows that str
ucturally eliminating outliers can accelerate convergence and improve model comp
ression. The code is avilable at \url{https://github.com/an-yongqi/systematic-ou
tliers}.
****
Neel Alex,Shoaib Ahmed Siddiqui,Amartya Sanyal,David Krueger
Protecting against simultaneous data poisoning attacks
Current backdoor defense methods are evaluated against a single attack at a time
. This is unrealistic, as powerful machine learning systems are trained on large
 datasets scraped from the internet, which may be attacked multiple times by one
 or more attackers. We demonstrate that multiple backdoors can be simultaneously
 installed in a single model through parallel data poisoning attacks without sub
stantially degrading clean accuracy. Furthermore, we show that existing backdoor
 defense methods do not effectively defend against multiple simultaneous attacks
. Finally, we leverage insights into the nature of backdoor attacks to develop a
 new defense, BaDLoss (**Ba**ckdoor **D**etection via **Loss** Dynamics), that i
s effective in the multi-attack setting. With minimal clean accuracy degradation
, BaDLoss attains an average attack success rate in the multi-attack setting of
7.98% in CIFAR-10, 10.29% in GTSRB, and 19.17% in Imagenette, compared to the av
erage of other defenses at 63.44%, 74.83%, and 41.74% respectively. BaDLoss scal
es to ImageNet-1k, reducing the average attack success rate from 88.57% to 15.61
%.
****
Jongmin Lee,Meiqi Sun,Pieter Abbeel
SEMDICE: Off-policy State Entropy Maximization via Stationary Distribution Corre
ction Estimation
In the unsupervised pre-training for reinforcement learning, the agent aims to l
earn a prior policy for downstream tasks without relying on task-specific reward
 functions. We focus on state entropy maximization (SEM), where the goal is to l
earn a policy that maximizes the entropy of the state's stationary distribution.
 In this paper, we introduce SEMDICE, a principled off-policy algorithm that com
putes an SEM policy from an arbitrary off-policy dataset, which optimizes the po
licy directly within the space of stationary distributions. SEMDICE computes a s
ingle, stationary Markov state-entropy-maximizing policy from an arbitrary off-p
olicy dataset. Experimental results demonstrate that SEMDICE outperforms baselin
e algorithms in maximizing state entropy while achieving the best adaptation eff
iciency for downstream tasks among SEM-based unsupervised RL pre-training method
s.
****
Ji■í N■me■ek,Tomáš Pevný,Jakub Marecek
Generating Likely Counterfactuals Using Sum-Product Networks
The need to explain decisions made by AI systems is driven by both recent regula
tion and user demand. The decisions are often explainable only post hoc. In coun
terfactual explanations, one may ask what constitutes the best counterfactual ex
planation. Clearly, multiple criteria must be taken into account, although "dist
ance from the sample" is a key criterion. Recent methods that consider the plaus
ibility of a counterfactual seem to sacrifice this original objective. Here, we
present a system that provides high-likelihood explanations that are, at the sam
e time, close and sparse. We show that the search for the most likely explanatio
ns satisfying many common desiderata for counterfactual explanations can be mode
led using Mixed-Integer Optimization (MIO). We use a Sum-Product Network (SPN) t
o estimate the likelihood of a counterfactual. To achieve that, we propose an MI
O formulation of an SPN, which can be of independent interest. The source code w
ith examples is available at https://github.com/Epanemu/LiCE.
****
Chase Walker,Sumit Kumar Jha,Rickard Ewetz
Metric-Driven Attributions for Vision Transformers
Attribution algorithms explain computer vision models by attributing the model r
esponse to pixels within the input. Existing attribution methods generate explan

ations by combining transformations of internal model representations such as cl
ass activation maps, gradients, attention, or relevance scores. The effectivenes
s of an attribution map is measured using attribution quality metrics. This lead
s us to pose the following question: if attribution methods are assessed using a
ttribution quality metrics, why are the metrics not used to generate the attribu
tions? In response to this question, we propose a Metric-Driven Attribution for
explaining Vision Transformers (ViT) called MDA. Guided by attribution quality m
etrics, the method creates attribution maps by performing patch order and patch
magnitude optimization across all patch tokens. The first step orders the patche
s in terms of importance and the second step assigns the magnitude to each patch
 while preserving the patch order. Moreover, MDA can provide a smooth trade-off
between sparse and dense attributions by modifying the optimization objective. E
xperimental evaluation demonstrates the proposed MDA method outperforms 7 exis
ting ViT attribution methods by an average of 12% across 12 attribution met
rics on the ImageNet dataset for the ViT-base 16×16, ViT-tiny $16 \time
s 16$, and ViT-base 32×32 models. Code is publicly available at https:/
/github.com/chasewalker26/MDA-Metric-Driven-Attributions-for-ViT.
****

Yuxuan Sun,Yunlong Zhang,Yixuan Si,Chenglu Zhu,Kai Zhang,Zhongyi Shui,Jingxiong
Li,Xuan Gong,XINHENG LYU,Tao Lin,Lin Yang
PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-ag
ent Collaboration
Vision Language Models (VLMs) like CLIP have attracted substantial attention in
pathology, serving as backbones for applications such as zero-shot image classif
ication and Whole Slide Image (WSI) analysis. Additionally, they can function as
 vision encoders when combined with large language models (LLMs) to support broa
der capabilities. Current efforts to train pathology VLMs rely on pathology imag
e-text pairs from platforms like PubMed, YouTube, and Twitter, which provide lim
ited, unscalable data with generally suboptimal image quality. In this work, we
leverage large-scale WSI datasets like TCGA to extract numerous high-quality ima
ge patches. We then train a large multimodal model (LMM) to generate captions fo
r extracted images, creating PathGen-1.6M, a dataset containing 1.6 million high
-quality image-caption pairs. Our approach involves multiple agent models collab
orating to extract representative WSI patches, generating and refining captions
to obtain high-quality image-text pairs. Extensive experiments show that integra
ting these generated pairs with existing datasets to train a pathology-specific
CLIP model, PathGen-CLIP, significantly enhances its ability to analyze patholog
ical images, with substantial improvements across nine pathology-related zero-sh
ot image classification tasks and three whole-slide image tasks. Furthermore, we
 construct 200K instruction-tuning data based on PathGen-1.6M and integrate Path
Gen-CLIP with the Vicuna LLM to create more powerful multimodal models through i
nstruction tuning. Overall, we provide a scalable pathway for high-quality data
generation in pathology, paving the way for next-generation general pathology mo
dels. Our dataset, code, and model are open-access at https://github.com/PathFou
ndation/PathGen-1.6M.
****

Tongzhou Liao,Barnabas Poczos
Greener GRASS: Enhancing GNNs with Encoding, Rewiring, and Attention
Graph Neural Networks (GNNs) have become important tools for machine learning on
 graph-structured data. In this paper, we explore the synergistic combination of
 graph encoding, graph rewiring, and graph attention, by introducing Graph Atten
tion with Stochastic Structures (GRASS), a novel GNN architecture. GRASS utilize
s relative random walk probabilities (RRWP) encoding and a novel decomposed vari
ant (D-RRWP) to efficiently capture structural information. It rewires the input
 graph by superimposing a random regular graph to enhance long-range information
 propagation. It also employs a novel additive attention mechanism tailored for
graph-structured data. Our empirical evaluations demonstrate that GRASS achieves
 state-of-the-art performance on multiple benchmark datasets, including a 20.3%
reduction in mean absolute error on the ZINC dataset.
****

FENG SHIBO,Wanjin Feng,Xingyu Gao,Peilin Zhao,Zhiqi Shen
TS-LIF: A Temporal Segment Spiking Neuron Network for Time Series Forecasting
Spiking Neural Networks (SNNs) offer a promising, biologically inspired approach for processing spatiotemporal data, particularly for time series forecasting.
However, conventional neuron models like the Leaky Integrate-and-Fire (LIF) struggle to capture long-term dependencies and effectively process multi-scale temporal dynamics.
To overcome these limitations, we introduce the Temporal Segment Leaky Integrate-and-Fire (TS-LIF) model, featuring a novel dual-compartment architecture.
The dendritic and somatic compartments specialize in capturing distinct frequency components, providing functional heterogeneity that enhances the neuron's ability to process both low- and high-frequency information.
Furthermore, the newly introduced direct somatic current injection reduces information loss during intra-neuronal transmission, while dendritic spike generation improves multi-scale information extraction.
We provide a theoretical stability analysis of the TS-LIF model and explain how each compartment contributes to distinct frequency response characteristics.
Experimental results show that TS-LIF outperforms traditional SNNs in time series forecasting, demonstrating better accuracy and robustness, even with missing data.
TS-LIF advances the application of SNNs in time-series forecasting, providing a biologically inspired approach that captures complex temporal dynamics and offers potential for practical implementation in diverse forecasting scenarios.
****
Dongping Chen,Ruoxi Chen,Shu Pu,Zhaoyi Liu,Yanru Wu,Caixi Chen,Benlin Liu,Yue Huang,Yao Wan,Pan Zhou,Ranjay Krishna
Interleaved Scene Graphs for Interleaved Text-and-Image Generation Assessment
Many real-world user queries (e.g. *"How do to make egg fried rice?"*) could benefit from systems capable of generating responses with both textual steps with accompanying images, similar to a cookbook.
Models designed to generate interleaved text and images face challenges in ensuring consistency within and across these modalities.
To address these challenges, we present ISG, a comprehensive evaluation framework for interleaved text-and-image generation. ISG leverages a scene graph structure to capture relationships between text and image blocks, evaluating responses on four levels of granularity: holistic, structural, block-level, and image-specific. This multi-tiered evaluation allows for a nuanced assessment of consistency, coherence, and accuracy, and provides interpretable question-answer feedback.
In conjunction with ISG, we introduce a benchmark, ISG-Bench, encompassing 1,150 samples across 8 categories and 21 subcategories. This benchmark dataset includes complex language-vision dependencies and golden answers to evaluate models effectively on vision-centric tasks such as style transfer, a challenging area for current models.
Using ISG-Bench, we demonstrate that recent unified vision-language models perform poorly on generating interleaved content. While compositional approaches that combine separate language and image models show a 111% improvement over unified models at the holistic level, their performance remains suboptimal at both block and image levels.
To facilitate future work, we develop ISG-Agent, a baseline agent employing a *"plan-execute-refine"* pipeline to invoke tools, achieving a 122% performance improvement.
****
Zhilu Zhang,Shuohao Zhang,Renlong Wu,Zifei Yan,Wangmeng Zuo
Exposure Bracketing Is All You Need For A High-Quality Image
It is highly desired but challenging to acquire high-quality photos with clear content in low-light environments. Although multi-image processing methods (using burst, dual-exposure, or multi-exposure images) have made significant progress in addressing this issue, they typically focus on specific restoration or enhancement problems, and do not fully explore the potential of utilizing multiple images. Motivated by the fact that multi-exposure images are complementary in denoi

sing, deblurring, high dynamic range imaging, and super-resolution, we propose t
o utilize exposure bracketing photography to get a high-quality image by combini
ng these tasks in this work. Due to the difficulty in collecting real-world pair
s, we suggest a solution that first pre-trains the model with synthetic paired d
ata and then adapts it to real-world unlabeled images. In particular, a temporal
ly modulated recurrent network (TMRNet) and self-supervised adaptation method ar
e proposed. Moreover, we construct a data simulation pipeline to synthesize pair
s and collect real-world images from 200 nighttime scenarios. Experiments on bot
h datasets show that our method performs favorably against the state-of-the-art
multi-image processing ones. Code and datasets are available at https://github.c
om/cszhilu1998/BracketIRE.
****

Xiang Li,Pengfei Li,Yupeng Zheng,Wei Sun,Yan Wang,yilun chen
Semi-Supervised Vision-Centric 3D Occupancy World Model for Autonomous Driving
Understanding world dynamics is crucial for planning in autonomous driving. Rece
nt methods attempt to achieve this by learning a 3D occupancy world model that f
orecasts future surrounding scenes based on current observation. However, 3D occ
upancy labels are still required to produce promising results. Considering the h
igh annotation cost for 3D outdoor scenes, we propose a semi-supervised vision-c
entric 3D occupancy world model, **PreWorld**, to leverage the potential of 2D l
abels through a novel two-stage training paradigm: the self-supervised pre-train
ing stage and the fully-supervised fine-tuning stage. Specifically, during the p
re-training stage, we utilize an attribute projection head to generate different
 attribute fields of a scene (e.g., RGB, density, semantic), thus enabling tempo
ral supervision from 2D labels via volume rendering techniques. Furthermore, we
introduce a simple yet effective state-conditioned forecasting module to recursi
vely forecast future occupancy and ego trajectory in a direct manner. Extensive
experiments on the nuScenes dataset validate the effectiveness and scalability o
f our method, and demonstrate that PreWorld achieves competitive performance acr
oss 3D occupancy prediction, 4D occupancy forecasting and motion planning tasks.
****

Jinyoung Choi,Junoh Kang,Bohyung Han
Enhanced Diffusion Sampling via Extrapolation with Multiple ODE Solutions
Diffusion probabilistic models (DPMs), while effective in generating high-qualit
y samples, often suffer from high computational costs due to the iterative sampl
ing process. To address this, we propose an enhanced ODE-based sampling method f
or DPMs inspired by Richardson extrapolation, which has been shown to reduce num
erical error and improve convergence rates. Our method, termed RX-DPM, utilizes
numerical solutions obtained over multiple denoising steps, leveraging the multi
ple ODE solutions to extrapolate the denoised prediction in DPMs. This signi■can
tly enhances the accuracy of estimations for the ■nal sample while preserving th
e number of function evaluations (NFEs). Unlike standard Richardson extrapolatio
n, which assumes uniform discretization of the time grid, we have developed a mo
re general formulation tailored to arbitrary time step scheduling, guided by the
 local truncation error derived from a baseline sampling method. The simplicity
of our approach facilitates accurate estimation of numerical solutions without a
dditional computational overhead, and allows for seamless and convenient integra
tion into various DPMs and solvers. Additionally, RX-DPM provides explicit error
 estimates, effectively illustrating the faster convergence achieved as the orde
r of the leading error term increases. Through a series of experiments, we demon
strate that the proposed method effectively enhances the quality of generated sa
mples without requiring additional sampling iterations.
****

Siavash Ameli,Siyuan Zhuang,Ion Stoica,Michael W. Mahoney
A Statistical Framework for Ranking LLM-based Chatbots
Large language models (LLMs) have transformed natural language processing, with
frameworks like Chatbot Arena providing pioneering platforms for evaluating thes
e models. By facilitating millions of pairwise comparisons based on human judgme
nts, Chatbot Arena has become a cornerstone in LLM evaluation, offering rich dat
asets for ranking models in open-ended conversational tasks. Building upon this

foundation, we propose a statistical framework that incorporates key advancements to address specific challenges in pairwise comparison analysis. First, we introduce a factored tie model that enhances the ability to handle ties—an integral aspect of human-judged comparisons—significantly improving the model's fit to observed data. Second, we extend the framework to model covariance between competitors, enabling deeper insights into performance relationships and facilitating intuitive groupings into performance tiers. Third, we resolve optimization challenges arising from parameter non-uniqueness by introducing novel constraints, ensuring stable and interpretable parameter estimation. Through rigorous evaluation and extensive experimentation, our framework demonstrates substantial improvements over existing methods in modeling pairwise comparison data. To support reproducibility and practical adoption, we release leaderbot, an open-source Python package implementing our models and analyses.
****
Xing Hu,Yuan Cheng,Dawei Yang,Zhixuan Chen,Zukang Xu,JiangyongYu,XUCHEN,Zhihang Yuan,Zhe jiang,Sifan Zhou
OSTQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting
Post-training quantization (PTQ) has emerged as a widely adopted technique for compressing and accelerating Large Language Models (LLMs).
The major challenge in LLM quantization is that uneven and heavy-tailed data distributions can expand the quantization range, thereby reducing bit precision for most values.
Recent methods attempt to eliminate outliers and balance inter-channel differences by employing linear transformations; however, they remain heuristic and are often overlook optimizing the data distribution across the entire quantization space.
In this paper, we introduce Quantization Space Utilization Rate (QSUR), a novel metric that effectively assesses the quantizability of transformed data by measuring the space utilization of the data in the quantization space. We complement QSUR with mathematical derivations that examine the effects and limitations of various transformations, guiding our development of Orthogonal and Scaling Transformation-based Quantization (OSTQuant). OSTQuant employs a learnable equivalent transformation, consisting of an orthogonal transformation and a scaling transformation, to optimize the distributions of weights and activations across the entire quantization space. Futhermore, we propose the KL-Top loss function, designed to mitigate noise during optimization while retaining richer semantic information within the limited calibration data imposed by PTQ.
OSTQuant outperforms existing work on various LLMs and benchmarks. In the W4-only setting, it retains 99.5\% of the floating-point accuracy. In the more challenging W4A4KV4 configuration, OSTQuant reduces the performance gap by 32\% on the LLaMA-3-8B model compared to state-of-the-art methods. Code will be available.
****
Ivona Najdenkoska,Mohammad Mahdi Derakhshani,Yuki M Asano,Nanne Van Noord,Marcel Worring,Cees G. M. Snoek
TULIP: Token-length Upgraded CLIP
We address the challenge of representing long captions in vision-language models, such as CLIP. By design these models are limited by fixed, absolute positional encodings, restricting inputs to a maximum of 77 tokens and hindering performance on tasks requiring longer descriptions. Although recent work has attempted to overcome this limit, their proposed approaches struggle to model token relationships over longer distances and simply extend to a fixed new token length. Instead, we propose a generalizable method, named TULIP, able to upgrade the token length to any length for CLIP-like models. We do so by improving the architecture with relative position encodings, followed by a training procedure that (i) distills the original CLIP text encoder into an encoder with relative position encodings and (ii) enhances the model for aligning longer captions with images. By effectively encoding captions longer than the default 77 tokens, our model outperforms baselines on cross-modal tasks such as retrieval and text-to-image generation. The code repository is available at https://github.com/ivonajdenkoska/tulip

****
Shawn Tan,Songlin Yang,Aaron Courville,Rameswar Panda,Yikang Shen
Scaling Stick-Breaking Attention: An Efficient Implementation and In-depth Study
The self-attention mechanism traditionally relies on the softmax operator, necessitating positional embeddings like RoPE, or position biases to account for token order.
But current methods using still face length generalisation challenges.
We investigate an alternative attention mechanism based on the stick-breaking process in larger scale settings.
The method works as follows: For each token before the current, we determine a break point, which represents the proportion of the stick, the weight of the attention, to allocate to the current token.
We repeat this on the remaining stick, until all tokens are allocated a weight, resulting in a sequence of attention weights.
This process naturally incorporates recency bias, which has linguistic motivations for grammar parsing (Shen et al., 2017).
We study the implications of replacing the conventional softmax-based attention mechanism with stick-breaking attention.
We then discuss implementation of numerically stable stick-breaking attention and adapt Flash Attention to accommodate this mechanism.
When used as a drop-in replacement for current softmax+RoPE attention systems, we find that stick-breaking attention performs competitively with current methods on length generalisation and downstream tasks.
Stick-breaking also performs well at length generalisation, allowing a model trained with 2^{11} context window to perform well at 2^{14} with perplexity improvements.
****
Songlin Yang,Jan Kautz,Ali Hatamizadeh
Gated Delta Networks: Improving Mamba2 with Delta Rule
Linear Transformers have gained attention as efficient alternatives to standard Transformers, but their performance in retrieval and long-context tasks has been limited. To address these limitations, recent work has explored two distinct mechanisms: gating for adaptive memory control and the delta update rule for precise memory modifications. We observe that these mechanisms are complementary—gating enables rapid memory erasure while the delta rule facilitates targeted updates. Building on this insight, we introduce the gated delta rule and develop a parallel training algorithm optimized for modern hardware. Our proposed architecture, Gated DeltaNet, consistently surpasses existing models like Mamba2 and DeltaNet across multiple benchmarks, including language modeling, common-sense reasoning, in-context retrieval, length extrapolation, and long-context understanding. We further enhance performance by developing hybrid architectures that combine Gated DeltaNet layers with sliding window attention or Mamba2 layers, achieving both improved training efficiency and superior task performance.
****
Daniel Paleka,Abhimanyu Pallavi Sudhir,Alejandro Alvarez,Vineeth Bhat,Adam Shen,Evan Wang,Florian Tramèr
Consistency Checks for Language Model Forecasters
Forecasting is a task that is difficult to evaluate: the ground truth can only be known in the future. Recent work showing LLM forecasters rapidly approaching human-level performance begs the question: how can we benchmark and evaluate these forecasters *instantaneously*? Following the consistency check framework, we measure the performance of forecasters in terms of the consistency of their predictions on different logically-related questions. We propose a new, general consistency metric based on *arbitrage*: for example, if a forecasting AI illogically predicts that both the Democratic and Republican parties have 60\% probability of winning the 2024 US presidential election, an arbitrageur could trade against the forecaster's predictions and make a profit. We build an automated evaluation system that generates a set of base questions, instantiates consistency checks from these questions, elicits the predictions of the forecaster, and measures the consistency of the predictions. We then build a standard, proper-scoring-rule

forecasting benchmark, and show that our (instantaneous) consistency metrics correlate strongly with LLM forecasters' ground truth Brier scores (which are only known in the future). We also release a consistency benchmark that resolves in 2028, providing a long-term evaluation tool for forecasting.
****
Connor Mooney,Zhongjian Wang,Jack Xin,Yifeng Yu
Global Well-posedness and Convergence Analysis of Score-based Generative Models via Sharp Lipschitz Estimates
We establish global well-posedness and convergence of the score-based generative models (SGM) under minimal general assumptions of initial data for score estimation. For the smooth case, we start from a Lipschitz bound of the score function with optimal time length. The optimality is validated by an example whose Lipschitz constant of scores is bounded at initial but blows up in finite time. This necessitates the separation of time scales in conventional bounds for non-log-concave distributions. In contrast, our follow up analysis only relies on a local Lipschitz condition and is valid globally in time. This leads to the convergence of numerical scheme without time separation. For the non-smooth case, we show that the optimal Lipschitz bound is $O(1/t)$ in the point-wise sense for distributions supported on a compact, smooth and low-dimensional manifold with boundary.
****
Zilin Fang,David Hsu,Gim Hee Lee
Neuralized Markov Random Field for Interaction-Aware Stochastic Human Trajectory Prediction
Interactive human motions and the continuously changing nature of intentions pose significant challenges for human trajectory prediction. In this paper, we present a neuralized Markov random field (MRF)-based motion evolution method for probabilistic interaction-aware human trajectory prediction. We use MRF to model each agent's motion and the resulting crowd interactions over time, hence is robust against noisy observations and enables group reasoning. We approximate the modeled distribution using two conditional variational autoencoders (CVAEs) for efficient learning and inference. Our proposed method achieves state-of-the-art performance on ADE/FDE metrics across two dataset categories: overhead datasets ETH/UCY, SDD, and NBA, and ego-centric JRDB. Furthermore, our approach allows for real-time stochastic inference in bustling environments, making it well-suited for a 30FPS video setting. We open-source our codes at: https://github.com/AdaCompNUS/NMRF_TrajectoryPrediction.git
****
Jianhao Huang,Zixuan Wang,Jason D. Lee
Transformers Learn to Implement Multi-step Gradient Descent with Chain of Thought
Chain of Thought (CoT) prompting has been shown to significantly improve the performance of large language models (LLMs), particularly in arithmetic and reasoning tasks, by instructing the model to produce intermediate reasoning steps. Despite the remarkable empirical success of CoT and its theoretical advantages in enhancing expressivity, the mechanisms underlying CoT training remain largely unexplored. In this paper, we study the training dynamics of transformers over a CoT objective on a in-context weight prediction task for linear regression. We prove that while a one-layer linear transformer without CoT can only implement a single step of gradient descent (GD) and fails to recover the ground-truth weight vector, a transformer with CoT prompting can learn to perform multi-step GD autoregressively, achieving near-exact recovery. Furthermore, we show that the trained transformer effectively generalizes on the unseen data. Empirically, we demonstrate that CoT prompting yields substantial performance improvements.
****
Zijing Shi,Meng Fang,Ling Chen
Monte Carlo Planning with Large Language Model for Text-Based Game Agents
Text-based games provide valuable environments for language-based autonomous agents. However, planning-then-learning paradigms, such as those combining Monte Carlo Tree Search (MCTS) and reinforcement learning (RL), are notably time-consuming due to extensive iterations. Additionally, these algorithms perform uncertain

ty-driven exploration but lack language understanding and reasoning abilities. In this paper, we introduce the Monte Carlo planning with Dynamic Memory-guided Large language model (MC-DML) algorithm. MC-DML leverages the language understanding and reasoning capabilities of Large Language Models (LLMs) alongside the exploratory advantages of tree search algorithms. Specifically, we enhance LLMs with in-trial and cross-trial memory mechanisms, enabling them to learn from past experiences and dynamically adjust action evaluations during planning. We conduct experiments on a series of text-based games from the Jericho benchmark. Our results demonstrate that the MC-DML algorithm significantly enhances performance across various games at the initial planning phase, outperforming strong contemporary methods that require multiple iterations. This demonstrates the effectiveness of our algorithm, paving the way for more efficient language-grounded planning in complex environments.
****

Priya Sundaresan,Hengyuan Hu,Quan Vuong,Jeannette Bohg,Dorsa Sadigh
What's the Move? Hybrid Imitation Learning via Salient Points
While imitation learning (IL) offers a promising framework for teaching robots various behaviors, learning complex tasks remains challenging. Existing IL policies struggle to generalize effectively across visual and spatial variations even for simple tasks. In this work, we introduce **SPHINX**: **S**alient **P**oint-based **H**ybrid **I**mitatio**N** and e**X**ecution, a flexible IL policy that leverages multimodal observations (point clouds and wrist images), along with a hybrid action space of low-frequency, sparse waypoints and high-frequency, dense end effector movements. Given 3D point cloud observations, SPHINX learns to infer task-relevant points within a point cloud, or *salient points*, which support spatial generalization by focusing on semantically meaningful features. These salient points serve as anchor points to predict waypoints for long-range movement, such as reaching target poses in free-space. Once near a salient point, SPHINX learns to switch to predicting dense end-effector movements given close-up wrist images for precise phases of a task. By exploiting the strengths of different input modalities and action representations for different manipulation phases, SPHINX tackles complex tasks in a sample-efficient, generalizable manner. Our method achieves **86.7%** success across 4 real-world and 2 simulated tasks, outperforming the next best state-of-the-art IL baseline by **41.1%** on average across **440** real world trials. SPHINX additionally generalizes to novel viewpoints, visual distractors, spatial arrangements, and execution speeds with a **1.7x** speedup over the most competitive baseline. Our website (http://sphinx-manip.github.io) provides open-sourced code for data collection, training, and evaluation, along with supplementary videos.
****

Lukas Miklautz,Timo Klein,Kevin Sidak,Collin Leiber,Thomas Lang,Andrii Shkabrii,Sebastian Tschiatschek,Claudia Plant
Breaking the Reclustering Barrier in Centroid-based Deep Clustering
This work investigates an important phenomenon in centroid-based deep clustering (DC) algorithms: Performance quickly saturates after a period of rapid early gains. Practitioners commonly address early saturation with periodic reclustering, which we demonstrate to be insufficient to address performance plateaus. We call this phenomenon the “reclustering barrier” and empirically show when the reclustering barrier occurs, what its underlying mechanisms are, and how it is possible to Break the Reclustering Barrier with our algorithm BRB. BRB avoids early over-commitment to initial clusterings and enables continuous adaptation to reinitialized clustering targets while remaining conceptually simple. Applying our algorithm to widely-used centroid-based DC algorithms, we show that (1) BRB consistently improves performance across a wide range of clustering benchmarks, (2) BRB enables training from scratch, and (3) BRB performs competitively against state-of-the-art DC algorithms when combined with a contrastive loss. We release our code and pre-trained models at https://github.com/Probabilistic-and-Interactive-ML/breaking-the-reclustering-barrier .
****

Shuai Wang,Malu Zhang,Dehao Zhang,Ammar Belatreche,Yichen Xiao,Yu Liang,Yimeng S

han,Qian Sun,Enqi Zhang,Yang Yang
Spiking Vision Transformer with Saccadic Attention
The combination of Spiking Neural Networks (SNNs) and Vision Transformers (ViTs) holds potential for achieving both energy efficiency and high performance, particularly suitable for edge vision applications. However, a significant performance gap still exists between SNN-based ViTs and their ANN counterparts. Here, we first analyze why SNN-based ViTs suffer from limited performance and identify a mismatch between the vanilla self-attention mechanism and spatio-temporal spike trains. This mismatch results in degraded spatial relevance and limited temporal interactions. To address these issues, we draw inspiration from biological saccadic attention mechanisms and introduce an innovative Saccadic Spike Self-Attention (SSSA) method. Specifically, in the spatial domain, SSSA employs a novel spike distribution-based method to effectively assess the relevance between Query and Key pairs in SNN-based ViTs. Temporally, SSSA employs a saccadic interaction module that dynamically focuses on selected visual areas at each timestep and significantly enhances whole scene understanding through temporal interactions. Building on the SSSA mechanism, we develop a SNN-based Vision Transformer (SNN-ViT). Extensive experiments across various visual tasks demonstrate that SNN-ViT achieves state-of-the-art performance with linear computational complexity. The effectiveness and efficiency of the SNN-ViT highlight its potential for power-critical edge vision applications.
****
Siyuan Wu,Leong Hou U,Panagiotis Karras
Balancing Bias in Two-sided Markets for Fair Stable Matchings
The Balanced Stable Marriage (BSM) problem aims to find a stable matching in a two-sided market that minimizes the maximum dissatisfaction among two sides. The classical Deferred Acceptance algorithm merely produces an unfair stable marriage, providing optimal partners for one side while partially assigning pessimal partners to the other. Solving BSM is NP-hard, thwarting attempts to resolve the problem exactly. As the instance size increases in practice, recent studies have explored heuristics for finding a fair stable marriage but have not found an exact optimal solution for BSM efficiently. Nevertheless, in this paper we propose an efficient algorithm, Isorropia, that returns the exact optimal solution to practical BSM problem instances. Isorropia constructs two sets of candidate rotations from which it builds three sets of promising antichains, and performs local search on those three sets of promising antichains. Our extensive experimental study shows that Isorropia surpasses the time-efficiency of baselines that return the exact solution by up to three orders of magnitude.
****
Junjie Chen,Xiangheng He,Yusuke Miyao,Danushka Bollegala
Improving Unsupervised Constituency Parsing via Maximizing Semantic Information
Unsupervised constituency parsers organize phrases within a sentence into a tree-shaped syntactic constituent structure that reflects the organization of sentence semantics.
However, the traditional objective of maximizing sentence log-likelihood (LL) does not explicitly account for the close relationship between the constituent structure and the semantics, resulting in a weak correlation between LL values and parsing accuracy.
In this paper, we introduce a novel objective that trains parsers by maximizing SemInfo, the semantic information encoded in constituent structures.
We introduce a bag-of-substrings model to represent the semantics and estimate the SemInfo value using the probability-weighted information metric.
We apply the SemInfo maximization objective to training Probabilistic Context-Free Grammar (PCFG) parsers and develop a Tree Conditional Random Field (TreeCRF)-based model to facilitate the training.
Experiments show that SemInfo correlates more strongly with parsing accuracy than LL, establishing SemInfo as a better unsupervised parsing objective.
As a result, our algorithm significantly improves parsing accuracy by an average of 7.85 sentence-F1 scores across five PCFG variants and in four languages, achieving state-of-the-art level results in three of the four languages.

****
Renjie Li,Panwang Pan,Bangbang Yang,Dejia Xu,Shijie Zhou,Xuanyang Zhang,Zeming Li,Achuta Kadambi,Zhangyang Wang,Zhengzhong Tu,Zhiwen Fan
4K4DGen: Panoramic 4D Generation at 4K Resolution
The blooming of virtual reality and augmented reality (VR/AR) technologies has driven an increasing demand for the creation of high-quality, immersive, and dynamic environments. However, existing generative techniques either focus solely on dynamic objects or perform outpainting from a single perspective image, failing to meet the requirements of VR/AR applications that need free-viewpoint, 360° virtual views where users can move in all directions. In this work, we tackle the challenging task of elevating a single panorama to an immersive 4D experience. For the first time, we demonstrate the capability to generate omnidirectional dynamic scenes with 360° views at 4K (4096 $\times$ 2048) resolution, thereby providing an immersive user experience. Our method introduces a pipeline that facilitates natural scene animations and optimizes a set of 3D Gaussians using efficient splatting techniques for real-time exploration. To overcome the lack of scene-scale annotated 4D data and models, especially in panoramic formats, we propose a novel Panoramic Denoiser that adapts generic 2D diffusion priors to animate consistently in 360° images, transforming them into panoramic videos with dynamic scenes at targeted regions. Subsequently, we propose Dynamic Panoramic Lifting to elevate the panoramic video into a 4D immersive environment while preserving spatial and temporal consistency. By transferring prior knowledge from 2D models in the perspective domain to the panoramic domain and the 4D lifting with spatial appearance and geometry regularization, we achieve high-quality Panorama-to-4D generation at a resolution of 4K for the first time. Project page: https://4k4dgen.github.io/.
****
Run Luo,Yunshui Li,Longze Chen,Wanwei He,Ting-En Lin,Ziqiang Liu,Lei Zhang,Zikai Song,Hamid Rokny,Xiaobo Xia,Tongliang Liu,Binyuan Hui,Min Yang
DEEM: Diffusion models serve as the eyes of large language models for image perception
The development of large language models (LLMs) has significantly advanced the emergence of large multimodal models (LMMs). While LMMs have achieved tremendous success by promoting the synergy between multimodal comprehension and creation, they often face challenges when confronted with out-of-distribution data, such as which can hardly distinguish orientation, quantity, color, structure, etc. This is primarily due to their reliance on image encoders trained to encode images into task-relevant features, which may lead them to disregard irrelevant details. Delving into the modeling capabilities of diffusion models for images naturally prompts the question: Can diffusion models serve as the eyes of large language models for image perception? In this paper, we propose DEEM, a simple but effective approach that utilizes the generative feedback of diffusion models to align the semantic distributions of the image encoder. This addresses the drawbacks of previous methods that solely relied on image encoders like CLIP-ViT, thereby enhancing the model's resilience against out-of-distribution samples and reducing visual hallucinations. Importantly, this is achieved without requiring additional training modules and with fewer training parameters. We extensively evaluated DEEM on both our newly constructed RobustVQA benchmark and other well-known benchmarks, POPE and MMVP, for visual hallucination and perception. In particular, DEEM improves LMM's visual perception performance to a large extent (e.g., 4\% ↑ on RobustVQA, 6.5\% ↑ on MMVP and 12.8 \% ↑ on POPE). Compared to the state-of-the-art interleaved content generation models, DEEM exhibits enhanced robustness and a superior capacity to alleviate model hallucinations while utilizing fewer trainable parameters, less pre-training data (10\%), and a smaller base model size. Extensive experiments demonstrate that DEEM enhances the performance of LMMs on various downstream tasks without inferior performance in the long term, including visual question answering, image captioning, and text-conditioned image synthesis.
****
Thieu Vo,Duy-Tung Pham,Xin T. Tong,Tan Minh Nguyen

Demystifying the Token Dynamics of Deep Selective State Space Models
Selective state space models (SSM), such as Mamba, have gained prominence for th
eir effectiveness in modeling sequential data. Despite their outstanding empiric
al performance, a comprehensive theoretical understanding of deep selective SSM
remains elusive, hindering their further development and adoption for applicatio
ns that need high fidelity. In this paper, we investigate the dynamical properti
es of tokens in a pre-trained Mamba model. In particular, we derive the dynamica
l system governing the continuous-time limit of the Mamba model and characterize
 the asymptotic behavior of its solutions. In the one-dimensional case, we prove
 that only one of the following two scenarios happens: either all tokens converg
e to zero, or all tokens diverge to infinity. We provide criteria based on mode
l parameters to determine when each scenario occurs. For the convergent scenario
, we empirically verify that this scenario negatively impacts the model's perfor
mance. For the divergent scenario, we prove that different tokens will diverge
to infinity at different rates, thereby contributing unequally to the updates du
ring model training. Based on these investigations, we propose two refinements
for the model: excluding the convergent scenario and reordering tokens based on
their importance scores, both aimed at improving practical performance. Our exp
erimental results validate these refinements, offering insights into enhancing M
amba's effectiveness in real-world applications.
****
Ian Connick Covert,Tony Sun,James Zou,Tatsunori Hashimoto
Locality Alignment Improves Vision-Language Models
Vision language models (VLMs) have seen growing adoption in recent years, but ma
ny still struggle with basic spatial reasoning errors. We hypothesize that this
is due to VLMs adopting pre-trained vision backbones, specifically vision transf
ormers (ViTs) trained with image-level supervision and minimal inductive biases.
 Such models may fail to encode the class contents at each position in the image
, and our goal is to resolve this with a vision backbone that effectively captur
es both local and global image semantics. Our main insight is that we do not req
uire new supervision to learn this capability – pre-trained models contain signi
ficant knowledge of local semantics that we can extract and use for scalable sel
f-supervision. We propose a new efficient post-training stage for ViTs called lo
cality alignment and a novel fine-tuning procedure called MaskEmbed that uses a
masked reconstruction loss to learn semantic contributions for each image patch.
 We first evaluate locality alignment with a vision-only benchmark, finding that
 it improves a model’s performance at patch-level semantic segmentation, especia
lly for strong backbones trained with image-caption pairs (e.g., CLIP and SigLIP
). We then train a series of VLMs with and without locality alignment, and show
that locality-aligned backbones improve performance across a range of benchmarks
, particularly ones that involve spatial understanding (e.g., RefCOCO, OCID-Ref,
 TallyQA, VSR, AI2D). Overall, we demonstrate that we can efficiently learn loca
l semantic extraction via a locality alignment stage, and that this procedure be
nefits VLM training recipes that use off-the-shelf vision backbones.
****
Sascha Marton,Tim Grams,Florian Vogt,Stefan Lüdtke,Christian Bartelt,Heiner Stuc
kenschmidt
Mitigating Information Loss in Tree-Based Reinforcement Learning via Direct Opti
mization
Reinforcement learning (RL) has seen significant success across various domains,
 but its adoption is often limited by the black-box nature of neural network pol
icies, making them difficult to interpret. In contrast, symbolic policies allow
representing decision-making strategies in a compact and interpretable way. Howe
ver, learning symbolic policies directly within on-policy methods remains challe
nging.
In this paper, we introduce SYMPOL, a novel method for SYMbolic tree-based on-PO
Licy RL. SYMPOL employs a tree-based model integrated with a policy gradient met
hod, enabling the agent to learn and adapt its actions while maintaining a high
level of interpretability.
We evaluate SYMPOL on a set of benchmark RL tasks, demonstrating its superiority

over alternative tree-based RL approaches in terms of performance and interpretability. Unlike existing methods, it enables gradient-based, end-to-end learning of interpretable, axis-aligned decision trees within standard on-policy RL algorithms. Therefore, SYMPOL can become the foundation for a new class of interpretable RL based on decision trees. Our implementation is available under: https://github.com/s-marton/sympol
****
Kartik Thakral,Rishabh Ranjan,Akanksha Singh,Akshat Jain,Mayank Vatsa,Richa Singh
ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset
The proliferation of deepfakes and AI-generated content has led to a surge in media forgeries and misinformation, necessitating robust detection systems. However, current datasets lack diversity across modalities, languages, and real-world scenarios. To address this gap, we present ILLUSION (Integration of Life-Like Unique Synthetic Identities and Objects from Neural Networks), a large-scale, multi-modal
deepfake dataset comprising 1.3 million samples spanning audio-visual forgeries, 26 languages, challenging noisy environments, and various manipulation protocols. Generated using 28 state-of-the-art generative techniques, ILLUSION includes faceswaps, audio spoofing, synchronized audio-video manipulations, and synthetic media while ensuring a balanced representation of gender and skin tone for unbiased evaluation. Using Jaccard Index and UpSet plot analysis, we demonstrate ILLUSION’s distinctiveness and minimal overlap with existing datasets, emphasizing its novel generative coverage. We benchmarked image, audio, video, and multi-modal detection models, revealing key challenges such as performance degradation in multilingual and multi-modal contexts, vulnerability to real-world distortions, and limited generalization to zero-day attacks. By bridging synthetic and real-world complexities, ILLUSION provides a challenging yet essential platform for advancing deepfake detection research. The dataset is publicly available at https://www.iab-rubric.org/illusion-database.
****
Chanwoo Park,Xiangyu Liu,Asuman E. Ozdaglar,Kaiqing Zhang
Do LLM Agents Have Regret? A Case Study in Online Learning and Games
Large language models (LLMs) have been increasingly employed for (interactive) decision-making, via the development of LLM-based autonomous agents. Despite their emerging successes, the performance of LLM agents in decision-making has not been fully investigated through quantitative metrics, especially in the multi-agent setting when they interact with each other, a typical scenario in real-world LLM-agent applications. To better understand the limits of LLM agents in these interactive environments, we propose to study their interactions in benchmark decision-making settings in online learning and game theory, through the performance metric of regret. We first empirically study the no-regret behaviors of LLMs in canonical non-stochastic online learning problems, as well as the emergence of equilibria when LLM agents interact through playing repeated games. We then provide some theoretical insights into the no-regret behaviors of LLM agents, under certain assumptions on the supervised pre-training and the rationality model of human decision-makers who generate the data. Notably, we also identify (simple) cases where advanced LLMs such as GPT-4 fail to be no-regret. To further promote the no-regret behaviors, we propose a novel unsupervised training loss of regret-loss, which, in contrast to the supervised pre-training loss, does not require the labels of (optimal) actions. Finally, we establish the statistical guarantee of generalization bound for regret-loss minimization, and more importantly, the optimization guarantee that minimizing such a loss may automatically lead to known no-regret learning algorithms, when single-layer self-attention models are used. Our further experiments demonstrate the effectiveness of our regret-loss, especially in addressing the above “regrettable” cases.
****
Lin Long,Xijun Gu,Xinjie Sun,Wentao Ye,Haobo Wang,Sai Wu,Gang Chen,Junbo Zhao
Bridging the Semantic Gap Between Text and Table: A Case Study on NL2SQL

The rise of Large Language Models (LLMs) has revolutionized numerous domains, ye
t these models still exhibit weakness in understanding structured tabular data.
Although the growing context window promises to accommodate a larger volume of t
able contents, it does not inherently improve the model's ability to understand
the underlying structure and semantics of tabular data.
To bridge the semantic gap between **T**ext and **T**able, we propose **T**n**T*
*, a table-language model that features multimodal table representations to empo
wer LLMs to effectively and efficiently abstract structure-enriched semantics fr
om tabular data.
Tn**T** also introduces a scalable and efficient training pipeline, featurin
g novel self-supervised tasks, to integrate abstract tabular knowledge into the
language modality.
Extensive experimental results on NL2SQL demonstrate a much better table underst
anding of **T**n**T**, which achieves up to **14.4** higher execution accuracy c
ompared with traditional text-based table representations.
****

Youn-Yeol Yu,Jeongwhan Choi,Jaehyeon Park,Kookjin Lee,Noseong Park
PIORF: Physics-Informed Ollivier-Ricci Flow for Long–Range Interactions in Mesh
Graph Neural Networks
Recently, data-driven simulators based on graph neural networks have gained atte
ntion in modeling physical systems on unstructured meshes. However, they struggl
e with long-range dependencies in fluid flows, particularly in refined mesh regi
ons. This challenge, known as the 'over-squashing' problem, hinders information
propagation. While existing graph rewiring methods address this issue to some ex
tent, they only consider graph topology, overlooking the underlying physical phe
nomena. We propose Physics-Informed Ollivier--Ricci Flow (PIORF), a novel rewiri
ng method that combines physical correlations with graph topology. PIORF uses Ol
livier--Ricci curvature (ORC) to identify bottleneck regions and connects these
areas with nodes in high-velocity gradient nodes, enabling long-range interactio
ns and mitigating over-squashing. Our approach is computationally efficient in r
ewiring edges and can scale to larger simulations. Experimental results on 3 flu
id dynamics benchmark datasets show that PIORF consistently outperforms baseline
 models and existing rewiring methods, achieving up to 26.2\% improvement.
****

Tianyuan Chen,Ronglong Cai,Faguo Wu,Xiao Zhang
ACTIVE: Offline Reinforcement Learning via Adaptive Imitation and In-sample V-
Ensemble
Offline reinforcement learning (RL) aims to learn from static datasets and thus
faces the challenge of value estimation errors for out-of-distribution actions.
The in-sample learning scheme addresses this issue by performing implicit TD bac
kups that does not query the values of unseen actions. However, pre-existing in-
sample value learning and policy extraction methods suffer from over-regularizat
ion, limiting their performance on suboptimal or compositional datasets. In this
 paper, we analyze key factors in in-sample learning that might potentially hind
er the use of a milder constraint. We propose Actor-Critic with Temperature adju
stment and In-sample Value Ensemble (ACTIVE), a novel in-sample offline RL algor
ithm that leverages an ensemble of V-functions for critic training and adaptiv
ely adjusts the constraint level using dual gradient descent. We theoretically s
how that the V-ensemble suppresses the accumulation of initial value errors, t
hereby mitigating overestimation. Our experiments on the D4RL benchmarks demonst
rate that ACTIVE alleviates overfitting of value functions and outperforms exist
ing in-sample methods in terms of learning stability and policy optimality.
****

Dingyi Zhuang,Chonghe Jiang,Yunhan Zheng,Shenhao Wang,Jinhua Zhao
GETS: Ensemble Temperature Scaling for Calibration in Graph Neural Networks
Graph Neural Networks (GNNs) deliver strong classification results but often suf
fer from poor calibration performance, leading to overconfidence or underconfide
nce. This is particularly problematic in high-stakes applications where accurate
 uncertainty estimates are essential. Existing post-hoc methods, such as tempera
ture scaling, fail to effectively utilize graph structures, while current GNN ca

libration methods often overlook the potential of leveraging diverse input information and model ensembles jointly. In the paper, we propose Graph Ensemble Temperature Scaling (GETS), a novel calibration framework that combines input and model ensemble strategies within a Graph Mixture-of-Experts (MoE) architecture. GETS integrates diverse inputs, including logits, node features, and degree embeddings, and adaptively selects the most relevant experts for each node’s calibration procedure. Our method outperforms state-of-the-art calibration techniques, reducing expected calibration error (ECE) by $\geq$ 25% across 10 GNN benchmark datasets. Additionally, GETS is computationally efficient, scalable, and capable of selecting effective input combinations for improved calibration performance. The implementation is available at https://github.com/ZhuangDingyi/GETS/.
****

Asad Aali,Giannis Daras,Brett Levac,Sidharth Kumar,Alex Dimakis,Jon Tamir
Ambient Diffusion Posterior Sampling: Solving Inverse Problems with Diffusion Models Trained on Corrupted Data
We provide a framework for solving inverse problems with diffusion models learned from linearly corrupted data. Firstly, we extend the Ambient Diffusion framework to enable training directly from measurements corrupted in the Fourier domain. Subsequently, we train diffusion models for MRI with access only to Fourier subsampled multi-coil measurements at acceleration factors R$=2, 4, 6, 8$. Secondly, we propose $\textit{Ambient Diffusion Posterior Sampling}$ (A-DPS), a reconstruction algorithm that leverages generative models pre-trained on one type of corruption (e.g. image inpainting) to perform posterior sampling on measurements from a different forward process (e.g. image blurring). For MRI reconstruction in high acceleration regimes, we observe that A-DPS models trained on subsampled data are better suited to solving inverse problems than models trained on fully sampled data. We also test the efficacy of A-DPS on natural image datasets (CelebA, FFHQ, and AFHQ) and show that A-DPS can sometimes outperform models trained on clean data for several image restoration tasks in both speed and performance.
****

Yaochen Zhu,Jing Ma,Liang Wu,Qi Guo,Liangjie Hong,Jundong Li
Causal Effect Estimation with Mixed Latent Confounders and Post-treatment Variables
Causal inference from observational data has attracted considerable attention among researchers. One main obstacle is the handling of confounders. As direct measurement of confounders may not be feasible, recent methods seek to address the confounding bias via proxy variables, i.e., covariates postulated to be conducive to the inference of latent confounders. However, the selected proxies may scramble both confounders and post-treatment variables in practice, which risks biasing the estimation by controlling for variables affected by the treatment. In this paper, we systematically investigate the bias due to latent post-treatment variables, i.e., latent post-treatment bias, in causal effect estimation. Specifically, we first derive the bias when selected proxies scramble both latent confounders and post-treatment variables, which we demonstrate can be arbitrarily bad. We then propose a Confounder-identifiable VAE (CiVAE) to address the bias. Based on a mild assumption that the prior of latent variables that generate the proxy belongs to a general exponential family with at least one invertible sufficient statistic in the factorized part, CiVAE individually identifies latent confounders and latent post-treatment variables up to bijective transformations. We then prove that with individual identification, the intractable disentanglement problem of latent confounders and post-treatment variables can be transformed into a tractable independence test problem despite arbitrary dependence may exist among them. Finally, we prove that the true causal effects can be unbiasedly estimated with transformed confounders inferred by CiVAE. Experiments on both simulated and real-world datasets demonstrate significantly improved robustness of CiVAE.
****

You Qin,Qilong Wu,Yicong Li,Wei Ji,Li Li,Pengcheng Cai,Lina Wei,Roger Zimmermann
Generalized Video Moment Retrieval
In this paper, we introduce the Generalized Video Moment Retrieval (GVMR) framew

ork, which extends traditional Video Moment Retrieval (VMR) to handle a wider ra
nge of query types. Unlike conventional VMR systems, which are often limited to
simple, single-target queries, GVMR accommodates both non-target and multi-targe
t queries. To support this expanded task, we present the NExT-VMR dataset, deriv
ed from the YFCC100M collection, featuring diverse query scenarios to enable mor
e robust model evaluation.
Additionally, we propose BCANet, a transformer-based model incorporating the nov
el Boundary-aware Cross Attention (BCA) module. The BCA module enhances boundary
 detection and uses cross-attention to achieve a comprehensive understanding of
video content in relation to queries. BCANet accurately predicts temporal video
segments based on natural language descriptions, outperforming traditional model
s in both accuracy and adaptability. Our results demonstrate the potential of th
e GVMR framework, the NExT-VMR dataset, and BCANet to advance VMR systems, setti
ng a new standard for future multimedia information retrieval research.
****

Giannis Daras,Yeshwanth Cherapanamjeri,Constantinos Costis Daskalakis
How Much is a Noisy Image Worth? Data Scaling Laws for Ambient Diffusion.
The quality of generative models depends on the quality of the data they are tra
ined on. Creating large-scale, high-quality datasets is often expensive and some
times impossible, e.g.~in certain scientific applications where there is no acce
ss to clean data due to physical or instrumentation constraints. Ambient Diffusi
on and related frameworks train diffusion models with solely corrupted data (whi
ch are usually cheaper to acquire) but ambient models significantly underperform
 models trained on clean data. We study this phenomenon at scale by training mor
e than 80 models on data with different corruption levels across three dataset
s ranging from $30,000$ to ≈ 1.3M samples. We show that it is impossible
, at these sample sizes, to match the performance of models trained on clean dat
a when only training on noisy data. Yet, a combination of a small set of clean d
ata (e.g.~10% of the total dataset) and a large set of highly noisy data suff
ices to reach the performance of models trained solely on similar-size datasets
of clean data, and in particular to achieve near state-of-the-art performance. W
e provide theoretical evidence for our findings by developing novel sample compl
exity bounds for learning from Gaussian Mixtures with heterogeneous variances. O
ur theoretical model suggests that, for large enough datasets, the effective mar
ginal utility of a noisy sample is exponentially worse that of a clean sample. P
roviding a small set of clean samples can significantly reduce the sample size r
equirements for noisy data, as we also observe in our experiments.
****

Yuejiang Liu,Jubayer Ibn Hamid,Annie Xie,Yoonho Lee,Max Du,Chelsea Finn
Bidirectional Decoding: Improving Action Chunking via Guided Test-Time Sampling
Predicting and executing a sequence of actions without intermediate replanning,
known as action chunking, is increasingly used in robot learning from human demo
nstrations. Yet, its effects on the learned policy remain inconsistent: some stu
dies find it crucial for achieving strong results, while others observe decrease
d performance. In this paper, we first dissect how action chunking impacts the d
ivergence between a learner and a demonstrator. We find that action chunking all
ows the learner to better capture the temporal dependencies in demonstrations bu
t at the cost of reduced reactivity to unexpected states. To address this tradeo
ff, we propose Bidirectional Decoding (BID), a test-time inference algorithm tha
t bridges action chunking with closed-loop adaptation. At each timestep, BID sam
ples multiple candidate predictions and searches for the optimal one based on tw
o criteria: (i) backward coherence, which favors samples that align with previou
s decisions; (ii) forward contrast, which seeks samples of high likelihood for f
uture plans. By coupling decisions within and across action chunks, BID promotes
 both long-term consistency and short-term reactivity. Experimental results show
 that our method boosts the performance of two state-of-the-art generative polic
ies across seven simulation benchmarks and two real-world tasks. Code and videos
 are available at https://bid-robot.github.io.
****

Jaehyun Pahk,Donghyeon Kwon,Seong Joon Oh,Suha Kwak

Decoupled Finetuning for Domain Generalizable Semantic Segmentation
Joint finetuning of a pretrained encoder and a randomly initialized decoder has been the de facto standard in semantic segmentation, but the vulnerability of this approach to domain shift has not been studied. We investigate the vulnerability issue of joint finetuning, and propose a novel finetuning framework called Decoupled FineTuning (DeFT) for domain generalization as a solution. DeFT operates in two stages. Its first stage warms up the decoder with the frozen, pretrained encoder so that the decoder learns task-relevant knowledge while the encoder preserves its generalizable features. In the second stage, it decouples finetuning of the encoder and decoder into two pathways, each of which concatenates an adaptive component (AC) and retentive component (RC); the encoder and decoder play different roles between AC and RC in different pathways. ACs are updated by gradients of the loss on the source domain, while RCs are updated by exponential moving average biased toward their initialization to retain their generalization capability. By the two separate optimization pathways with opposite AC-RC configurations, DeFT reduces the number of learnable parameters virtually, and decreases the distance between learned parameters and their initialization, leading to improved generalization capability. DeFT significantly outperformed existing methods in various domain shift scenarios, and its performance was further boosted by incorporating a simple distance regularization.
****
Guanzheng Chen,Xin Li,Michael Shieh,Lidong Bing
LongPO: Long Context Self-Evolution of Large Language Models through Short-to-Long Preference Optimization
Large Language Models (LLMs) have demonstrated remarkable capabilities through pretraining and alignment. However, superior short-context LLMs may underperform in long-context scenarios due to insufficient long-context alignment. This alignment process remains challenging due to the impracticality of human annotation for extended contexts and the difficulty in balancing short- and long-context performance. To address these challenges, we introduce LongPO, that enables short-context LLMs to self-evolve to excel on long-context tasks by internally transferring short-context capabilities. LongPO harnesses LLMs to learn from self-generated short-to-long preference data, comprising paired responses generated for identical instructions with long-context inputs and their compressed short-context counterparts, respectively. This preference reveals capabilities and potentials of LLMs cultivated during short-context alignment that may be diminished in under-aligned long-context scenarios. Additionally, LongPO incorporates a short-to-long KL constraint to mitigate short-context performance decline during long-context alignment. When applied to Mistral-7B-Instruct-v0.2 from 128K to 512K context lengths, LongPO fully retains short-context performance and largely outperforms naive SFT and DPO in both long- and short-context tasks. Specifically, LongPO-trained models can achieve results on long-context benchmarks comparable to, or even surpassing, those of superior LLMs (e.g., GPT-4-128K) that involve extensive long-context annotation and larger parameter scales. Our code is available at https://github.com/DAMO-NLP-SG/LongPO.
****
Can Zhang,Gim Hee Lee
econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians
The primary focus of most recent works on open-vocabulary neural fields is extracting precise semantic features
from the VLMs and then consolidating them efficiently into a multi-view consistent 3D neural fields
representation. However, most existing works over-trusted SAM to regularize image-level CLIP without any further refinement. Moreover, several existing works improved efficiency by dimensionality reduction of semantic features from 2D VLMs before fusing with 3DGS semantic fields, which inevitably leads to multi-view inconsistency. In this work, we propose econSG for open-vocabulary semantic segmentation with 3DGS. Our econSG consists of: 1) A Confidence-region Guided Regularization (CRR) that mutually refines SAM and CLIP to get the best of both worlds f

or precise semantic features with complete and precise boundaries. 2) A low dimensional contextual space to enforce 3D multi-view consistency while improving computational efficiency by fusing backprojected multi-view 2D features and follow by dimensional reduction directly on the fused 3D features instead of operating on each 2D view separately. Our econSG show state-of-the-art performance on four benchmark datasets compared to the existing methods. Furthermore, we are also the most efficient training among all the methods.
****
Yangtao Chen,Zixuan Chen,Junhui Yin,Jing Huo,Pinzhuo Tian,Jieqi Shi,Yang Gao
GravMAD: Grounded Spatial Value Maps Guided Action Diffusion for Generalized 3D Manipulation
Robots' ability to follow language instructions and execute diverse 3D manipulation tasks is vital in robot learning. Traditional imitation learning-based methods perform well on seen tasks but struggle with novel, unseen ones due to variability. Recent approaches leverage large foundation models to assist in understanding novel tasks, thereby mitigating this issue. However, these methods lack a task-specific learning process, which is essential for an accurate understanding of 3D environments, often leading to execution failures. In this paper, we introduce GravMAD, a sub-goal-driven, language-conditioned action diffusion framework that combines the strengths of imitation learning and foundation models. Our approach breaks tasks into sub-goals based on language instructions, allowing auxiliary guidance during both training and inference. During training, we introduce Sub-goal Keypose Discovery to identify key sub-goals from demonstrations. Inference differs from training, as there are no demonstrations available, so we use pre-trained foundation models to bridge the gap and identify sub-goals for the current task. In both phases, GravMaps are generated from sub-goals, providing GravMAD with more flexible 3D spatial guidance compared to fixed 3D positions. Empirical evaluations on RLBench show that GravMAD significantly outperforms state-of-the-art methods, with a 28.63\% improvement on novel tasks and a 13.36\% gain on tasks encountered during training. Evaluations on real-world robotic tasks further show that GravMAD can reason about real-world tasks, associate them with relevant visual information, and generalize to novel tasks. These results demonstrate GravMAD's strong multi-task learning and generalization in 3D manipulation. Video demonstrations are available at: https://gravmad.github.io.
****
Peiwen Sun,Sitong Cheng,Xiangtai Li,Zhen Ye,Huadai Liu,Honggang Zhang,Wei Xue,Yike Guo
Both Ears Wide Open: Towards Language-Driven Spatial Audio Generation
Recently, diffusion models have achieved great success in mono-channel audio generation.
However, when it comes to stereo audio generation, the soundscapes often have a complex scene of multiple objects and directions.
Controlling stereo audio with spatial contexts remains challenging due to high data costs and unstable generative models.
To the best of our knowledge, this work represents the first attempt to address these issues.
We first construct a large-scale, simulation-based, and GPT-assisted dataset, BEWO-1M, with abundant soundscapes and descriptions even including moving and multiple sources.
Beyond text modality, we have also acquired a set of images and rationally paired stereo audios through retrieval to advance multimodal generation.
Existing audio generation models tend to generate rather random and indistinct spatial audio.
To provide accurate guidance for Latent Diffusion Models, we introduce the SpatialSonic model utilizing spatial-aware encoders and azimuth state matrices to reveal reasonable spatial guidance.
By leveraging spatial guidance, our model not only achieves the objective of generating immersive and controllable spatial audio from text but also extends to other modalities as the pioneer attempt.
Finally, under fair settings, we conduct subjective and objective evaluations on

simulated and real-world data to compare our approach with prevailing methods. The results demonstrate the effectiveness of our method, highlighting its capability to generate spatial audio that adheres to physical rules.
****

Johannes Kaiser,Kristian Schwethelm,Daniel Rueckert,Georgios Kaissis
Laplace Sample Information: Data Informativeness Through a Bayesian Lens
Accurately estimating the informativeness of individual samples in a dataset is an important objective in deep learning, as it can guide sample selection, which can improve model efficiency and accuracy by removing redundant or potentially harmful samples.
We propose $\text{\textit{Laplace Sample Information}}$ (LSI) measure of sample informativeness grounded in information theory widely applicable across model architectures and learning settings.
LSI leverages a Bayesian approximation to the weight posterior and the KL divergence to measure the change in the parameter distribution induced by a sample of interest from the dataset.
We experimentally show that LSI is effective in ordering the data with respect to typicality, detecting mislabeled samples, measuring class-wise informativeness, and assessing dataset difficulty.
We demonstrate these capabilities of LSI on image and text data in supervised and unsupervised settings.
Moreover, we show that LSI can be computed efficiently through probes and transfers well to the training of large models.
****

Irtaza Khalid,Steven Schockaert
Systematic Relational Reasoning With Epistemic Graph Neural Networks
Developing models that can learn to reason is a notoriously challenging problem. We focus on reasoning in relational domains, where the use of Graph Neural Networks (GNNs) seems like a natural choice. However, previous work has shown that regular GNNs lack the ability to systematically generalize from training examples on test graphs requiring longer inference chains, which fundamentally limits their reasoning abilities. A common solution relies on neuro-symbolic methods that systematically reason by learning rules, but their scalability is often limited and they tend to make unrealistically strong assumptions, e.g.\ that the answer can always be inferred from a single relational path. We propose the Epistemic GNN (EpiGNN), a novel parameter-efficient and scalable GNN architecture with an epistemic inductive bias for systematic reasoning. Node embeddings in EpiGNNs are treated as epistemic states, and message passing is implemented accordingly. We show that EpiGNNs achieve state-of-the-art results on link prediction tasks that require systematic reasoning. Furthermore, for inductive knowledge graph completion, EpiGNNs rival the performance of state-of-the-art specialized approaches. Finally, we introduce two new benchmarks that go beyond standard relational reasoning by requiring the aggregation of information from multiple paths. Here, existing neuro-symbolic approaches fail, yet EpiGNNs learn to reason accurately. Code and datasets are available at https://github.com/erg0dic/gnn-sg.
****

Bhaskar Mukhoty,Hilal AlQuabeh,Bin Gu
Improving Generalization and Robustness in SNNs Through Signed Rate Encoding and Sparse Encoding Attacks
Rate-encoded spiking neural networks (SNNs) are known to offer superior adversarial robustness compared to direct-encoded SNNs but have relatively poor generalization on clean input. While the latter offers good generalization on clean input it suffers poor adversarial robustness under standard training. A key reason for this difference is the input noise introduced by the rate encoding, which encodes a pixel intensity with T independent Bernoulli samples. To improve the generalization of rate-encoded SNNs, we propose the *signed rate encoding* (sRATE) that allows mean centering of the input and helps reduce the randomness introduced by the encoding, resulting in improved clean accuracy. In contrast to rate encoding, where input restricted to $[0,1]^d$ is encoded in $\\{0,1\\}^{d\times T}$, the signed rate encoding allows input in $[-1,1]^d$ to be encoded with spike

s in $\\{-1,0,1\\}^{d\times T}$, where positive (negative) inputs are encoded wi
th positive (negative) spikes. We further construct efficient \textit{Sparse Enc
oding Attack} (SEA) on standard and signed rate encoded input, which performs $l
_0$-norm restricted adversarial attack in the discrete encoding space. We prove
the theoretical optimality of the attack under the first-order approximation of
the loss and compare it empirically with the existing attacks on the input space
. Adversarial training performed with SEA, under signed rate encoding, offers su
perior adversarial robustness to the existing attacks and itself. Experiments co
nducted on standard datasets show the effectiveness of sign rate encoding in imp
roving accuracy across all settings including adversarial robustness. The code i
s available at https://github.com/BhaskarMukhoty/SignedRateEncoding
****

Alejandro Newell,Peiyun Hu,Lahav Lipson,Stephan Richter,Vladlen Koltun
CoMotion: Concurrent Multi-person 3D Motion
We introduce an approach for detecting and tracking detailed 3D poses of multipl
e people from a single monocular camera stream. Our system maintains temporally
coherent predictions in crowded scenes filled with difficult poses and occlusion
s. Our model performs both strong per-frame detection and a learned pose update
to track people from frame to frame. Rather than match detections across time, p
oses are updated directly from a new input image, which enables online tracking
through occlusion. We train on numerous image and video datasets leveraging pseu
do-labeled annotations to produce a model that matches state-of-the-art systems
in 3D pose estimation accuracy while being faster and more accurate in tracking
multiple people through time.
****

Levi Rauchwerger,Stefanie Jegelka,Ron Levie
Generalization, Expressivity, and Universality of Graph Neural Networks on Attri
buted Graphs
We analyze the universality and generalization of graph neural networks (GNNs) o
n attributed graphs, i.e., with node attributes. To this end, we propose pseudom
etrics over the space of all attributed graphs that describe the fine-grained ex
pressivity of GNNs. Namely, GNNs are both Lipschitz continuous with respect to o
ur pseudometrics and can separate attributed graphs that are distant in the metr
ic. Moreover, we prove that the space of all attributed graphs is relatively com
pact with respect to our metrics. Based on these properties, we prove a universa
l approximation theorem for GNNs and generalization bounds for GNNs on any data
distribution of attributed graphs. The proposed metrics compute the similarity b
etween the structures of attributed graphs via a hierarchical optimal transport
 between computation trees. Our work extends and unites previous approaches whic
h either derived theory only for graphs with no attributes, derived compact metr
ics under which GNNs are continuous but without separation power, or derived met
rics under which GNNs are continuous and separate points but the space of graphs
 is not relatively compact, which prevents universal approximation and generaliz
ation analysis.
****

Chejian Xu,Jiawei Zhang,Zhaorun Chen,Chulin Xie,Mintong Kang,Yujin Potter,Zhun W
ang,Zhuowen Yuan,Alexander Xiong,Zidi Xiong,Chenhui Zhang,Lingzhi Yuan,Yi Zeng,P
eiyang Xu,Chengquan Guo,Andy Zhou,Jeffrey Ziwei Tan,Xuandong Zhao,Francesco Pint
o,Zhen Xiang,Yu Gai,Zinan Lin,Dan Hendrycks,Bo Li,Dawn Song
MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models
Multimodal foundation models (MMFMs) play a crucial role in various applications
, including autonomous driving, healthcare, and virtual assistants. However, sev
eral studies have revealed vulnerabilities in these models, such as generating u
nsafe content by text-to-image models. Existing benchmarks on multimodal models
either predominantly assess the helpfulness of these models, or only focus on li
mited perspectives such as fairness and privacy. In this paper, we present the f
irst unified platform, MMDT (Multimodal DecodingTrust), designed to provide a co
mprehensive safety and trustworthiness evaluation for MMFMs. Our platform assess
es models from multiple perspectives, including safety, hallucination, fairness/
bias, privacy, adversarial robustness, and out-of-distribution (OOD) generalizat

ion. We have designed various evaluation scenarios and red teaming algorithms under different tasks for each perspective to generate challenging data, forming a high-quality benchmark. We evaluate a range of multimodal models using MMDT, and our findings reveal a series of vulnerabilities and areas for improvement across these perspectives. This work introduces the first comprehensive and unique safety and trustworthiness evaluation platform for MMFMs, paving the way for developing safer and more reliable MMFMs and systems. Our platform and benchmark are available at https://mmdecodingtrust.github.io/.
****
Juyeon Heo,Christina Heinze-Deml,Oussama Elachqar,Kwan Ho Ryan Chan,Shirley You Ren,Andrew Miller,Udhyakumar Nallasamy,Jaya Narain
Do LLMs ``know'' internally when they follow instructions?
Instruction-following is crucial for building AI agents with large language models (LLMs), as these models must adhere strictly to user-provided constraints and guidelines.
However, LLMs often fail to follow even simple and clear instructions.
To improve instruction-following behavior and prevent undesirable outputs, a deeper understanding of how LLMs' internal states relate to these outcomes is required.
In this work, we investigate whether LLMs encode information in their representations that correlates with instruction-following success—a property we term ``knowing internally''.
Our analysis identifies a direction in the input embedding space, termed the instruction-following dimension, that predicts whether a response will comply with a given instruction.
We find that this dimension generalizes well across unseen tasks but not across unseen instruction types.
We demonstrate that modifying representations along this dimension improves instruction-following success rates compared to random changes, without compromising response quality.
Further investigation reveals that this dimension is more closely related to the phrasing of prompts rather than the inherent difficulty of the task or instructions.
This discovery also suggests explanations for why LLMs sometimes fail to follow clear instructions and why prompt engineering is often effective, even when the content remains largely unchanged.
This work provides insight into the internal workings of LLMs' instruction-following, paving the way for reliable LLM agents.
****
Anh Khoa Nguyen Vu,Truong Quoc Truong,Vinh-Tiep Nguyen,Thanh Duc Ngo,Thanh-Toan Do,Tam V. Nguyen
Multi-Perspective Data Augmentation for Few-shot Object Detection
Recent few-shot object detection (FSOD) methods have focused on augmenting synthetic samples for novel classes, show promising results to the rise of diffusion models. However, the diversity of such datasets is often limited in representativeness because they lack awareness of typical and hard samples, especially in the context of foreground and background relationships. To tackle this issue, we propose a Multi-Perspective Data Augmentation (MPAD) framework. In terms of foreground-foreground relationships, we propose in-context learning for object synthesis (ICOS) with bounding box adjustments to enhance the detail and spatial information of synthetic samples. Inspired by the large margin principle, support samples play a vital role in defining class boundaries. Therefore, we design a Harmonic Prompt Aggregation Scheduler (HPAS) to mix prompt embeddings at each time step of the generation process in diffusion models, producing hard novel samples. For foreground-background relationships, we introduce a Background Proposal method (BAP) to sample typical and hard backgrounds. Extensive experiments on multiple FSOD benchmarks demonstrate the effectiveness of our approach. Our framework significantly outperforms traditional methods, achieving an average increase of 17.5% in nAP50 over the baseline on PASCAL VOC.
****

Linus Bao,Emily Jin,Michael M. Bronstein,Ismail Ilkan Ceylan,Matthias Lanzinger
Homomorphism Counts as Structural Encodings for Graph Learning
Graph Transformers are popular neural networks that extend the well-known Transformer architecture to the graph domain. These architectures operate by applying self-attention on graph nodes and incorporating graph structure through the use of positional encodings (e.g., Laplacian positional encoding) or structural encodings (e.g., random-walk structural encoding). The quality of such encodings is critical, since they provide the necessary \emph{graph inductive biases} to condition the model on graph structure. In this work, we propose \emph{motif structural encoding} (MoSE) as a flexible and powerful structural encoding framework based on counting graph homomorphisms. Theoretically, we compare the expressive power of MoSE to random-walk structural encoding and relate both encodings to the expressive power of standard message passing neural networks. Empirically, we observe that MoSE outperforms other well-known positional and structural encodings across a range of architectures, and it achieves state-of-the-art performance on a widely studied molecular property prediction dataset.
****
Raul Fernandez-Diaz,Hoang Thanh Lam,Vanessa López,Denis C. Shields
A new framework for evaluating model out-of-distribution generalisation for the biochemical domain
Quantifying model generalization to out-of-distribution data has been a longstanding challenge in machine learning. Addressing this issue is crucial for leveraging machine learning in scientific discovery, where models must generalize to new molecules or materials. Current methods typically split data into train and test sets using various criteria — temporal, sequence identity, scaffold, or random cross-validation — before evaluating model performance. However, with so many splitting criteria available, existing approaches offer limited guidance on selecting the most appropriate one, and they do not provide mechanisms for incorporating prior knowledge about the target deployment distribution(s).

To tackle this problem, we have developed a novel metric, AU-GOOD, which quantifies expected model performance under conditions of increasing dissimilarity between train and test sets, while also accounting for prior knowledge about the target deployment distribution(s), when available. This metric is broadly applicable to biochemical entities, including proteins, small molecules, nucleic acids, or cells; as long as a relevant similarity function is defined for them. Recognizing the wide range of similarity functions used in biochemistry, we propose criteria to guide the selection of the most appropriate metric for partitioning. We also introduce a new partitioning algorithm that generates more challenging test sets, and we propose statistical methods for comparing models based on AU-GOOD.

Finally, we demonstrate the insights that can be gained from this framework by applying it to two different use cases: developing predictors for pharmaceutical properties of small molecules, and using protein language models as embeddings to build biophysical property predictors.
****
Klas Wijk,Ricardo Vinuesa,Hossein Azizpour
SFESS: Score Function Estimators for k-Subset Sampling
Are score function estimators a viable approach to learning with k-subset sampling? Sampling k-subsets is a fundamental operation that is not amenable to differentiable parametrization, impeding gradient-based optimization. Previous work has favored approximate pathwise gradients or relaxed sampling, dismissing score function estimators because of their high variance. Inspired by the success of score function estimators in variational inference and reinforcement learning, we revisit them for k-subset sampling. We demonstrate how to efficiently compute the distribution's score function using a discrete Fourier transform and reduce the estimator's variance with control variates. The resulting estimator provides both k-hot samples and unbiased gradient estimates while being applicable to non-differentiable downstream models, unlike existing methods. We validate our approach experimentally and find that it produces results comparable to those

of recent state-of-the-art pathwise gradient estimators across a range of tasks
.
****
Pegah Golestaneh,Mahsa Taheri,Johannes Lederer
How many samples are needed to train a deep neural network?
Even though neural networks have become standard tools in many areas, many impor
tant statistical questions remain open. This paper studies the question of how m
uch data are needed to train a ReLU feed-forward neural network. Our theoretical
and empirical results suggest that the generalization error of ReLU feed-forwar
d neural networks scales at the rate $1/\sqrt{n}$ in the sample size n-rather
than the "parametric rate" $1/n$, which might be suggested by traditional stati
stical theories. Thus, broadly speaking, our results underpin the common belief
that neural networks need "many" training samples. Along the way, we also establ
ish new technical insights, such as the first lower bounds of the entropy of ReL
U feed-forward networks.
****
Xianyang Zhang,Huijuan Zhou
Generalization Bounds and Model Complexity for Kolmogorov–Arnold Networks
Kolmogorov–Arnold Network (KAN) is a network structure recently proposed in Liu
et al. (2024) that offers improved interpretability and a more parsimonious desi
gn in many science-oriented tasks compared to multi-layer perceptrons. This work
provides a rigorous theoretical analysis of KAN by establishing generalization
bounds for KAN equipped with activation functions that are either represented by
linear combinations of basis functions or lying in a low-rank Reproducing Kerne
l Hilbert Space (RKHS). In the first case, the generalization bound accommodates
various choices of basis functions in forming the activation functions in each
layer of KAN and is adapted to different operator norms at each layer. For a par
ticular choice of operator norms, the bound scales with the l_1 norm of the co
efficient matrices and the Lipschitz constants for the activation functions, and
it has no dependence on combinatorial parameters (e.g., number of nodes) outsid
e of logarithmic factors. Moreover, our result does not require the boundedness
assumption on the loss function and, hence, is applicable to a general class of
regression-type loss functions. In the low-rank case, the generalization bound s
cales polynomially with the underlying ranks as well as the Lipschitz constants
of the activation functions in each layer. These bounds are empirically investig
ated for KANs trained with stochastic gradient descent on simulated and real dat
a sets. The numerical results demonstrate the practical relevance of these bound
s.
****
Haotian Tang,Yecheng Wu,Shang Yang,Enze Xie,Junsong Chen,Junyu Chen,Zhuoyang Zha
ng,Han Cai,Yao Lu,Song Han
HART: Efficient Visual Generation with Hybrid Autoregressive Transformer
We introduce Hybrid Autoregressive Transformer (HART), the first autoregressive
(AR) visual generation model capable of directly generating 1024x1024 images, ri
valing diffusion models in image generation quality. Existing AR models face lim
itations due to the poor image reconstruction quality of their discrete tokenize
rs and the prohibitive training costs associated with generating 1024px images.
To address these challenges, we present the hybrid tokenizer, which decomposes t
he continuous latents from the autoencoder into two components: discrete tokens
representing the big picture and continuous tokens representing the residual co
mponents that cannot be represented by the discrete tokens. The discrete compone
nt is modeled by a scalable-resolution discrete AR model, while the continuous c
omponent is learned with a lightweight residual diffusion module with only 37M p
arameters. Compared with the discrete-only VAR tokenizer, our hybrid approach im
proves reconstruction FID from 2.11 to 0.30 on MJHQ-30K, leading to a 31% genera
tion FID improvement from 7.85 to 5.38. HART also outperforms state-of-the-art d
iffusion models in both FID and CLIP score, with 4.5-7.7$\times$ higher throughp
ut and 6.9-13.4$\times$ lower MACs. Our code is open sourced at https://github.c
om/mit-han-lab/hart.
****

Nilay Yilmaz,Maitreya Patel,Yiran Lawrence Luo,Tejas Gokhale,Chitta Baral,Suren Jayasuriya,Yezhou Yang
VOILA: Evaluation of MLLMs For Perceptual Understanding and Analogical Reasoning
Multimodal Large Language Models (MLLMs) have become a powerful tool for integrating visual and textual information. Despite their exceptional performance on visual understanding benchmarks, measuring their ability to reason abstractly across multiple images remains a significant challenge. To address this, we introduce VOILA, a large-scale, open-ended, dynamic benchmark designed to evaluate MLLMs' perceptual understanding and abstract relational reasoning. VOILA employs an analogical mapping approach in the visual domain, requiring models to generate an image that completes an analogy between two given image pairs, reference and application, without relying on predefined choices. Our experiments demonstrate that the analogical reasoning tasks in VOILA present a challenge to MLLMs. Through multi-step analysis, we reveal that current MLLMs struggle to comprehend inter-image relationships and exhibit limited capabilities in high-level relational reasoning. Notably, we observe that performance improves when following a multi-step strategy of least-to-most prompting. Comprehensive evaluations on open-source models and GPT-4o show that on text-based answers, the best accuracy for challenging scenarios is 13% (LLaMa 3.2) and even for simpler tasks is only 29% (GPT-4o), while human performance is significantly higher at 70% across both difficulty levels.
****
Yaowen Ye,Cassidy Laidlaw,Jacob Steinhardt
Iterative Label Refinement Matters More than Preference Optimization under Weak Supervision
Language model (LM) post-training relies on two stages of human supervision: task demonstrations for supervised finetuning (SFT), followed by preference comparisons for reinforcement learning from human feedback (RLHF). As LMs become more capable, the tasks they are given become harder to supervise. Will post-training remain effective under unreliable supervision? To test this, we simulate unreliable demonstrations and comparison feedback using small LMs and time-constrained humans. We find that in the presence of unreliable supervision, SFT still retains some effectiveness, but DPO (a common RLHF algorithm) fails to improve the model beyond SFT. To address this, we propose *iterative label refinement* (ILR) as an alternative to RLHF. ILR improves the SFT data by using comparison feedback to decide whether human demonstrations should be replaced by model-generated alternatives, then retrains the model via SFT on the updated data. SFT+ILR outperforms SFT+DPO on several tasks with unreliable supervision (math, coding, and safe instruction-following). Our findings suggest that as LMs are used for complex tasks where human supervision is unreliable, RLHF may no longer be the best use of human comparison feedback; instead, it is better to direct feedback towards improving the training *data* rather than continually training the *model*. Our code and data are available at https://github.com/helloelwin/iterative-label-refinement.
****
Yaxin Luo,Gen Luo,Jiayi Ji,Yiyi Zhou,Xiaoshuai Sun,Zhiqiang Shen,Rongrong Ji
$\gamma-$MoD: Exploring Mixture-of-Depth Adaptation for Multimodal Large Language Models
Despite the significant progress in multimodal large language models (MLLMs), their high computational cost remains a barrier to real-world deployment. Inspired by the mixture of depths (MoDs) in natural language processing, we aim to address this limitation from the perspective of ``activated tokens''. Our key insight is that if most tokens are redundant for the layer computation, then can be skipped directly via the MoD layer. However, directly converting the dense layers of MLLMs to MoD layers leads to substantial performance degradation. To address this issue, we propose an innovative MoD adaptation strategy for existing MLLMs called γ-MoD. In γ-MoD, a novel metric is proposed to guide the deployment of MoDs in the MLLM, namely rank of attention maps (ARank). Through ARank, we can effectively identify which layer is redundant and should be replaced with the MoD layer. Based on ARank, we further propose two novel designs to

maximize the computational sparsity of MLLM while maintaining its performance, namely shared vision-language router and masked routing learning. With these designs, more than 90% dense layers of the MLLM can be effectively converted to the MoD ones. To validate our method, we apply it to three popular MLLMs, and conduct extensive experiments on 9 benchmark datasets. Experimental results not only validate the significant efficiency benefit of γ-MoD to existing MLLMs but also confirm its generalization ability on various MLLMs. For example, with a minor performance drop, i.e., -1.5%, γ-MoD can reduce the training and inference time of LLaVA-HR by 31.0% and 53.2%, respectively.
****

Yushi Guan,Daniel Kwan,Jean Sebastien Dandurand,Xi Yan,Ruofan Liang,Yuxuan Zhang,Nilesh Jain,Nilesh Ahuja,Selvakumar Panneer,Nandita Vijaykumar
Retri3D: 3D Neural Graphics Representation Retrieval
Learnable 3D Neural Graphics Representations (3DNGR) have emerged as promising 3D representations for reconstructing 3D scenes from 2D images. Numerous works, including Neural Radiance Fields (NeRF), 3D Gaussian Splatting (3DGS), and their variants, have significantly enhanced the quality of these representations. The ease of construction from 2D images, suitability for online viewing/sharing, and applications in game/art design downstream tasks make it a vital 3D representation, with potential creation of large numbers of such 3D models. This necessitates large data stores, local or online, to save 3D visual data in these formats. However, no existing framework enables accurate retrieval of stored 3DNGRs. In this work, we propose, Retri3D, a framework that enables accurate and efficient retrieval of 3D scenes represented as NGRs from large data stores using text queries. We introduce a novel Neural Field Artifact Analysis technique, combined with a Smart Camera Movement Module, to select clean views and navigate pre-trained 3DNGRs. These techniques enable accurate retrieval by selecting the best viewing directions in the 3D scene for high-quality visual feature embeddings. We demonstrate that Retri3D is compatible with any NGR representation. On the LERF and ScanNet++ datasets, we show significant improvement in retrieval accuracy compared to existing techniques, while being orders of magnitude faster and storage efficient.
****

Zhixuan Lin,Evgenii Nikishin,Xu He,Aaron Courville
Forgetting Transformer: Softmax Attention with a Forget Gate
An essential component of modern recurrent sequence models is the forget gate. While Transformers do not have an explicit recurrent form, we show that a forget gate can be naturally incorporated into Transformers by down-weighting the unnormalized attention scores in a data-dependent way. We name this attention mechanism the Forgetting Attention and the resulting model the Forgetting Transformer (FoX). We show that FoX outperforms the Transformer on long-context language modeling, length extrapolation, and short-context downstream tasks, while performing on par with the Transformer on long-context downstream tasks. Moreover, it is compatible with the FlashAttention algorithm and does not require any positional embeddings. Several analyses, including the needle-in-the-haystack test, show that FoX also retains the Transformer's superior long-context capabilities over recurrent sequence models such as Mamba-2, HGRN2, and DeltaNet. We also introduce a ``Pro'' block design that incorporates some common architectural components in recurrent sequence models and find it significantly improves the performance of both FoX and the Transformer.
Our code is available at [`https://github.com/zhixuan-lin/forgetting-transformer`](https://github.com/zhixuan-lin/forgetting-transformer).
****

Jian Wu,Linyi Yang,Zhen Wang,Manabu Okumura,Yue Zhang
CofCA: A STEP-WISE Counterfactual Multi-hop QA benchmark
While Large Language Models (LLMs) excel in question-answering (QA) tasks, their real reasoning abilities on multiple evidence retrieval and integration on Multi-hop QA tasks remain less explored. Firstly, LLMs sometimes generate answers that rely on internal memory rather than retrieving evidence and reasoning in the given context, which brings concerns about the evaluation quality of real reason

ing abilities. Although previous counterfactual QA benchmarks can separate the i
nternal memory of LLMs, they focus solely on final QA performance, which is insu
fficient for reporting LLMs' real reasoning abilities. Because LLMs are expected
 to engage in intricate reasoning processes that involve evidence retrieval and
answering a series of sub-questions from given passages. Moreover, current factu
al Multi-hop QA (MHQA) benchmarks are annotated on open-source corpora such as W
ikipedia, although useful for multi-step reasoning evaluation, they show limitat
ions due to the potential data contamination in LLMs' pre-training stage. To add
ress these issues, we introduce the Step-wise and Counterfactual benchmark (CofC
A), a novel evaluation benchmark consisting of factual data and counterfactual d
ata that reveals LLMs' real reasoning abilities on multi-step reasoning and reas
oning chain evaluation. Our experimental results reveal a significant performanc
e gap of several LLMs between Wikipedia-based factual data and counterfactual da
ta, deeming data contamination issues in existing benchmarks. Moreover, we obser
ve that LLMs usually bypass the correct reasoning chain, showing an inflated mul
ti-step reasoning performance. We believe that our CofCA benchmark will enhance
and facilitate the evaluations of trustworthy LLMs.
****

Ziheng Chen,Yue Song,Xiaojun Wu,Gaowen Liu,Nicu Sebe
Understanding Matrix Function Normalizations in Covariance Pooling through the L
ens of Riemannian Geometry
Global Covariance Pooling (GCP) has been demonstrated to improve the performance
 of Deep Neural Networks (DNNs) by exploiting second-order statistics of high-le
vel representations. GCP typically performs classification of the covariance mat
rices by applying matrix function normalization, such as matrix logarithm or pow
er, followed by a Euclidean classifier. However, covariance matrices inherently
lie in a Riemannian manifold, known as the Symmetric Positive Definite (SPD) man
ifold. The current literature does not provide a satisfactory explanation of why
 Euclidean classifiers can be applied directly to Riemannian features after the
normalization of the matrix power. To mitigate this gap, this paper provides a c
omprehensive and unified understanding of the matrix logarithm and power from a
Riemannian geometry perspective. The underlying mechanism of matrix functions in
 GCP is interpreted from two perspectives: one based on tangent classifiers (Euc
lidean classifiers on the tangent space) and the other based on Riemannian class
ifiers. Via theoretical analysis and empirical validation through extensive expe
riments on fine-grained and large-scale visual classification datasets, we concl
ude that the working mechanism of the matrix functions should be attributed to t
he Riemannian classifiers they implicitly respect. The code is available at http
s://github.com/GitZH-Chen/RiemGCP.git.
****

Lizhe Fang,Yifei Wang,Khashayar Gatmiry,Lei Fang,Yisen Wang
Rethinking Invariance in In-context Learning
In-Context Learning (ICL) has emerged as a pivotal capability of auto-regressive
 large language models, yet it is hindered by a notable sensitivity to the order
ing of context examples regardless of their mutual independence. To address this
 issue, recent studies have introduced several variant algorithms of ICL that ac
hieve permutation invariance. However, many of these do not exhibit comparable p
erformance with the standard auto-regressive ICL algorithm. In this work, we ide
ntify two crucial elements in the design of an invariant ICL algorithm: informat
ion non-leakage and context interdependence, which are not simultaneously achiev
ed by any of the existing methods. These investigations lead us to the proposed
\emph{Invariant ICL (InvICL)}, a methodology designed to achieve invariance in I
CL while ensuring the two properties. Empirically, our findings reveal that InvI
CL surpasses previous models, both invariant and non-invariant, in most benchmar
k datasets, showcasing superior generalization capabilities across varying input
 lengths. Code is available at https://github.com/PKU-ML/InvICL.
****

Peihao Wang,Ruisi Cai,Yuehao Wang,Jiajun Zhu,Pragya Srivastava,Zhangyang Wang,Pa
n Li
Understanding and Mitigating Bottlenecks of State Space Models through the Lens

of Recency and Over-smoothing
Structured State Space Models (SSMs) have emerged as alternatives to transformer s.
While SSMs are often regarded as effective in capturing long-sequence dependenci es, we rigorously demonstrate that they are inherently limited by strong recency bias.
Our empirical studies also reveal that this bias impairs the models' ability to recall distant information and introduces robustness issues. Our scaling experim ents then discovered that deeper structures in SSMs can facilitate the learning of long contexts.
However, subsequent theoretical analysis reveals that as SSMs increase in depth, they exhibit another inevitable tendency toward over-smoothing, e.g., token rep resentations becoming increasingly indistinguishable.
This *fundamental dilemma* between recency and over-smoothing hinders the scalab ility of existing SSMs.
Inspired by our theoretical findings, we propose to *polarize* two channels of t he state transition matrices in SSMs, setting them to zero and one, respectively , simultaneously addressing recency bias and over-smoothing.
Experiments demonstrate that our polarization technique consistently enhances th e associative recall accuracy of long-range tokens and unlocks SSMs to benefit f urther from deeper architectures.
All source codes are released at https://github.com/VITA-Group/SSM-Bottleneck.
****
Xu-Hui Liu,Yali Du,Jun Wang,Yang Yu
On the Optimization Landscape of Low Rank Adaptation Methods for Large Language Models
Training Large Language Models (LLMs) poses significant memory challenges, makin g low-rank adaptation methods an attractive solution. Previously, Low-Rank Adapt ation (LoRA) addressed this by adding a trainable low-rank matrix to the frozen pre-trained weights in each layer, reducing the number of trainable parameters a nd optimizer states. GaLore, which compresses the gradient matrix instead of the weight matrix, has demonstrated superior performance to LoRA with faster conver gence and reduced memory consumption. Despite their empirical success, the perfo rmance of these methods has not been fully understood or explained theoretically . In this paper, we analyze the optimization landscapes of LoRA, GaLore, and ful l-rank methods, revealing that GaLore benefits from fewer spurious local minima and a larger region that satisfies the \pl, a variant of Polyak-■ojasiewicz (PL) condition, leading to faster convergence. Our analysis leads to a novel method, GaRare, which further improves GaLore by using gradient random projection to re duce computational overhead. Practically, GaRare achieves strong performance in both pre-training and fine-tuning tasks, offering a more efficient approach to l arge-scale model adaptation.
****
Jiahui Gao,Renjie Pi,Jipeng Zhang,Jiacheng Ye,Wanjun Zhong,Yufei Wang,Lanqing HO NG,Jianhua Han,Hang Xu,Zhenguo Li,Lingpeng Kong
G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model
Large language models (LLMs) have shown remarkable proficiency in human-level re asoning and generation capabilities, which encourages extensive research on thei r application in mathematical problem solving. However, current work has been la rgely focused on text-based mathematical problems, with limited investigation in problems involving multi-modal geometric information. Addressing this gap, we a im to enable LLMs to solve geometric problems by understanding image input. We f irst identify the limitations of current Multimodal Large Language Models (MLLMs) in this area: they struggle to accurately comprehend basic geometric elements and their relationships. To address these challenges, we leverage the inherent a ttribute of logical structure compactness in geometric figures, utilizing text-o nly Large Language Models (LLMs) to curate a comprehensive multimodal geometry d ataset. This dataset, named Geo170k, contains more than 170K geometric image-cap tion and question-answer pairs. Utilizing the Geo170k dataset, we introduce G-LL aVA, a model that demonstrates exceptional performance in solving geometric prob

lems. It significantly outperforms GPT4-V on the geometry task of MathVista benc
hmark with only 7B parameters.
****
Damien MARTINS GOMES,Yanlei Zhang,Eugene Belilovsky,Guy Wolf,Mahdi S. Hosseini
AdaFisher: Adaptive Second Order Optimization via Fisher Information
First-order optimization methods are currently the mainstream in training deep n
eural networks (DNNs). Optimizers like Adam incorporate limited curvature inform
ation by employing the diagonal matrix preconditioning of the stochastic gradien
t during the training. Despite their widespread, second-order optimization algor
ithms exhibit superior convergence properties compared to their first-order coun
terparts e.g. Adam and SGD. However, their practicality in training DNNs is stil
l limited due to increased per-iteration computations compared to the first-orde
r methods. We present *AdaFisher*--an adaptive second-order optimizer that lever
ages a *diagonal block-Kronecker* approximation of the Fisher information matrix
 for adaptive gradient preconditioning. AdaFisher aims to bridge the gap between
 enhanced *convergence/generalization* capabilities and computational efficiency
 in second-order optimization framework for training DNNs. Despite the slow pace
 of second-order optimizers, we showcase that AdaFisher can be reliably adopted
for image classification, language modeling and stands out for its stability and
 robustness in hyper-parameter tuning. We demonstrate that AdaFisher **outperfor
ms the SOTA optimizers** in terms of both accuracy and convergence speed. Code i
s available from https://github.com/AtlasAnalyticsLab/AdaFisher.
****
Yuwei Cheng,Fan Yao,Xuefeng Liu,Haifeng Xu
Learning from Imperfect Human Feedback: A Tale from Corruption-Robust Dueling
This paper studies Learning from Imperfect Human Feedback (LIHF), addressing the
 potential irrationality or imperfect perception when learning from comparative
human feedback. Building on evidences that human's imperfection decays over time
 (i.e., humans learn to improve), we cast this problem as a concave-utility con
tinuous-action dueling bandit but under a restricted form of corruption: i.e.,
 the corruption scale is decaying over time as $t^{\rho-1}$ for some ``imperfect
ion rate'' $\rho \in [0, 1]$.
With T as the total number of iterations, we establish a regret lower bound of
 $ \Omega(\max\{\sqrt{T}, T^{\rho}\})$ for LIHF, even when ρ is known. For
the same setting, we develop the Robustified Stochastic Mirror Descent for Imper
fect Dueling (RoSMID) algorithm, which achieves nearly optimal regret $\tilde{\m
athcal{O}}(\max\{\sqrt{T}, T^{\rho}\})$. Core to our analysis is a novel framewo
rk for analyzing gradient-based algorithms for dueling bandit under corruption,
and we demonstrate its general applicability by showing how this framework can b
e easily applied to obtain corruption-robust guarantees for other popular gradie
nt-based dueling bandit algorithms. Our theoretical results are validated by ext
ensive experiments.
****
Jiabo Ye,Haiyang Xu,Haowei Liu,Anwen Hu,Ming Yan,Qi Qian,Ji Zhang,Fei Huang,Jing
ren Zhou
mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Langu
age Models
Multi-modal Large Language Models have demonstrated remarkable capabilities in e
xecuting instructions for a variety of single-image tasks. Despite this progress
, significant challenges remain in modeling long image sequences. In this work,
we introduce the versatile multi-modal large language model, mPLUG-Owl3, which e
nhances the capability for long image-sequence understanding in scenarios that i
ncorporate retrieved image-text knowledge, multimodal in-context examples, and l
engthy videos. Specifically, we propose novel hyper attention blocks to efficien
tly integrate vision and language into a common language-guided semantic space,
thereby facilitating the processing of extended multi-image scenarios. We conduc
t evaluations on 21 benchmarks that cover single/multi-image, and short/long vid
eo understanding. mPLUG-Owl3 achieves competitive performance with the state-of-
the-art methods while reducing inference time and memory usage by 87.8\% and 48.
5\% in average. Moreover, we propose a Distractor Resistance evaluation to asses

s the ability of models to maintain focus amidst distractions. mPLUG-Owl3 also d
emonstrates outstanding performance in distractor resistance on ultra-long visua
l sequence inputs. We hope that mPLUG-Owl3 can contribute to the development of
more efficient and powerful multimodal large language models.
****

Yanjie Li,Kaisheng Liang,Bin Xiao

UV-Attack: Physical-World Adversarial Attacks on Person Detection via Dynamic-Ne
RF-based UV Mapping

Recent works have attacked person detectors using adversarial patches or static-
3D-model-based texture modifications. However, these methods suffer from low att
ack success rates when faced with significant human movements. The primary chall
enge stems from the highly non-rigid nature of the human body and clothing. Curr
ent attacks fail to model these 3D non-rigid deformations caused by varied actio
ns.

Fortunately, recent research has shown significant progress in using NeRF for dy
namic human modeling.

In this paper, we introduce \texttt{UV-Attack}, a novel physical adversarial att
ack achieving high attack success rates in scenarios involving extensive and uns
een actions. We address the challenges above by leveraging dynamic-NeRF-based UV
 mapping. Our method can generate human images across diverse actions and viewpo
ints and even create novel unseen actions by sampling from the SMPL parameter sp
ace. While dynamic NeRF models are capable of modeling human bodies, modifying t
heir clothing textures is challenging due to the texture being embedded within n
eural network parameters.

To overcome this, \texttt{UV-Attack} generates UV maps instead of RGB images and
 modifies the texture stacks. This approach enables real-time texture edits and
makes attacks more practical. Finally, we propose a novel Expectation over Pose
Transformation loss (EoPT) to improve the evasion success rate on unseen poses a
nd views.

Our experiments show that \texttt{UV-Attack} achieves a 92.7\% attack success ra
te against the FastRCNN model across varied poses in dynamic video settings, sig
nificantly outperforming the state-of-the-art AdvCaT attack, which only had a 28
.5\% ASR. Moreover, we achieve 49.5\% ASR on the latest YOLOv8 detector in black
-box settings. The code is available at https://github.com/PolyLiYJ/UV-Attack
****

Tuo Xu

Towards a Complete Logical Framework for GNN Expressiveness

Designing expressive Graph neural networks (GNNs) is an important topic in graph
 machine learning fields. Traditionally, the Weisfeiler-Lehman (WL) test has bee
n the primary measure for evaluating GNN expressiveness. However, high-order WL
tests can be obscure, making it challenging to discern the specific graph patter
ns captured by them. Given the connection between WL tests and first-order logic
, some have explored the logical expressiveness of Message Passing Neural Networ
ks. This paper aims to establish a comprehensive and systematic relationship bet
ween GNNs and logic. We propose a framework for identifying the equivalent logic
al formulas for arbitrary GNN architectures, which not only explains existing mo
dels, but also provides inspiration for future research. As case studies, we ana
lyze multiple classes of prominent GNNs within this framework, unifying differen
t subareas of the field. Additionally, we conduct a detailed examination of homo
morphism expressivity from a logical perspective and present a general method fo
r determining the homomorphism expressivity of arbitrary GNN models, as well as
addressing several open problems.
****

Zikun Zhang,Zixiang Chen,Quanquan Gu

Convergence of Score-Based Discrete Diffusion Models: A Discrete-Time Analysis

Diffusion models have achieved great success in generating high-dimensional samp
les across various applications. While the theoretical guarantees for continuous
-state diffusion models have been extensively studied, the convergence analysis
of the discrete-state counterparts remains under-explored. In this paper, we stu
dy the theoretical aspects of score-based discrete diffusion models under the Co

ntinuous Time Markov Chain (CTMC) framework. We introduce a discrete-time sampling algorithm in the general state space $[S]^d$ that utilizes score estimators at predefined time points. We derive convergence bounds for the Kullback-Leibler (KL) divergence and total variation (TV) distance between the generated sample distribution and the data distribution, considering both scenarios with and without early stopping under reasonable assumptions. Notably, our KL divergence bounds are nearly linear in the dimension d, aligning with state-of-the-art results for diffusion models. Our convergence analysis employs a Girsanov-based method and establishes key properties of the discrete score function, which are essential for characterizing the discrete-time sampling process.
****

Anton Xue,Avishree Khare,Rajeev Alur,Surbhi Goel,Eric Wong
Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference
We study how to subvert large language models (LLMs) from following prompt-specified rules.
We first formalize rule-following as inference in propositional Horn logic, a mathematical system in which rules have the form "if P and Q, then R" for some propositions P, Q, and R.
Next, we prove that although small transformers can faithfully follow such rules, maliciously crafted prompts can still mislead both theoretical constructions and models learned from data.
Furthermore, we demonstrate that popular attack algorithms on LLMs find adversarial prompts and induce attention patterns that align with our theory.
Our novel logic-based framework provides a foundation for studying LLMs in rule-based settings, enabling a formal analysis of tasks like logical reasoning and jailbreak attacks.
****

Xianming LI,Zongxi Li,Jing Li,Haoran Xie,Qing Li
ESE: Espresso Sentence Embeddings
High-quality sentence embeddings are fundamental in many natural language processing (NLP) tasks, such as semantic textual similarity (STS) and retrieval-augmented generation (RAG). However, most existing methods leverage fixed-length sentence embeddings from full-layer language models, which lack the scalability to accommodate the diverse available resources across various applications. Viewing this gap, we propose a novel sentence embedding model Espresso Sentence Embeddings (ESE) with two learning processes. First, the learn-to-express process encodes more salient representations to shallow layers. Second, the learn-to-compress process compacts essential features into the initial dimensions using Principal Component Analysis (PCA). This way, ESE can scale model depth via the former process and embedding size via the latter. Extensive experiments on STS and RAG suggest that ESE can effectively produce high-quality sentence embeddings with less model depth and embedding size, enhancing inference efficiency. The code is available at https://github.com/SeanLee97/AnglE/blob/main/README_ESE.md.
****

Jiawei Xu,Rui Yang,Shuang Qiu,Feng Luo,Meng Fang,Baoxiang Wang,Lei Han
Tackling Data Corruption in Offline Reinforcement Learning via Sequence Modeling
Learning policy from offline datasets through offline reinforcement learning (RL) holds promise for scaling data-driven decision-making while avoiding unsafe and costly online interactions. However, real-world data collected from sensors or humans often contains noise and errors, posing a significant challenge for existing offline RL methods, particularly when the real-world data is limited. Our study reveals that prior research focusing on adapting predominant offline RL methods based on temporal difference learning still falls short under data corruption when the dataset is limited. In contrast, we discover that vanilla sequence modeling methods, such as Decision Transformer, exhibit robustness against data corruption, even without specialized modifications. To unlock the full potential of sequence modeling, we propose **R**obust **D**ecision **T**ransformer (**RDT**) by incorporating three simple yet effective robust techniques: embedding dropout to improve the model's robustness against erroneous inputs, Gaussian weighted learning to mitigate the effects of corrupted labels, and iterative data corre

ction to eliminate corrupted data from the source.
Extensive experiments on MuJoCo, Kitchen, and Adroit tasks demonstrate RDT's sup erior performance under various data corruption scenarios compared to prior meth ods. Furthermore, RDT exhibits remarkable robustness in a more challenging setti ng that combines training-time data corruption with test-time observation pertur bations. These results highlight the potential of sequence modeling for learning from noisy or corrupted offline datasets, thereby promoting the reliable applic ation of offline RL in real-world scenarios.
Our code is available at https://github.com/jiawei415/RobustDecisionTransformer■
****

Vasiliki Tassopoulou,Haochang Shou,Christos Davatzikos
Adaptive Shrinkage Estimation for Personalized Deep Kernel Regression in Modelin g Brain Trajectories
Longitudinal biomedical studies monitor individuals over time to capture dynamic s in brain development, disease progression, and treatment effects. However, est imating trajectories of brain biomarkers is challenging due to biological variab ility, inconsistencies in measurement protocols (e.g., differences in MRI scanne rs) as well as scarcity and irregularity in longitudinal measurements. Herein, we introduce a novel personalized deep kernel regression framework for forecasti ng brain biomarkers, with application to regional volumetric measurements. Our a pproach integrates two key components: a population model that captures brain tr ajectories from a large and diverse cohort, and a subject-specific model that ca ptures individual trajectories. To optimally combine these, we propose Adaptive Shrinkage Estimation, which effectively balances population and subject-specific models. We assess our model’s performance through predictive accuracy metrics, uncertainty quantification, and validation against external clinical studies. Be nchmarking against state-of-the-art statistical and machine learning models—incl uding linear mixed effects models, generalized additive models, and deep learnin g methods—demonstrates the superior predictive performance of our approach. Addi tionally, we apply our method to predict trajectories of composite neuroimaging biomarkers, which highlights the versatility of our approach in modeling the pro gression of longitudinal neuroimaging biomarkers. Furthermore, validation on thr ee external neuroimaging studies confirms the robustness of our method across di fferent clinical contexts. We make the code available at https://github.com/vata ss/AdaptiveShrinkageDKGP.
****

Sebastian Sanokowski,Wilhelm Franz Berghammer,Haoyu Peter Wang,Martin Ennemoser, Sepp Hochreiter,Sebastian Lehner
Scalable Discrete Diffusion Samplers: Combinatorial Optimization and Statistical Physics
Learning to sample from complex unnormalized distributions over discrete domains emerged as a promising research direction with applications in statistical phys ics, variational inference, and combinatorial optimization. Recent work has demo nstrated the potential of diffusion models in this domain. However, existing met hods face limitations in memory scaling and thus the number of attainable diffus ion steps since they require backpropagation through the entire generative proce ss. To overcome these limitations we introduce two novel training methods for di screte diffusion samplers, one grounded in the policy gradient theorem and the o ther one leveraging Self-Normalized Neural Importance Sampling (SN-NIS). These m ethods yield memory-efficient training and achieve state-of-the-art results in u nsupervised combinatorial optimization.
Numerous scientific applications additionally require the ability of unbiased sa mpling. We introduce adaptations of SN-NIS and Neural Markov Chain Monte Carlo t hat enable for the first time the application of discrete diffusion models to th is problem. We validate our methods on Ising model benchmarks and find that they outperform popular autoregressive approaches. Our work opens new avenues for ap plying diffusion models to a wide range of scientific applications in discrete d omains that were hitherto restricted to exact likelihood models.
****

Shireen Kudukkil Manchingal,Muhammad Mubashar,Kaizheng Wang,Keivan Shariatmadar,

Fabio Cuzzolin
Random-Set Neural Networks
Machine learning is increasingly deployed in safety-critical domains where erron
eous predictions may lead to potentially catastrophic consequences, highlighting
 the need for learning systems to be aware of how confident they are in their ow
n predictions: in other words, 'to know when they do not know’. In this paper, w
e propose a novel Random-Set Neural Network (RS-NN) approach to classification w
hich predicts *belief functions* (rather than classical probability vectors) ove
r the class list using the mathematics of *random sets*, i.e., distributions ove
r the collection of *sets* of classes. RS-NN encodes the 'epistemic' uncertainty
 induced by training sets that are insufficiently representative or limited in s
ize via the size of the convex set of probability vectors associated with a pred
icted belief function. Our approach outperforms state-of-the-art Bayesian and En
semble methods in terms of accuracy, uncertainty estimation and out-of-distribut
ion (OoD) detection on multiple benchmarks (CIFAR-10 vs SVHN/Intel-Image, MNIST
vs FMNIST/KMNIST, ImageNet vs ImageNet-O). RS-NN also scales up effectively to l
arge-scale architectures (e.g. WideResNet-28-10, VGG16, Inception V3, EfficientN
etB2 and ViT-Base-16),
exhibits remarkable robustness to adversarial attacks and can provide statistica
l guarantees in a conformal learning setting.
****
Yuze Ge,Rujun Jiang
SOREL: A Stochastic Algorithm for Spectral Risks Minimization
The spectral risk has wide applications in machine learning, especially in real-
world decision-making, where people are concerned with more than just average mo
del performance. By assigning different weights to the losses of different sampl
e points, rather than the same weights as in the empirical risk, it allows the m
odel's performance to lie between the average performance and the worst-case per
formance. In this paper, we propose SOREL, the first stochastic gradient-based a
lgorithm with convergence guarantees for spectral risks minimization. Previous a
pproaches often rely on smoothing the spectral risk by adding a strongly concave
 function, thereby lacking convergence guarantees for the original spectral risk
. We theoretically prove that our algorithm achieves a near-optimal rate of $\w
idetilde{O}(1/\sqrt{\epsilon})$ to obtain an ϵ-optimal solution in term
s ϵ. Experiments on real datasets show that our algorithm outperforms e
xisting ones in most cases, both in terms of runtime and sample complexity.
****
Donghwan Rho,Taeseong Kim,Minje Park,Jung Woo Kim,Hyunsik Chae,Ernest K. Ryu,Jun
g Hee Cheon
Encryption-Friendly LLM Architecture
Large language models (LLMs) offer personalized responses based on user interact
ions, but this use case raises serious privacy concerns. Homomorphic encryption
(HE) is a cryptographic protocol supporting arithmetic computations in encrypted
 states and provides a potential solution for privacy-preserving machine learnin
g (PPML). However, the computational intensity of transformers poses challenges
for applying HE to LLMs. In this work, we propose a modified HE-friendly transfo
rmer architecture with an emphasis on inference following personalized (private)
 fine-tuning. Utilizing LoRA fine-tuning and Gaussian kernels, we achieve signif
icant computational speedups---6.94$\times$ for fine-tuning and 2.3$\times$ for
inference---while maintaining performance comparable to plaintext models. Our fi
ndings provide a viable proof of concept for offering privacy-preserving LLM ser
vices in areas where data protection is crucial. Our code is available on GitHub
.
****
Jeremy Andrew Irvin,Emily Ruoyu Liu,Joyce C. Chen,Ines Dormoy,Jinyoung Kim,Samar
 Khanna,Zhuo Zheng,Stefano Ermon
TEOChat: A Large Vision-Language Assistant for Temporal Earth Observation Data
Large vision and language assistants have enabled new capabilities for interpret
ing natural images. These approaches have recently been adapted to earth observa
tion data, but they are only able to handle single image inputs, limiting their

use for many real-world tasks. In this work, we develop a new vision and languag
e assistant called TEOChat that can engage in conversations about temporal seque
nces of earth observation data. To train TEOChat, we curate an instruction-follo
wing dataset composed of many single image and temporal tasks including building
 change and damage assessment, semantic change detection, and temporal scene cla
ssification. We show that TEOChat can perform a wide variety of spatial and temp
oral reasoning tasks, substantially outperforming previous vision and language a
ssistants, and even achieving comparable or better performance than several spec
ialist models trained to perform specific tasks. Furthermore, TEOChat achieves i
mpressive zero-shot performance on a change detection and change question answer
ing dataset, outperforms GPT-4o and Gemini 1.5 Pro on multiple temporal tasks, a
nd exhibits stronger single image capabilities than a comparable single image in
struction-following model on scene classification, visual question answering, an
d captioning. We publicly release our data, models, and code at https://github.c
om/ermongroup/TEOChat .
****
Di Wu,Hongwei Wang,Wenhao Yu,Yuwei Zhang,Kai-Wei Chang,Dong Yu
LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory
Recent large language model (LLM)-driven chat assistant systems have integrated
memory components to track user-assistant chat histories, enabling more accurate
 and personalized responses. However, their long-term memory capabilities in sus
tained interactions remain underexplored. We introduce LongMemEval, a comprehens
ive benchmark designed to evaluate five core long-term memory abilities of chat
assistants: information extraction, multi-session reasoning, temporal reasoning,
 knowledge updates, and abstention. With 500 meticulously curated questions embe
dded within freely scalable user-assistant chat histories, LongMemEval presents
a significant challenge to existing long-term memory systems, with commercial ch
at assistants and long-context LLMs showing a 30% accuracy drop on memorizing in
formation across sustained interactions. We then present a unified framework tha
t breaks down the long-term memory design into three stages: indexing, retrieval
, and reading. Built upon key experimental insights, we propose several memory d
esign optimizations including session decomposition for value granularity, fact-
augmented key expansion for indexing, and time-aware query expansion for refinin
g the search scope. Extensive experiments show that these optimizations greatly
improve both memory recall and downstream question answering on LongMemEval. Ove
rall, our study provides valuable resources and guidance for advancing the long-
term memory capabilities of LLM-based chat assistants, paving the way toward mor
e personalized and reliable conversational AI. Our benchmark and code are public
ly available at https://github.com/xiaowu0162/LongMemEval.
****
Guy Tevet,Sigal Raab,Setareh Cohan,Daniele Reda,Zhengyi Luo,Xue Bin Peng,Amit Ha
im Bermano,Michiel van de Panne
CLoSD: Closing the Loop between Simulation and Diffusion for multi-task characte
r control
Motion diffusion models and Reinforcement Learning (RL) based control for physic
s-based simulations have complementary strengths for human motion generation. Th
e former is capable of generating a wide variety of motions, adhering to intuiti
ve control such as text, while the latter offers physically plausible motion and
 direct interaction with the environment. In this work, we present a method that
 combines their respective strengths. CLoSD is a text-driven RL physics-based co
ntroller, guided by diffusion generation for various tasks. Our key insight is t
hat motion diffusion can serve as an on-the-fly universal planner for a robust R
L controller. To this end, CLoSD maintains a closed-loop interaction between two
 modules – a Diffusion Planner (DiP), and a tracking controller. DiP is a fast-r
esponding autoregressive diffusion model, controlled by textual prompts and targ
et locations, and the controller is a simple and robust motion imitator that con
tinuously receives motion plans from DiP and provides feedback from the environm
ent. CLoSD is capable of seamlessly performing a sequence of different tasks, in
cluding navigation to a goal location, striking an object with a hand or foot as
 specified in a text prompt, sitting down, and getting up.

****
Core Francisco Park,Andrew Lee,Ekdeep Singh Lubana,Yongyi Yang,Maya Okawa,Kento Nishi,Martin Wattenberg,Hidenori Tanaka
ICLR: In-Context Learning of Representations
Recent work demonstrates that structured patterns in pretraining data influence how representations of different concepts are organized in a large language model's (LLM) internals, with such representations then driving downstream abilities. Given the open-ended nature of LLMs, e.g., their ability to in-context learn novel tasks, we ask whether models can flexibly alter their semantically grounded organization of concepts. Specifically, if we provide in-context exemplars wherein a concept plays a different role than what the pretraining data suggests, can models infer these novel semantics and reorganize representations in accordance with them? To answer this question, we define a toy "graph tracing" task where in the nodes of the graph are referenced via concepts seen during training (e.g., apple, bird, etc.), and the connectivity of the graph is defined via some predefined structure (e.g., a square grid). Given exemplars that indicate traces of random walks on the graph, we analyze intermediate representations of the model and find that as the amount of context is scaled, there is a sudden re-organization of representations according to the graph's structure. Further, we find that when reference concepts have correlations in their semantics (e.g., Monday, Tuesday, etc.), the context-specified graph structure is still present in the representations, but is unable to dominate the pretrained structure. To explain these results, we analogize our task to energy minimization for a predefined graph topology, which shows getting non-trivial performance on the task requires for the model to infer a connected component. Overall, our findings indicate context-size may be an underappreciated scaling axis that can flexibly re-organize model representations, unlocking novel capabilities.
****
Yazheng Liu,Sihong Xie
Explanations of GNN on Evolving Graphs via Axiomatic Layer edges
Graphs are ubiquitous in social networks, chemical molecules, and financial data, where Graph Neural Networks (GNNs) achieve superior predictive accuracy. Graphs can be
evolving, while understanding how GNN predictions respond to the evolution provides significant insight and trust.
We explore the problem of explaining evolving GNN predictions due to continuously changing edge weights.
We introduce a layer edge-based explanation to balance
explanation fidelity and interpretability.
We propose a novel framework to address the challenges of axiomatic attribution and the entanglement of multiple computational graph paths due to continuous change of edge weights. We first design an axiomatic attribution of the evolution of the model prediction to message flows, then develop Shapley value to fairly map message flow contributions to layer edges.
We formulate a novel optimization problem to find the critical layer edges based on KL-divergence minimization. Extensive experiments on eight datasets for node classification, link prediction, and graph classification tasks with evolving graphs demonstrate the better fidelity and interpretability of the proposed method over the baseline methods. The code is available at https://github.com/yazhengliu/Axiomatic-Layer-Edges/tree/main.
****
Rui Li,Marcus Klasson,Arno Solin,Martin Trapp
Streamlining Prediction in Bayesian Deep Learning
The rising interest in Bayesian deep learning (BDL) has led to a plethora of methods for estimating the posterior distribution. However, efficient computation of inferences, such as predictions, has been largely overlooked with Monte Carlo integration remaining the standard. In this work we examine streamlining prediction in BDL through a single forward pass without sampling. For this, we use local linearisation of activation functions and local Gaussian approximations at linear layers. Thus allowing us to analytically compute an approximation of the pos

terior predictive distribution. We showcase our approach for both MLP and transf
ormers, such as ViT and GPT-2, and assess its performance on regression and clas
sification tasks.

Open-source library: https://github.com/AaltoML/SUQ.
****
Chengyu Du,Jinyi Han,Yizhou Ying,Aili Chen,Qianyu He,Haokun Zhao,Haoran Guo,Siru
i Xia,Jiaqing Liang,Zulong Chen,Liangyue Li,Yanghua Xiao
Think Thrice Before You Act: Progressive Thought Refinement in Large Language Mo
dels
Recent advancements in large language models (LLMs) have demonstrated that progr
essive refinement,
  rather than providing a single answer, results in more accurate and thoughtful
 outputs. However, existing methods often rely heavily on supervision signals to
 evaluate previous responses, making it difficult to effectively assess output q
uality in more open-ended scenarios. Additionally, these methods are typically d
esigned for specific tasks, which limits their generalization to new domains. To
 address these limitations, we propose Progressive Thought Refinement (PTR), a f
ramework that enables LLMs to progressively refine their responses. PTR operates
 in two phases: (1) Thought data construction stage: We propose a weak and stro
ng model collaborative selection strategy to build a high-quality progressive re
finement dataset to ensure logical consistency from thought to answers, and the
answers are gradually refined in each round. (2) Thought-Mask Fine-Tuning Phase:
 We design a training structure to mask the "thought" and adjust loss weights to
 encourage LLMs to refine prior thought, teaching them to implicitly understand
"how to improve" rather than "what is correct." Experimental results show that P
TR significantly enhances LLM performance across ten diverse tasks (avg. from 49
.6% to 53.5%) without task-specific fine-tuning. Notably, in more open-ended tas
ks, LLMs also demonstrate substantial improvements in the quality of responses b
eyond mere accuracy, suggesting that PTR truly teaches LLMs to self-improve over
 time. Our work is now open-source. https://github.com/cydu24/Progressive-Thoug
ht-Refinement
****
Yunfei Liu,Lei Zhu,Lijian Lin,Ye Zhu,Ailing Zhang,Yu Li
TEASER: Token Enhanced Spatial Modeling for Expressions Reconstruction
3D facial reconstruction from a single in-the-wild image is a crucial task in hu
man-centered computer vision tasks. While existing methods can recover accurate
facial shapes, there remains significant space for improvement in fine-grained e
xpression capture. Current approaches struggle with irregular mouth shapes, exa
ggerated expressions, and asymmetrical facial movements. We present TEASER (Toke
n EnhAnced Spatial modeling for Expressions Reconstruction), which addresses the
se challenges and enhances 3D facial geometry performance■■. TEASER tackles two
main limitations of existing methods: insufficient photometric loss for self-rec
onstruction and inaccurate localization of subtle expressions. We introduce a mu
lti-scale tokenizer to extract facial appearance information. Combined with a ne
ural renderer, these tokens provide precise geometric guidance for expression re
construction. Furthermore, TEASER incorporates a pose-dependent landmark loss to
 further improve geometric performance■. Our approach not only significantly enh
ances expression reconstruction quality but also offers interpretable tokens sui
table for various downstream applications, such as photorealistic facial video d
riving, expression transfer, and identity swapping. Quantitative and qualitative
 experimental results across multiple datasets demonstrate that TEASER achieves
state-of-the-art performance in precise expression reconstruction.
****
Ze Peng,Jian Zhang,Yisen Wang,Lei Qi,Yinghuan Shi,Yang Gao
Leveraging Flatness to Improve Information-Theoretic Generalization Bounds for S
GD
Information-theoretic (IT) generalization bounds have been used to study the gen
eralization of learning algorithms. These bounds are intrinsically data- and alg
orithm-dependent so that one can exploit the properties of data and algorithm to

derive tighter bounds. However, we observe that although the flatness bias is c
rucial for SGD’s generalization, these bounds fail to capture the improved gener
alization under better flatness and are also numerically loose. This is caused b
y the inadequate leverage of SGD's flatness bias in existing IT bounds. This pap
er derives a more flatness-leveraging IT bound for the flatness-favoring SGD. Th
e bound indicates the learned models generalize better if the large-variance dir
ections of the final weight covariance have small local curvatures in the loss l
andscape. Experiments on deep neural networks show our bound not only correctly
reflects the better generalization when flatness is improved, but is also numeri
cally much tighter. This is achieved by a flexible technique called "omniscient
trajectory". When applied to Gradient Descent’s minimax excess risk on convex-Li
pschitz-Bounded problems, it improves representative IT bounds’ $\Omega(1)$ rate
s to $O(1/\sqrt{n})$. It also implies a by-pass of memorization-generalization t
rade-offs. Codes are available at [https://github.com/peng-ze/omniscient-bounds]
(https://github.com/peng-ze/omniscient-bounds).
****

Juan Viguera Diez,Mathias Jacob Schreiner,Ola Engkvist,Simon Olsson
Boltzmann priors for Implicit Transfer Operators
Accurate prediction of thermodynamic properties is essential in drug discovery a
nd materials science. Molecular dynamics (MD) simulations provide a principled a
pproach to this task, yet they typically rely on prohibitively long sequential s
imulations. Implicit Transfer Operator (ITO) Learning offers a promising approac
h to address this limitation by enabling stable simulation with time steps order
s of magnitude larger than MD. However, to train ITOs, we need extensive, unbias
ed MD data, limiting the scope of this framework. Here, we introduce Boltzmann P
riors for ITO (BoPITO) to enhance ITO learning in two ways. First, BoPITO enable
s more efficient data generation, and second, it embeds inductive biases for lon
g-term dynamical behavior, simultaneously improving sample efficiency by one ord
er of magnitude and guaranteeing asymptotically unbiased equilibrium statistics.
Furthermore, we showcase the use of BoPITO in a new tunable sampling protocol i
nterpolating between ITOs trained on off-equilibrium simulations and an equilibr
ium model by incorporating unbiased correlation functions. Code is available at
https://github.com/olsson-group/bopito.
****

Baiting Luo,Ava Pettet,Aron Laszka,Abhishek Dubey,Ayan Mukhopadhyay
Scalable Decision-Making in Stochastic Environments through Learned Temporal Abs
traction
Sequential decision-making in high-dimensional continuous action spaces, particu
larly in stochastic environments, faces significant computational challenges. We
explore this challenge in the traditional offline RL setting, where an agent mu
st learn how to make decisions based on data collected through a stochastic beha
vior policy. We present \textit{Latent Macro Action Planner} (L-MAP), which addr
esses this challenge by learning a set of temporally extended macro-actions thro
ugh a state-conditional Vector Quantized Variational Autoencoder (VQ-VAE), effec
tively reducing action dimensionality. L-MAP employs a (separate) learned prior
model that acts as a latent transition model and allows efficient sampling of pl
ausible actions. During planning, our approach accounts for stochasticity in bot
h the environment and the behavior policy by using Monte Carlo tree search (MCTS
). In offline RL settings, including stochastic continuous control tasks, L-MAP
efficiently searches over discrete latent actions to yield high expected returns
.
Empirical results demonstrate that L-MAP maintains low decision latency despite
increased action dimensionality. Notably, across tasks ranging from continuous c
ontrol with inherently stochastic dynamics to high-dimensional robotic hand mani
pulation, L-MAP significantly outperforms existing model-based methods and perfo
rms on par with strong model-free actor-critic baselines, highlighting the effec
tiveness of the proposed approach in planning in complex and stochastic environm
ents with high-dimensional action spaces.
****

Hexu Zhao,Haoyang Weng,Daohan Lu,Ang Li,Jinyang Li,Aurojit Panda,Saining Xie

On Scaling Up 3D Gaussian Splatting Training
3D Gaussian Splatting (3DGS) is increasingly popular for 3D reconstruction due to its superior visual quality and rendering speed. However, 3DGS training currently occurs on a single GPU, limiting its ability to handle high-resolution and large-scale 3D reconstruction tasks due to memory constraints. We introduce Grendel, a distributed system designed to partition 3DGS parameters and parallelize computation across multiple GPUs. As each Gaussian affects a small, dynamic subset of rendered pixels, Grendel employs sparse all-to-all communication to transfer the necessary Gaussians to pixel partitions and performs dynamic load balancing. Unlike existing 3DGS systems that train using one camera view image at a time, Grendel supports batched training with multiple views. We explore various optimization hyperparameter scaling strategies and find that a simple sqrt(batch-size) scaling rule is highly effective. Evaluations using large-scale, high-resolution scenes show that Grendel enhances rendering quality by scaling up 3DGS parameters across multiple GPUs. On the 4K ``Rubble'' dataset, we achieve a test PSNR of 27.28 by distributing 40.4 million Gaussians across 16 GPU, compared to a PSNR of 26.28 using 11.2 million Gaussians on a single GPU. Grendel is an open-source project available at: https://github.com/nyu-systems/Grendel-GS
****
Lu Yi,Zhewei Wei
Scalable and Certifiable Graph Unlearning: Overcoming the Approximation Error Barrier
Graph unlearning has emerged as a pivotal research area for ensuring privacy protection, given the widespread adoption of Graph Neural Networks (GNNs) in applications involving sensitive user data. Among existing studies, certified graph unlearning is distinguished by providing robust privacy guarantees. However, current certified graph unlearning methods are impractical for large-scale graphs because they necessitate the costly re-computation of graph propagation for each unlearning request. Although numerous scalable techniques have been developed to accelerate graph propagation for GNNs, their integration into certified graph unlearning remains uncertain as these scalable approaches introduce approximation errors into node embeddings. In contrast, certified graph unlearning demands bounded model error on exact node embeddings to maintain its certified guarantee.

To address this challenge, we present ScaleGUN, the first approach to scale certified graph unlearning to billion-edge graphs. ScaleGUN integrates the approximate graph propagation technique into certified graph unlearning, offering certified guarantees for three unlearning scenarios: node feature, edge and node unlearning.
Extensive experiments on real-world datasets demonstrate the efficiency and unlearning efficacy of ScaleGUN. Remarkably, ScaleGUN accomplishes $(\epsilon,\delta)=(1,10^{-4})$ certified unlearning on the billion-edge graph ogbn-papers100M in 20 seconds for a 5,000 random edge removal request -- of which only 5 seconds are required for updating the node embeddings -- compared to 1.91 hours for retraining and 1.89 hours for re-propagation. Our code is available at https://github.com/luyi256/ScaleGUN.
****
Han Qi,Haocheng Yin,Heng Yang
Control-oriented Clustering of Visual Latent Representation
We initiate a study of the geometry of the visual representation space ---the information channel from the vision encoder to the action decoder--- in an image-based control pipeline learned from behavior cloning. Inspired by the phenomenon of *neural collapse* (NC) in image classification, we empirically demonstrate the prevalent emergence of a similar *law of clustering* in the visual representation space. Specifically,

- In discrete image-based control (e.g., Lunar Lander), the visual representations cluster according to the natural discrete action labels;

- In continuous image-based control (e.g., Planar Pushing and Block Stacking), t

he clustering emerges according to ``control-oriented'' classes that are based o
n (a) the relative pose between the object and the target in the input or (b) th
e relative pose of the object induced by expert actions in the output. Each of t
he classes corresponds to one relative pose orthant (REPO).

Beyond empirical observation, we show such a law of clustering can be leveraged
as an algorithmic tool to improve test-time performance when training a policy w
ith limited expert demonstrations. Particularly, we pretrain the vision encoder
using NC as a regularization to encourage control-oriented clustering of the vis
ual features. Surprisingly, such an NC-pretrained vision encoder, when finetuned
 end-to-end with the action decoder, boosts the test-time performance by 10% to
35%. Real-world vision-based planar pushing experiments confirmed the surprising
 advantage of control-oriented visual representation pretraining.
****
Stewart Slocum,Asher Parker-Sartori,Dylan Hadfield-Menell
Diverse Preference Learning for Capabilities and Alignment
As LLMs increasingly impact society, their ability to represent diverse perspect
ives is critical. However, recent studies reveal that alignment algorithms such
 as RLHF and DPO significantly reduce the diversity of LLM outputs. Not only do
aligned LLMs generate text with repetitive structure and word choice, they also
approach problems in more uniform ways, and their responses reflect a narrower r
ange of societal perspectives. We attribute this problem to the KL divergence re
gularizer employed in preference learning algorithms. This causes the model to o
verweight majority opinions and sacrifice diversity in exchange for optimal rewa
rd. To address this, we propose Soft Preference Learning, which decouples the en
tropy and cross-entropy terms in the KL penalty — allowing for fine-grained cont
rol over LLM generation diversity. From a capabilities perspective, LLMs trained
 using Soft Preference Learning attain higher accuracy on difficult repeated sam
pling tasks and produce outputs with greater semantic and lexical diversity. Fro
m an alignment perspective, they are capable of representing a wider range of so
cietal viewpoints and display improved logit calibration. Notably, Soft Preferen
ce Learning resembles, but is a Pareto improvement over, standard temperature sc
aling.
****
Zhenhan Fang,Aixin Tan,Jian Huang
CONTRA: Conformal Prediction Region via Normalizing Flow Transformation
Density estimation and reliable prediction regions for outputs are crucial in su
pervised and unsupervised learning. While conformal prediction effectively gener
ates coverage-guaranteed regions, it struggles with multi-dimensional outputs du
e to reliance on one-dimensional nonconformity scores. To address this, we intro
duce CONTRA: CONformal prediction region via normalizing flow TRAnsformation. CO
NTRA utilizes the latent spaces of normalizing flows to define nonconformity sco
res based on distances from the center. This allows for the mapping of high-dens
ity regions in latent space to sharp prediction regions in the output space, sur
passing traditional hyperrectangular or elliptical conformal regions. Further, f
or scenarios where other predictive models are favored over flow-based models, w
e extend CONTRA to enhance any such model with a reliable prediction region by t
raining a simple normalizing flow on the residuals. We demonstrate that both CON
TRA and its extension maintain guaranteed coverage probability and outperform ex
isting methods in generating accurate prediction regions across various datasets
. We conclude that CONTRA is an effective tool for (conditional) density estimat
ion, addressing the under-explored challenge of delivering multi-dimensional pre
diction regions.
****
Fanqi Lin,Yingdong Hu,Pingyue Sheng,Chuan Wen,Jiacheng You,Yang Gao
Data Scaling Laws in Imitation Learning for Robotic Manipulation
Data scaling has revolutionized fields like natural language processing and comp
uter vision, providing models with remarkable generalization capabilities. In th
is paper, we investigate whether similar data scaling laws exist in robotics, pa
rticularly in robotic manipulation, and whether appropriate data scaling can yie

ld single-task robot policies that can be deployed zero-shot for any object with in the same category in any environment. To this end, we conduct a comprehensive empirical study on data scaling in imitation learning. By collecting data acros s numerous environments and objects, we study how a policy’s generalization perf ormance changes with the number of training environments, objects, and demonstra tions. Throughout our research, we collect over 40,000 demonstrations and execut e more than 15,000 real-world robot rollouts under a rigorous evaluation protoco l. Our findings reveal several intriguing results: the generalization performanc e of the policy follows a roughly power-law relationship with the number of envi ronments and objects. The diversity of environments and objects is far more impo rtant than the absolute number of demonstrations; once the number of demonstrati ons per environment or object reaches a certain threshold, additional demonstrat ions have minimal effect. Based on these insights, we propose an efficient data collection strategy. With four data collectors working for one afternoon, we col lect sufficient data to enable the policies for two tasks to achieve approximate ly 90\% success rates in novel environments with unseen objects.
****

Ziyun Cui,Ziyang Zhang,Guangzhi Sun,Wen Wu,Chao Zhang
Bayesian WeakS-to-Strong from Text Classification to Generation
Advances in large language models raise the question of how alignment techniques will adapt as models become increasingly complex and humans will only be able t o supervise them weakly. Weak-to-Strong mimics such a scenario where weak model supervision attempts to harness the full capabilities of a much stronger model. This work extends Weak-to-Strong to WeakS-to-Strong by exploring an ensemble of weak models which simulate the variability in human opinions. Confidence scores are estimated using a Bayesian approach to guide the WeakS-to-Strong generalizat ion. Furthermore, we extend the application of WeakS-to-Strong from text classif ication tasks to text generation tasks where more advanced strategies are invest igated for supervision. Moreover, direct preference optimization is applied to a dvance the student model's preference learning, beyond the basic learning framew ork of teacher forcing. Results demonstrate the effectiveness of the proposed ap proach for the reliability of a strong student model, showing potential for supe ralignment.
****

Anastasiia Filippova,Angelos Katharopoulos,David Grangier,Ronan Collobert
No Need to Talk: Asynchronous Mixture of Language Models
We introduce SMALLTALK LM, an innovative method for training a mixture of langua ge models in an almost asynchronous manner. Each
model of the mixture specializes in distinct parts of the data distribution, wit hout the need of high-bandwidth communication between the nodes training each mo del. At inference, a lightweight router directs a given sequence to a single exp ert, according to a short prefix. This inference scheme naturally uses a fractio n of the parameters from the overall mixture model. Unlike prior works on asynch ronous LLM training, our routing method does not rely on full corpus clustering or access to metadata, making it more suitable for real-world applications. Our experiments on language modeling demonstrate that SMALLTALK LM achieves signifi cantly lower perplexity than dense model baselines for the same total training F LOPs and an almost identical inference cost. Finally, in our downstream evaluati ons we outperform the dense baseline on 75% of the tasks.
****

Konstantin Hemker,Nikola Simidjievski,Mateja Jamnik
Multimodal Lego: Model Merging and Fine-Tuning Across Topologies and Modalities in Biomedicine
Learning holistic computational representations in physical, chemical or biologi cal systems requires the ability to process information from different distribut ions and modalities within the same model. Thus, the demand for multimodal machi ne learning models has sharply risen for modalities that go beyond vision and la nguage, such as sequences, graphs, time series, or tabular data. While there are many available multimodal fusion and alignment approaches, most of them require end-to-end training, scale quadratically with the number of modalities, cannot

handle cases of high modality imbalance in the training set, or are highly topol
ogy-specific, making them too restrictive for many biomedical learning tasks. Th
is paper presents Multimodal Lego (MM-Lego), a general-purpose fusion framework
to turn any set of encoders into a competitive multimodal model with no or minim
al fine-tuning. We achieve this by introducing a wrapper for any unimodal encode
r that enforces shape consistency between modality representations. It harmonise
s these representations by learning features in the frequency domain to enable m
odel merging with little signal interference. We show that MM-Lego 1) can be use
d as a model merging method which achieves competitive performance with end-to-e
nd fusion models without any fine-tuning, 2) can operate on any unimodal encoder
, and 3) is a model fusion method that, with minimal fine-tuning, surpasses all
benchmarks in five out of seven datasets.
****
Domokos M. Kelen,Ádám Jung,Péter Kersch,Andras A Benczur
Distribution-Free Data Uncertainty for Neural Network Regression
Quantifying uncertainty is an essential part of predictive modeling, especially
in the context of high-stakes decision-making. While classification output inclu
des data uncertainty by design in the form of class probabilities, the regressio
n task generally aims only to predict the expected value of the target variable.
 Probabilistic extensions often assume parametric distributions around the expec
ted value, optimizing the likelihood over the resulting explicit densities. Howe
ver, using parametric distributions can limit practical applicability, making it
 difficult for models to capture skewed, multi-modal, or otherwise complex distr
ibutions. In this paper, we propose optimizing a novel nondeterministic neural n
etwork regression architecture for loss functions derived from a sample-based ap
proximation of the continuous ranked probability score (CRPS), enabling a truly
distribution-free approach by learning to sample from the target's aleatoric dis
tribution, rather than predicting explicit densities. Our approach allows the mo
del to learn well-calibrated, arbitrary uni- and multivariate output distributio
ns. We evaluate the method on a variety of synthetic and real-world tasks, inclu
ding uni- and multivariate problems, function inverse approximation, and standar
d regression uncertainty benchmarks. Finally, we make all experiment code public
ly available.
****
Yong-Hyun Park,Chieh-Hsin Lai,Satoshi Hayakawa,Yuhta Takida,Yuki Mitsufuji
Jump Your Steps: Optimizing Sampling Schedule of Discrete Diffusion Models
Diffusion models have seen notable success in continuous domains, leading to the
 development of discrete diffusion models (DDMs) for discrete variables. Despite
 recent advances, DDMs face the challenge of slow sampling speeds. While paralle
l sampling methods like τ-leaping accelerate this process, they introduce _
Compounding Decoding Error_ (CDE), where discrepancies arise between the true di
stribution and the approximation from parallel token generation, leading to degr
aded sample quality. In this work, we present _Jump Your Steps_ (JYS), a novel a
pproach that optimizes the allocation of discrete sampling timesteps by minimizi
ng CDE without extra computational cost. More precisely, we derive a practical u
pper bound on CDE and propose an efficient algorithm for searching for the optim
al sampling schedule. Extensive experiments across image, music, and text genera
tion show that JYS significantly improves sampling quality, establishing it as a
 versatile framework for enhancing DDM performance for fast sampling.
****
Arjun V Sudhakar,Hadi Nekoei,Mathieu Reymond,Miao Liu,Janarthanan Rajendran,Sara
th Chandar
A Generalist Hanabi Agent
Traditional multi-agent reinforcement learning (MARL) systems can develop cooper
ative strategies through repeated interactions. However, these systems are unabl
e to perform well on any other setting than the one they have been trained on, a
nd struggle to successfully cooperate with unfamiliar collaborators. This is par
ticularly visible in the Hanabi benchmark, a popular 2-to-5 player cooperative c
ard-game which requires complex reasoning and precise assistance to other agents
. Current MARL agents for Hanabi can only learn one specific game-setting (e.g.,

2-player games), and play with the same algorithmic agents. This is in stark co
ntrast to humans, who can quickly adjust their strategies to work with unfamilia
r partners or situations. In this paper, we introduce Recurrent Replay Relevance
Distributed DQN (R3D2), a generalist agent for Hanabi, designed to overcome the
se limitations. We reformulate the task using text, as language has been shown t
o improve transfer. We then propose a distributed MARL algorithm that copes with
the resulting dynamic observation- and action-space. In doing so, our agent is
the first that can play all game settings concurrently, and extend strategies le
arned from one setting to other ones. As a consequence, our agent also demonstra
tes the ability to collaborate with different algorithmic agents ---agents that
are themselves unable to do so.
****

Charles Jones,Fabio De Sousa Ribeiro,Mélanie Roschewitz,Daniel C. Castro,Ben Glo
cker

Rethinking Fair Representation Learning for Performance-Sensitive Tasks

We investigate the prominent class of fair representation learning methods for b
ias mitigation. Using causal reasoning to define and formalise different sources
of dataset bias, we reveal important implicit assumptions inherent to these met
hods. We prove fundamental limitations on fair representation learning when eval
uation data is drawn from the same distribution as training data and run experim
ents across a range of medical modalities to examine the performance of fair rep
resentation learning under distribution shifts. Our results explain apparent con
tradictions in the existing literature and reveal how rarely considered causal a
nd statistical aspects of the underlying data affect the validity of fair repres
entation learning. We raise doubts about current evaluation practices and the ap
plicability of fair representation learning methods in performance-sensitive set
tings. We argue that fine-grained analysis of dataset biases should play a key r
ole in the field moving forward.
****

Seonghwan Seo,Minsu Kim,Tony Shen,Martin Ester,Jinkyoo Park,Sungsoo Ahn,Woo Youn
Kim

Generative Flows on Synthetic Pathway for Drug Design

Generative models in drug discovery have recently gained attention as efficient
alternatives to brute-force virtual screening. However, most existing models do
not account for synthesizability, limiting their practical use in real-world sce
narios. In this paper, we propose RxnFlow, which sequentially assembles molecule
s using predefined molecular building blocks and chemical reaction templates to
constrain the synthetic chemical pathway. We then train on this sequential gener
ating process with the objective of generative flow networks (GFlowNets) to gene
rate both highly rewarded and diverse molecules. To mitigate the large action sp
ace of synthetic pathways in GFlowNets, we implement a novel action space subsam
pling method. This enables RxnFlow to learn generative flows over extensive acti
on spaces comprising combinations of 1.2 million building blocks and 71 reaction
templates without significant computational overhead. Additionally, RxnFlow can
employ modified or expanded action spaces for generation without retraining, al
lowing for the introduction of additional objectives or the incorporation of new
ly discovered building blocks. We experimentally demonstrate that RxnFlow outper
forms existing reaction-based and fragment-based models in pocket-specific optim
ization across various target pockets. Furthermore, RxnFlow achieves state-of-th
e-art performance on CrossDocked2020 for pocket-conditional generation, with an
average Vina score of −8.85 kcal/mol and 34.8% synthesizability. Code is availab
le at https://github.com/SeonghwanSeo/RxnFlow.
****

Lun Wang

Revisit Micro-batch Clipping: Adaptive Data Pruning via Gradient Manipulation

Micro-batch clipping, a gradient clipping method, has recently shown potential i
n enhancing auto-speech recognition (ASR) model performance. However, the underl
ying mechanism behind this improvement remains mysterious, particularly the obse
rvation that only certain micro-batch sizes are beneficial. In this paper, we ma
ke the first attempt to explain this phenomenon. Inspired by recent data pruning

research, we assume that specific training samples may impede model convergence during certain training phases. Under this assumption, the convergence analysis shows that micro-batch clipping can improve the convergence rate asymptotically at the cost of an additional constant bias that does not diminish with more training iterations. The bias is dependent on a few factors and can be minimized at specific micro-batch size, thereby elucidating the existence of the sweet-spot micro-batch size observed previously. We also verify the effectiveness of micro-batch clipping beyond speech models on vision and language models, and show promising performance gains in these domains. An exploration of potential limitations shows that micro-batch clipping is less effective when training data originates from multiple distinct domains.
****

Zhipei Xu,Xuanyu Zhang,Runyi Li,Zecheng Tang,Qing Huang,Jian Zhang
FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models
The rapid development of generative AI is a double-edged sword, which not only facilitates content creation but also makes image manipulation easier and more difficult to detect. Although current image forgery detection and localization (IFDL) methods are generally effective, they tend to face two challenges: \textbf{1)} black-box nature with unknown detection principle, \textbf{2)} limited generalization across diverse tampering methods (e.g., Photoshop, DeepFake, AIGC-Editing). To address these issues, we propose the explainable IFDL task and design FakeShield, a multi-modal framework capable of evaluating image authenticity, generating tampered region masks, and providing a judgment basis based on pixel-level and image-level tampering clues. Additionally, we leverage GPT-4o to enhance existing IFDL datasets, creating the Multi-Modal Tamper Description dataSet (MMTD-Set) for training FakeShield's tampering analysis capabilities. Meanwhile, we incorporate a Domain Tag-guided Explainable Forgery Detection Module (DTE-FDM) and a Multi-modal Forgery Localization Module (MFLM) to address various types of tamper detection interpretation and achieve forgery localization guided by detailed textual descriptions. Extensive experiments demonstrate that FakeShield effectively detects and localizes various tampering techniques, offering an explainable and superior solution compared to previous IFDL methods. The code is available at https://github.com/zhipeixu/FakeShield.
****

Alexander Nikulin,Ilya Zisman,Alexey Zemtsov,Vladislav Kurenkov
XLand-100B: A Large-Scale Multi-Task Dataset for In-Context Reinforcement Learning
Following the success of the in-context learning paradigm in large-scale language and computer vision models, the recently emerging field of in-context reinforcement learning is experiencing a rapid growth. However, its development has been held back by the lack of challenging benchmarks, as all the experiments have been carried out in simple environments and on small-scale datasets. We present **XLand-100B**, a large-scale dataset for in-context reinforcement learning based on the XLand-MiniGrid environment, as a first step to alleviate this problem. It contains complete learning histories for nearly $30,000$ different tasks, covering 100B transitions and 2.5B episodes. It took $50,000$ GPU hours to collect the dataset, which is beyond the reach of most academic labs. Along with the dataset, we provide the utilities to reproduce or expand it even further. We also benchmark common in-context RL baselines and show that they struggle to generalize to novel and diverse tasks. With this substantial effort, we aim to democratize research in the rapidly growing field of in-context reinforcement learning and provide a solid foundation for further scaling.
****

Jihan Yao,Wenxuan Ding,Shangbin Feng,Lucy Lu Wang,Yulia Tsvetkov
Varying Shades of Wrong: Aligning LLMs with Wrong Answers Only
In the absence of abundant reliable annotations for challenging tasks and contexts, how can we expand the frontier of LLM capabilities with potentially wrong answers? We focus on two research questions: (1) Can LLMs generate reliable preferences among wrong options? And if so, (2) Would alignment with such wrong-over-w

rong preferences be helpful? We employ methods based on self-consistency, token probabilities, and LLM-as-a-judge to elicit wrong-over-wrong preferences, and fine-tune language models with preference optimization approaches using these synthesized preferences. Extensive experiments with seven LLMs and eight datasets demonstrate that (1) LLMs do have preliminary capability in distinguishing various shades of wrong, achieving up to 20.9% higher performance than random guess; (2) Alignment with wrong-over-wrong preferences helps LLMs to produce less wrong and sometimes even outright correct answers, while improving overall model calibration. Code and data are publicly available at https://github.com/yaojh18/Varying-Shades-of-Wrong.
****
David Grangier,Simin Fan,Skyler Seto,Pierre Ablin
Task-Adaptive Pretrained Language Models via Clustered-Importance Sampling
Specialist language models (LMs) focus on a specific task or domain on which they often outperform generalist LMs of the same size. However, the specialist data needed to pretrain these models is only available in limited amount for most tasks. In this work, we build specialist models from large generalist training sets instead. We adjust the training distribution of the generalist data with guidance from the limited domain-specific data. We explore several approaches, with clustered importance sampling standing out. This method clusters the generalist dataset and samples from these clusters based on their frequencies in the smaller specialist dataset. It is scalable, suitable for pretraining and continued pretraining, it works well in multi-task settings. Our findings demonstrate improvements across different domains in terms of language modeling perplexity and accuracy on multiple-choice question tasks. We also present ablation studies that examine the impact of dataset sizes, clustering configurations, and model sizes.
****
Zhiyuan Liu,Yanchen Luo,Han Huang,Enzhi Zhang,Sihang Li,Junfeng Fang,Yaorui Shi,Xiang Wang,Kenji Kawaguchi,Tat-Seng Chua
NExT-Mol: 3D Diffusion Meets 1D Language Modeling for 3D Molecule Generation
3D molecule generation is crucial for drug discovery and material design. While prior efforts focus on 3D diffusion models for their benefits in modeling continuous 3D conformers, they overlook the advantages of 1D SELFIES-based Language Models (LMs), which can generate 100\% valid molecules and leverage the billion-scale 1D molecule datasets. To combine these advantages for 3D molecule generation, we propose a foundation model -- NExT-Mol: 3D Diffusion Meets 1D Language Modeling for 3D Molecule Generation. NExT-Mol uses an extensively pretrained molecule LM for 1D molecule generation, and subsequently predicts the generated molecule's 3D conformers with a 3D diffusion model. We enhance NExT-Mol's performance by scaling up the LM's model size, refining the diffusion neural architecture, and applying 1D to 3D transfer learning. Notably, our 1D molecule LM significantly outperforms baselines in distributional similarity while ensuring validity, and our 3D diffusion model achieves leading performances in conformer prediction. Given these improvements in 1D and 3D modeling, NExT-Mol achieves a 26\% relative improvement in 3D FCD for de novo 3D generation on GEOM-DRUGS, and a 13\% average relative gain for conditional 3D generation on QM9-2014. Our codes and pretrained checkpoints are available at https://github.com/acharkq/NExT-Mol.
****
Sagar Shrestha,Xiao Fu
Content-Style Learning from Unaligned Domains: Identifiability under Unknown Latent Dimensions
Understanding identifiability of latent content and style variables from unaligned multi-domain data is essential for tasks such as
domain translation and data generation. Existing works on content-style identification were often developed under somewhat stringent conditions, e.g., that all latent components are mutually independent and that the dimensions of the content and style variables are known. We introduce a new analytical framework via cross-domain *latent distribution matching* (LDM), which establishes content-style identifiability under substantially more relaxed conditions. Specifically, we show that restrictive assumptions such as component-wise independence of the laten

t variables can be removed. Most notably, we prove that prior knowledge of the c
ontent and style dimensions is not necessary for ensuring identifiability, if sp
arsity constraints are properly imposed onto the learned latent representations.
 Bypassing the knowledge of the exact latent dimension has been a longstanding a
spiration in unsupervised representation learning---our analysis is the first to
 underpin its theoretical and practical viability. On the implementation side, w
e recast the LDM formulation into a regularized multi-domain GAN loss with coupl
ed latent variables. We show that the reformulation is equivalent to LDM under m
ild conditions---yet requiring considerably less computational resource. Experim
ents corroborate with our theoretical claims.
****
Chen Chen,Daochang Liu,Mubarak Shah,Chang Xu
Exploring Local Memorization in Diffusion Models via Bright Ending Attention
Text-to-image diffusion models have achieved unprecedented proficiency in genera
ting realistic images. However, their inherent tendency to memorize and replicat
e training data during inference raises significant concerns, including potentia
l copyright infringement. In response, various methods have been proposed to eva
luate, detect, and mitigate memorization. Our analysis reveals that existing app
roaches significantly underperform in handling local memorization, where only sp
ecific image regions are memorized, compared to global memorization, where the e
ntire image is replicated. Also, they cannot locate the local memorization regio
ns, making it hard to investigate locally. To address these, we identify a novel
 "bright ending" (BE) anomaly in diffusion models prone to memorizing training i
mages. BE refers to a distinct cross-attention pattern observed in text-to-image
 diffusion models, where memorized image patches exhibit significantly greater a
ttention to the final text token during the last inference step than non-memoriz
ed patches. This pattern highlights regions where the generated image replicates
 training data and enables efficient localization of memorized regions. Equipped
 with this, we propose a simple yet effective method to integrate BE into existi
ng frameworks, significantly improving their performance by narrowing the perfor
mance gap caused by local memorization. Our results not only validate the succes
sful execution of the new localization task but also establish new state-of-the-
art performance across all existing tasks, underscoring the significance of the
BE phenomenon.
****
Qingjie Zhang,Han Qiu,Di Wang,Yiming Li,Tianwei Zhang,Wenyu Zhu,Haiqin Weng,Liu
Yan,Chao Zhang
A Benchmark for Semantic Sensitive Information in LLMs Outputs
Large language models (LLMs) can output sensitive information, which has emerged
 as a novel safety concern. Previous works focus on structured sensitive informa
tion (e.g. personal identifiable information).
However, we notice that sensitive information can also be at semantic level, i.e
. semantic sensitive information (SemSI).
Particularly, *simple natural questions* can let state-of-the-art (SOTA) LLMs ou
tput SemSI.
%which is hard to be detected compared with structured ones.
Compared to previous work of structured sensitive information in LLM's outputs,
SemSI are hard to define and are rarely studied.
Therefore, we propose a novel and large-scale investigation on the existence of
SemSI in SOTA LLMs induced by simple natural questions.
First, we construct a comprehensive and labeled dataset of semantic sensitive in
formation, SemSI-Set, by including three typical categories of SemSI.
Then, we propose a large-scale benchmark, SemSI-Bench, to systematically evaluat
e semantic sensitive information in 25 SOTA LLMs.
Our finding reveals that SemSI widely exists in SOTA LLMs' outputs by querying w
ith simple natural questions.
We open-source our project at https://semsi-project.github.io/.
****
Hulingxiao He,Geng Li,Zijun Geng,Jinglin Xu,Yuxin Peng
Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-mo

dal Large Language Models
Multi-modal large language models (MLLMs) have shown remarkable abilities in var ious visual understanding tasks. However, MLLMs still struggle with fine-grained visual recognition (FGVR), which aims to identify subordinate-level categories from images. This can negatively impact more advanced capabilities of MLLMs, suc h as object-centric visual question answering and reasoning. In our study, we re visit three quintessential capabilities of MLLMs for FGVR, including object info rmation extraction, category knowledge reserve, object-category alignment, and p osition of the root cause as a misalignment problem. To address this issue, we p resent Finedefics, an MLLM that enhances the model's FGVR capability by incorpor ating informative attribute descriptions of objects into the training phase. We employ contrastive learning on object-attribute pairs and attribute-category pai rs simultaneously and use examples from similar but incorrect categories as hard negatives, naturally bringing representations of visual objects and category na mes closer. Extensive evaluations across multiple popular FGVR datasets demonstr ate that Finedefics outperforms existing MLLMs of comparable parameter sizes, sh owcasing its remarkable efficacy. The code is available at [https://github.com/P KU-ICST-MIPL/Finedefics_ICLR2025](https://github.com/PKU-ICST-MIPL/Finedefics_IC LR2025).
****
Daeho Um,Jongin Lim,Sunoh Kim,Yuneil Yeo,Yoonho Jung
Spreading Out-of-Distribution Detection on Graphs
Node-level out-of-distribution (OOD) detection on graphs has received significan t attention from the machine learning community. However, previous approaches ar e evaluated using unrealistic benchmarks that consider only randomly selected OO D nodes, failing to reflect the interactions among nodes. In this paper, we intr oduce a new challenging task to model the interactions of OOD nodes in a graph, termed spreading OOD detection, where a newly emerged OOD node spreads its prope rty to neighboring nodes. We curate realistic benchmarks by employing the epidem ic spreading models that simulate the spreading of OOD nodes on the graph. We al so showcase a ``Spreading COVID-19" dataset to demonstrate the applicability of spreading OOD detection in real-world scenarios. Furthermore, to effectively det ect spreading OOD samples under the proposed benchmark setup, we present a new a pproach called energy distribution-based detector (EDBD), which includes a novel energy-aggregation scheme. EDBD is designed to mitigate undesired mixing of OOD scores between in-distribution (ID) and OOD nodes. Our extensive experimental r esults demonstrate the superiority of our approach over state-of-the-art methods in both spreading OOD detection and conventional node-level OOD detection tasks across seven benchmark datasets. The source code is available at https://github .com/daehoum1/edbd.
****
Tianqu Zhuang,Hongyao Yu,Yixiang Qiu,Hao Fang,Bin Chen,Shu-Tao Xia
Stealthy Shield Defense: A Conditional Mutual Information-Based Post-Processing against Black-Box Model Inversion Attacks
Model inversion attacks (MIAs) aim to reconstruct the private training data by a ccessing a public model, raising concerns about privacy leakage. Black-box MIAs, where attackers can only query the model and obtain outputs, are closer to real -world scenarios. The latest black-box attacks have outperformed the state-of-th e-art white-box attacks, and existing defenses cannot resist them effectively. T o fill this gap, we propose Stealthy Shield Defense (SSD), a post-processing alg orithm against black-box MIAs. Our idea is to modify the model's outputs to mini mize the conditional mutual information (CMI). We mathematically prove that CMI is a special case of information bottlenecks (IB), and thus inherits the advanta ges of IB---making predictions less dependent on inputs and more dependent on gr ound truths. This theoretically guarantees our effectiveness, both in resisting MIAs and preserving utility. For minimizing CMI, we formulate a convex optimizat ion problem and solve it via the water-filling method. Adaptive rate-distortion is introduced to constrain the modification to the outputs, and the water-fillin g is implemented on GPUs to address computation cost. Without the need to retrai n the model, our algorithm is plug-and-play and easy to deploy. Experimental res

ults indicate that SSD outperforms existing defenses, in terms of MIA resistance and model's utility, across various attack algorithms, training datasets, and model architectures. Our code is available at https://github.com/ZhuangQu/Stealthy-Shield-Defense.
****
Calvin Luo,Zilai Zeng,Yilun Du,Chen Sun
Solving New Tasks by Adapting Internet Video Knowledge
Video generative models demonstrate great promise in robotics by serving as visual planners or as policy supervisors. When pretrained on internet-scale data, such video models intimately understand alignment with natural language, and can thus facilitate generalization to novel downstream behavior through text-conditioning. However, they may not be sensitive to the specificities of the particular environment the agent inhabits. On the other hand, training video models on in-domain examples of robotic behavior naturally encodes environment-specific intricacies, but the scale of available demonstrations may not be sufficient to support generalization to unseen tasks via natural language specification. In this work, we investigate different adaptation techniques that integrate in-domain information with large-scale pretrained video models, and explore the extent to which they enable novel text-conditioned generalization for robotic tasks, while also considering their independent data and resource considerations. We successfully demonstrate across robotic environments that adapting powerful video models with small scales of example data can successfully facilitate generalization to novel behaviors. In particular, we present a novel adaptation strategy, termed *Inverse Probabilistic Adaptation*, that not only consistently achieves strong generalization performance across robotic tasks and settings, but also exhibits robustness to the quality of adaptation data, successfully solving novel tasks even when only suboptimal in-domain demonstrations are available.
****
Mingzhou Liu,Ching-Wen Lee,Xinwei Sun,Xueqing Yu,Yu QIAO,Yizhou Wang
Learning Causal Alignment for Reliable Disease Diagnosis
Aligning the decision-making process of machine learning algorithms with that of experienced radiologists is crucial for reliable diagnosis. While existing methods have attempted to align their prediction behaviors to those of radiologists reflected in the training data, this alignment is primarily associational rather than causal, resulting in pseudo-correlations that may not transfer well. In this paper, we propose a causality-based alignment framework towards aligning the model's decision process with that of experts. Specifically, we first employ counterfactual generation to identify the causal chain of model decisions. To align this causal chain with that of experts, we propose a causal alignment loss that enforces the model to focus on causal factors underlying each decision step in the whole causal chain. To optimize this loss that involves the counterfactual generator as an implicit function of the model's parameters, we employ the implicit function theorem equipped with the conjugate gradient method for efficient estimation. We demonstrate the effectiveness of our method on two medical diagnosis applications, showcasing faithful alignment to radiologists.
****
Mingyang Chen,sunhaoze,Tianpeng Li,Fan Yang,Hao Liang,KeerLu,Bin CUI,Wentao Zhang,Zenan Zhou,weipeng chen
Facilitating Multi-turn Function Calling for LLMs via Compositional Instruction Tuning
Large Language Models (LLMs) have exhibited significant potential in performing diverse tasks, including the ability to call functions or use external tools to enhance their performance. While current research on function calling by LLMs primarily focuses on single-turn interactions, this paper addresses the overlooked necessity for LLMs to engage in multi-turn function calling—critical for handling compositional, real-world queries that require planning with functions but not only use functions. To facilitate this, we introduce an approach, BUTTON, which generates synthetic compositional instruction tuning data via bottom-up instruction construction and top-down trajectory generation. In the bottom-up phase, we generate simple atomic tasks based on real-world scenarios and build compositi

onal tasks using heuristic strategies based on atomic tasks. Corresponding funct
ion definitions are then synthesized for these compositional tasks. The top-down
 phase features a multi-agent environment where interactions among simulated hum
ans, assistants, and tools are utilized to gather multi-turn function calling tr
ajectories. This approach ensures task compositionality and allows for effective
 function and trajectory generation by examining atomic tasks within composition
al tasks. We produce a dataset BUTTONInstruct comprising 8k data points and demo
nstrate its effectiveness through extensive experiments across various LLMs.
****
Minghao Xu,Yunteng Geng,Yihang Zhang,Ling Yang,Jian Tang,Wentao Zhang
GlycanML: A Multi-Task and Multi-Structure Benchmark for Glycan Machine Learning
Glycans are basic biomolecules and perform essential functions within living org
anisms. The rapid increase of functional glycan data provides a good opportunity
 for machine learning solutions to glycan understanding. However, there still la
cks a standard machine learning benchmark for glycan property and function predi
ction. In this work, we fill this blank by building a comprehensive benchmark fo
r Glycan Machine Learning (GlycanML). The GlycanML benchmark consists of diverse
 types of tasks including glycan taxonomy prediction, glycan immunogenicity pred
iction, glycosylation type prediction, and protein-glycan interaction prediction
. Glycans can be represented by both sequences and graphs in GlycanML, which ena
bles us to extensively evaluate sequence-based models and graph neural networks
(GNNs) on benchmark tasks. Furthermore, by concurrently performing eight glycan
taxonomy prediction tasks, we introduce the GlycanML-MTL testbed for multi-task
learning (MTL) algorithms. Also, we evaluate how taxonomy prediction can boost o
ther three function prediction tasks by MTL. Experimental results show the super
iority of modeling glycans with multi-relational GNNs, and suitable MTL methods
can further boost model performance. We provide all datasets and source codes at
 https://github.com/GlycanML/GlycanML and maintain a leaderboard at https://Glyc
anML.github.io/project
****
Logan Cross,Violet Xiang,Agam Bhatia,Daniel LK Yamins,Nick Haber
Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large
Language Models
Multi-agent reinforcement learning (MARL) methods struggle with the non-stationa
rity of multi-agent systems and fail to adaptively learn online when tested with
 novel agents. Here, we leverage large language models (LLMs) to create an auton
omous agent that can handle these challenges. Our agent, Hypothetical Minds, con
sists of a cognitively-inspired architecture, featuring modular components for p
erception, memory, and hierarchical planning over two levels of abstraction. We
introduce the Theory of Mind module that scaffolds the high-level planning proce
ss by generating hypotheses about other agents' strategies in natural language.
It then evaluates and iteratively refines these hypotheses by reinforcing hypoth
eses that make correct predictions about the other agents' behavior. Hypothetica
l Minds significantly improves performance over previous LLM-agent and RL baseli
nes on a range of competitive, mixed motive, and collaborative domains in the Me
lting Pot benchmark, including both dyadic and population-based environments. Ad
ditionally, comparisons against LLM-agent baselines and ablations reveal the imp
ortance of hypothesis evaluation and refinement for succeeding on complex scenar
ios.
****
Martin Klissarov,Mikael Henaff,Roberta Raileanu,Shagun Sodhani,Pascal Vincent,Am
y Zhang,Pierre-Luc Bacon,Doina Precup,Marlos C. Machado,Pierluca D'Oro
MaestroMotif: Skill Design from Artificial Intelligence Feedback
Describing skills in natural language has the potential to provide an accessible
 way to inject human knowledge about decision-making into an AI system. We prese
nt MaestroMotif, a method for AI-assisted skill design, which yields high-perfor
ming and adaptable agents. MaestroMotif leverages the capabilities of Large Lang
uage Models (LLMs) to effectively create and reuse skills. It first uses an LLM'
s feedback to automatically design rewards corresponding to each skill, starting
 from their natural language description. Then, it employs an LLM's code generat

ion abilities, together with reinforcement learning, for training the skills and combining them to implement complex behaviors specified in language. We evaluate MaestroMotif using a suite of complex tasks in the NetHack Learning Environment (NLE), demonstrating that it surpasses existing approaches in both performance and usability.
****
Qi Fan,Xin Tao,Lei Ke,Mingqiao Ye,Di ZHANG,Pengfei Wan,Yu-Wing Tai,Chi-Keung Tang
Stable Segment Anything Model
The Segment Anything Model (SAM) achieves remarkable promptable segmentation given high-quality prompts which, however, often require good skills to specify. To make SAM robust to casual prompts, this paper presents the first comprehensive analysis on SAM’s segmentation stability across a diverse spectrum of prompt qualities, notably imprecise bounding boxes and insufficient points. Our key finding reveals that given such low-quality prompts, SAM’s mask decoder tends to activate image features that are biased towards the background or confined to specific object parts. To mitigate this issue, our key idea consists of calibrating solely SAM’s mask attention by adjusting the sampling locations and amplitudes of image features, while the original SAM model architecture and weights remain unchanged. Consequently, our deformable sampling plugin (DSP) enables SAM to adaptively shift attention to the prompted target regions in a data-driven manner. During inference, dynamic routing plugin (DRP) is proposed that toggles SAM between the deformable and regular grid sampling modes, conditioned on the input prompt quality. Thus, our solution, termed Stable-SAM, offers several advantages: 1) improved SAM’s segmentation stability across a wide range of prompt qualities, while 2) retaining SAM’s powerful promptable segmentation efficiency and generality, with 3) minimal learnable parameters (0.08 M) and fast adaptation. Extensive experiments validate the effectiveness and advantages of our approach, underscoring Stable-SAM as a more robust solution for segmenting anything. Codes are at https://github.com/fanq15/Stable-SAM.
****
Najwa Laabid,Severi Rissanen,Markus Heinonen,Arno Solin,Vikas Garg
Equivariant Denoisers Cannot Copy Graphs: Align Your Graph Diffusion Models
Graph diffusion models, dominant in graph generative modeling, remain underexplored for graph-to-graph translation tasks like chemical reaction prediction. We demonstrate that standard permutation equivariant denoisers face fundamental limitations in these tasks due to their inability to break symmetries in noisy inputs. To address this, we propose \emph{aligning} input and target graphs to break input symmetries while preserving permutation equivariance in non-matching graph portions. Using retrosynthesis (i.e., the task of predicting precursors for synthesis of a given target molecule) as our application domain, we show how alignment dramatically improves discrete diffusion model performance from 5\% to a SOTA-matching 54.7\% top-1 accuracy. Code is available at https://github.com/Aalto-QuML/DiffAlign.
****
Zhe Li,Bicheng Ying,Zidong Liu,Chaosheng Dong,Haibo Yang
Achieving Dimension-Free Communication in Federated Learning via Zeroth-Order Optimization
Federated Learning (FL) offers a promising framework for collaborative and privacy-preserving machine learning across distributed data sources.
However, the substantial communication costs associated with FL significantly challenge its efficiency.
Specifically, in each communication round, the communication costs scale linearly with the model's dimension, which presents a formidable obstacle, especially in large model scenarios.
Despite various communication-efficient strategies, the intrinsic dimension-dependent communication cost remains a major bottleneck for current FL implementations.
This paper proposes a novel dimension-free communication algorithm - DeComFL, which leverages the zeroth-order optimization techniques and reduces the communica

tion cost from $\mathcal{O}(d)$ to $\mathcal{O}(1)$ by transmitting only a const
ant number of scalar values between clients and the server in each round, regard
less of the dimension d of the model parameters.
Theoretically, in non-convex functions, we prove that our algorithm achieves sta
te-of-the-art rates, which show a linear speedup of the number of clients and lo
cal steps under standard assumptions. With additional low effective rank assumpt
ion, we can further show that the convergence rate is independent of the model d
imension d as well.
Empirical evaluations, encompassing both classic deep learning training and larg
e language model fine-tuning, demonstrate significant reductions in communicatio
n overhead.
Notably, DeComFL achieves this by transmitting only around 1MB of data in total
between the server and a client to fine-tune a model with billions of parameters
.
The code is available at https://github.com/ZidongLiu/DeComFL.
****
Artyom Gadetsky,Andrei Atanov,Yulun Jiang,Zhitong Gao,Ghazal Hosseini Mighan,Ami
r Zamir,Maria Brbic
Large (Vision) Language Models are Unsupervised In-Context Learners
Recent advances in large language and vision-language models have enabled zero-s
hot inference, allowing models to solve new tasks without task-specific training
. Various adaptation techniques such as prompt engineering, In-Context Learning
(ICL), and supervised fine-tuning can further enhance the model’s performance on
 a downstream task, but they require substantial manual effort to construct effe
ctive prompts or labeled examples. In this work, we introduce a joint inference
framework for fully unsupervised adaptation, eliminating the need for manual pro
mpt engineering and labeled examples. Unlike zero-shot inference, which makes in
dependent predictions, the joint inference makes predictions simultaneously for
all inputs in a given task. Since direct joint inference involves computationall
y expensive optimization, we develop efficient approximation techniques, leading
 to two unsupervised adaptation methods: unsupervised fine-tuning and unsupervis
ed ICL. We demonstrate the effectiveness of our methods across diverse tasks and
 models, including language-only Llama-3.1 on natural language processing tasks,
 reasoning-oriented Qwen2.5-Math on grade school math problems, vision-language
OpenFlamingo on vision tasks, and the API-only access GPT-4o model on massive mu
lti-discipline tasks. Our experiments demonstrate substantial improvements over
the standard zero-shot approach, including 39% absolute improvement on the chall
enging GSM8K math reasoning dataset. Remarkably, despite being fully unsupervise
d, our framework often performs on par with supervised approaches that rely on g
round truth labels.
****
Manuel Faysse,Hugues Sibille,Tony Wu,Bilel Omrani,Gautier Viaud,CELINE HUDELOT,P
ierre Colombo
ColPali: Efficient Document Retrieval with Vision Language Models
Documents are visually rich structures that convey information through text, but
 also figures, page layouts, tables, or even fonts. Since modern retrieval syste
ms mainly rely on the textual information they extract from document pages to in
dex documents -often through lengthy and brittle processes-, they struggle to ex
ploit key visual cues efficiently. This limits their capabilities in many practi
cal document retrieval applications such as Retrieval Augmented Generation (RAG)
.
To benchmark current systems on visually rich document retrieval, we introduce t
he Visual Document Retrieval Benchmark $\textit{ViDoRe}$, composed of various pa
ge-level retrieval tasks spanning multiple domains, languages, and practical set
tings.
The inherent complexity and performance shortcomings of modern systems motivate
a new concept; doing document retrieval by directly embedding the images of the
document pages. We release $\textit{ColPali}$, a Vision Language Model trained t
o produce high-quality multi-vector embeddings from images of document pages. Co
mbined with a late interaction matching mechanism, $\textit{ColPali}$ largely ou

tperforms modern document retrieval pipelines while being drastically simpler, f
aster and end-to-end trainable.
We release models, data, code and benchmarks under open licenses at https://hf.c
o/vidore.
****
Dohyeong Kim,Mineui Hong,Jeongho Park,Songhwai Oh
Conflict-Averse Gradient Aggregation for Constrained Multi-Objective Reinforceme
nt Learning
In real-world applications, a reinforcement learning (RL) agent should consider
multiple objectives and adhere to safety guidelines.
To address these considerations, we propose a constrained multi-objective RL alg
orithm named constrained multi-objective gradient aggregator (CoMOGA).
In the field of multi-objective optimization, managing conflicts between the gra
dients of the multiple objectives is crucial to prevent policies from converging
 to local optima.
It is also essential to efficiently handle safety constraints for stable trainin
g and constraint satisfaction.
We address these challenges straightforwardly by treating the maximization of mu
ltiple objectives as a constrained optimization problem (COP), where the constra
ints are defined to improve the original objectives.
Existing safety constraints are then integrated into the COP, and the policy is
updated by solving the COP, which ensures the avoidance of gradient conflicts.
Despite its simplicity, CoMOGA guarantees convergence to global optima in a tabu
lar setting.
Through various experiments, we have confirmed that preventing gradient conflict
s is critical, and the proposed method achieves constraint satisfaction across a
ll tasks.
****
Zechun Liu,Changsheng Zhao,Igor Fedorov,Bilge Soran,Dhruv Choudhary,Raghuraman K
rishnamoorthi,Vikas Chandra,Yuandong Tian,Tijmen Blankevoort
SpinQuant: LLM Quantization with Learned Rotations
Post-training quantization (PTQ) techniques applied to weights, activations, and
 the KV cache greatly reduce memory usage, latency, and power consumption of Lar
ge Language Models (LLMs), but may lead to large quantization errors when outlie
rs are present. Rotating activation or weight matrices helps remove outliers and
 benefits quantization. In this work, we identify a collection of applicable rot
ation parameterizations that lead to identical outputs in full-precision Transfo
rmer architectures while enhancing quantization accuracy. In addition, we find t
hat some random rotations lead to much better quantization than others, with an
up to 13 points difference in downstream zero-shot reasoning performance. As a r
esult, we propose SpinQuant, a novel approach that incorporates learned rotation
 matrices for optimal quantized network accuracy. With 4-bit quantization of wei
ght, activation, and KV-cache, SpinQuant narrows the accuracy gap on zero-shot r
easoning tasks with full precision to merely 2.9 points on the LLaMA-2 7B model,
 surpassing LLM-QAT by 19.1 points and SmoothQuant by 25.0 points. Furthermore,
SpinQuant also outperforms concurrent work QuaRot, which applies random rotation
s to remove outliers. In particular, for LLaMA-3 8B models that are hard to quan
tize, SpinQuant reduces the gap to full precision by up to 45.1% relative to Qua
Rot. Code is available at https://github.com/facebookresearch/SpinQuant.
****
Zhengqing Wu,Berfin Simsek,François Gaston Ged
Loss Landscape of Shallow ReLU-like Neural Networks: Stationary Points, Saddle E
scape, and Network Embedding
In this paper, we study the loss landscape of one-hidden-layer neural networks w
ith ReLU-like activation functions trained with the empirical squared loss using
 gradient descent (GD). We identify the stationary points of such networks, whic
h significantly slow down loss decrease during training. To capture such points
while accounting for the non-differentiability of the loss, the stationary point
s that we study are directional stationary points, rather than other notions lik
e Clarke stationary points. We show that, if a stationary point does not contain

"escape neurons", which are defined with first-order conditions, it must be a l
ocal minimum. Moreover, for the scalar-output case, the presence of an escape ne
uron guarantees that the stationary point is not a local minimum. Our results re
fine the description of the *saddle-to-saddle* training process starting from in
finitesimally small (vanishing) initialization for shallow ReLU-like networks: B
y precluding the saddle escape types that previous works did not rule out, we ad
vance one step closer to a complete picture of the entire dynamics. Moreover, we
 are also able to fully discuss how network embedding, which is to instantiate a
 narrower network with a wider network, reshapes the stationary points.
****

Yan Scholten,Stephan Günnemann
Provably Reliable Conformal Prediction Sets in the Presence of Data Poisoning
Conformal prediction provides model-agnostic and distribution-free uncertainty q
uantification through prediction sets that are guaranteed to include the ground
truth with any user-specified probability. Yet, conformal prediction is not reli
able under poisoning attacks where adversaries manipulate both training and cali
bration data, which can significantly alter prediction sets in practice. As a so
lution, we propose reliable prediction sets (RPS): the first efficient method fo
r constructing conformal prediction sets with provable reliability guarantees un
der poisoning. To ensure reliability under training poisoning, we introduce smoo
thed score functions that reliably aggregate predictions of classifiers trained
on distinct partitions of the training data. To ensure reliability under calibra
tion poisoning, we construct multiple prediction sets, each calibrated on distin
ct subsets of the calibration data. We then aggregate them into a majority predi
ction set, which includes a class only if it appears in a majority of the indivi
dual sets. Both proposed aggregations mitigate the influence of datapoints in th
e training and calibration data on the final prediction set. We experimentally v
alidate our approach on image classification tasks, achieving strong reliability
 while maintaining utility and preserving coverage on clean data. Overall, our a
pproach represents an important step towards more trustworthy uncertainty quanti
fication in the presence of data poisoning.
****

Christophe Roux,Max Zimmer,Sebastian Pokutta
On the Byzantine-Resilience of Distillation-Based Federated Learning
Federated Learning (FL) algorithms using Knowledge Distillation (KD) have receiv
ed increasing attention due to their favorable properties with respect to privac
y, non-i.i.d. data and communication cost. These methods depart from transmittin
g model parameters and instead communicate information about a learning task by
sharing predictions on a public dataset. In this work, we study the performance
of such approaches in the byzantine setting, where a subset of the clients act i
n an adversarial manner aiming to disrupt the learning process. We show that KD-
based FL algorithms are remarkably resilient and analyze how byzantine clients c
an influence the learning process. Based on these insights, we introduce two new
 byzantine attacks and demonstrate their ability to break existing byzantine-res
ilient methods. Additionally, we propose a novel defence method which enhances t
he byzantine resilience of KD-based FL algorithms. Finally, we provide a general
 framework to obfuscate attacks, making them significantly harder to detect, the
reby improving their effectiveness.
****

Silong Yong,Venkata Nagarjun Pudureddiyur Manivannan,Bernhard Kerbl,Zifu Wan,Sim
on Stepputtis,Katia P. Sycara,Yaqi Xie
OMG: Opacity Matters in Material Modeling with Gaussian Splatting
Decomposing geometry, materials and lighting from a set of images, namely invers
e rendering, has been a long-standing problem in computer vision and graphics. R
ecent advances in neural rendering enable photo-realistic and plausible inverse
rendering results. The emergence of 3D Gaussian Splatting has boosted it to the
next level by showing real-time rendering potentials. An intuitive finding is t
hat the models used for inverse rendering do not take into account the dependenc
y of opacity w.r.t. material properties, namely cross section, as suggested by o
ptics. Therefore, we develop a novel approach that adds this dependency to the m

odeling itself. Inspired by radiative transfer, we augment the opacity term by i
ntroducing a neural network that takes as input material properties to provide m
odeling of cross section and a physically correct activation function. The gradi
ents for material properties are therefore not only from color but also from opa
city, facilitating a constraint for their optimization. Therefore, the proposed
method incorporates more accurate physical properties compared to previous works
. We implement our method into 3 different baselines that use Gaussian Splatting
 for inverse rendering and achieve significant improvements universally in terms
 of novel view synthesis and material modeling.
****

Yulun Wu,Louis McConnell,Claudia Iriondo
Counterfactual Generative Modeling with Variational Causal Inference
Estimating an individual's potential outcomes under counterfactual treatments is
 a challenging task for traditional causal inference and supervised learning app
roaches when the outcome is high-dimensional (e.g. gene expressions, facial imag
es) and covariates are relatively limited. In this case, to predict one's outcom
es under counterfactual treatments, it is crucial to leverage individual informa
tion contained in the observed outcome in addition to the covariates. Prior work
s using variational inference in counterfactual generative modeling have been fo
cusing on neural adaptations and model variants within the conditional variation
al autoencoder formulation, which we argue is fundamentally ill-suited to the no
tion of counterfactual in causal inference. In this work, we present a novel var
iational Bayesian causal inference framework and its theoretical backings to pro
perly handle counterfactual generative modeling tasks, through which we are able
 to conduct counterfactual supervision end-to-end during training without any co
unterfactual samples, and encourage disentangled exogenous noise abduction that
aids the correct identification of causal effect in counterfactual generations.
In experiments, we demonstrate the advantage of our framework compared to state-
of-the-art models in counterfactual generative modeling on multiple benchmarks.
****

Somnath Basu Roy Chowdhury,Krzysztof Marcin Choromanski,Arijit Sehanobish,Kumar
Avinava Dubey,Snigdha Chaturvedi
Towards Scalable Exact Machine Unlearning Using Parameter-Efficient Fine-Tuning
Machine unlearning is the process of efficiently removing the influence of a tra
ining data instance from a trained machine learning model without retraining it
from scratch. A popular subclass of unlearning approaches is exact machine unlea
rning, which focuses on techniques that explicitly guarantee the removal of the
influence of a data instance from a model. Exact unlearning approaches use a ma
chine learning model in which individual components are trained on disjoint subs
ets of the data. During deletion, exact unlearning approaches only retrain the a
ffected components rather than the entire model. While existing approaches reduc
e retraining costs, it can still be expensive for an organization to retrain a m
odel component as it requires halting a system in production, which leads to ser
vice failure and adversely impacts customers. To address these challenges, we i
ntroduce an exact unlearning framework -- Sequence-aware Sharded Sliced Training
  (S3T), which is designed to enhance the deletion capabilities of an exact unle
arning system while minimizing the impact on model's performance. At the core of
 S3T, we utilize a lightweight parameter-efficient fine-tuning approach that ena
bles parameter isolation by sequentially training layers with disjoint data slic
es. This enables efficient unlearning by simply deactivating the layers affected
 by data deletion. Furthermore, to reduce the retraining cost and improve model
performance, we train the model on multiple data sequences, which allows S3T to
handle an increased number of deletion requests. Both theoretically and empirica
lly, we demonstrate that S3T attains superior deletion capabilities and enhanced
 performance compared to baselines across a wide range of settings.
****

Orion Weller,Benjamin Van Durme,Dawn Lawrie,Ashwin Paranjape,Yuhao Zhang,Jack He
ssel
Promptriever: Instruction-Trained Retrievers Can Be Prompted Like Language Model
s

Instruction-tuned language models (LM) are able to respond to imperative command s, providing a more natural user interface compared to their base counterparts. In this work, we present Promptriever, the first retrieval model able to be prom pted like an LM. To train Promptriever, we curate and release a new instance-lev el instruction training set from MS MARCO, spanning nearly 500k instances. Promp triever not only achieves strong performance on standard retrieval tasks, but al so follows instructions. We observe: (1) large gains (reaching SoTA) on followin g detailed relevance instructions (+14.3 p-MRR / +3.1 nDCG on FollowIR), (2) sig nificantly increased robustness to lexical choices/phrasing in the query+instruc tion (+12.9 Robustness@10 on InstructIR), and (3) the ability to perform hyper-p arameter search via prompting to reliably improve retrieval performance (+1.4 av erage increase on BEIR). Promptriever demonstrates that retrieval models can be controlled with prompts on a per-query basis, setting the stage for future work aligning LM prompting techniques with information retrieval.
****

Esben Kran,Hieu Minh Nguyen,Akash Kundu,Sami Jawhar,Jinsuk Park,Mateusz Maria Ju rewicz

DarkBench: Benchmarking Dark Patterns in Large Language Models

We introduce DarkBench, a comprehensive benchmark for detecting dark design patt erns—manipulative techniques that influence user behavior—in interactions with l arge language models (LLMs). Our benchmark comprises 660 prompts across six cate gories: brand bias, user retention, sycophancy, anthropomorphism, harmful genera tion, and sneaking. We evaluate models from five leading companies (OpenAI, Anth ropic, Meta, Mistral, Google) and find that some LLMs are explicitly designed to favor their developers' products and exhibit untruthful communication, among ot her manipulative behaviors. Companies developing LLMs should recognize and mitig ate the impact of dark design patterns to promote more ethical Al.
****

Tony Alex,Sara Atito,Armin Mustafa,Muhammad Awais,Philip J B Jackson

SSLAM: Enhancing Self-Supervised Models with Audio Mixtures for Polyphonic Sound scapes

Self-supervised pre-trained audio networks have seen widespread adoption in real -world systems, particularly in multi-modal large language models. These network s are often employed in a frozen state, under the assumption that the self-super vised pre-training has sufficiently equipped them to handle real-world audio. Ho wever, a critical question remains: how well do these models actually perform in real-world conditions, where audio is typically polyphonic and complex, involvi ng multiple overlapping sound sources? Current audio self-supervised learning (S SL) methods are often benchmarked on datasets predominantly featuring monophonic audio, such as environmental sounds, and speech. As a result, the ability of SS L models to generalize to polyphonic audio, a common characteristic in natural s cenarios, remains underexplored. This limitation raises concerns about the pract ical robustness of SSL models in more realistic audio settings. To address this gap, we introduce Self-Supervised Learning from Audio Mixtures (SSLAM), a novel direction in audio SSL research, designed to improve the model’s ability to lea rn from polyphonic data while maintaining strong performance on monophonic data. We thoroughly evaluate SSLAM on standard audio SSL benchmark datasets which are predominantly monophonic and conduct a comprehensive comparative analysis again st state-of-the-art (SOTA) methods using a range of high-quality, publicly avail able polyphonic datasets. SSLAM not only improves model performance on polyphoni c audio, but also maintains or exceeds performance on standard audio SSL benchma rks. Notably, it achieves up to a 3.9% improvement on the AudioSet-2M(AS-2M), re aching a mean average precision (mAP) of 50.2. For polyphonic datasets, SSLAM se ts new SOTA in both linear evaluation and fine-tuning regimes with performance i mprovements of up to 9.1%(mAP). These results demonstrate SSLAM's effectiveness in both polyphonic and monophonic soundscapes, significantly enhancing the perfo rmance of audio SSL models. Code and pre-trained models are available at https:/ /github.com/ta012/SSLAM.
****

Zhen Zhang,Xin Liu,Shaoli Wang,Jiaye Teng

Minimax Optimal Two-Stage Algorithm For Moment Estimation Under Covariate Shift
Covariate shift occurs when the distribution of input features differs between t
he training and testing phases. In covariate shift, estimating an unknown funct
ion's moment is a classical problem that remains under-explored, despite its com
mon occurrence in real-world scenarios. In this paper, we investigate the minima
x lower bound of the problem when the source and target distributions are known.
To achieve the minimax optimal bound (up to a logarithmic factor), we propose a
two-stage algorithm. Specifically, it first trains an optimal estimator for the
function under the source distribution, and then uses a likelihood ratio reweig
hting procedure to calibrate the moment estimator. In practice, the source and t
arget distributions are typically unknown, and estimating the likelihood ratio m
ay be unstable. To solve this problem, we propose a truncated version of the est
imator that ensures double robustness and provide the corresponding upper bound.
Extensive numerical studies on synthetic examples confirm our theoretical findi
ngs and further illustrate the effectiveness of our proposed method.
****
Changming Xu,Debangshu Banerjee,Deepak Vasisht,Gagandeep Singh
Support is All You Need for Certified VAE Training
Variational Autoencoders (VAEs) have become increasingly popular and deployed in
safety-critical applications. In such applications, we want to give certified p
robabilistic guarantees on performance under adversarial attacks. We propose a n
ovel method, CIVET, for certified training of VAEs. CIVET depends on the key ins
ight that we can bound worst-case VAE error by bounding the error on carefully c
hosen support sets at the latent layer. We show this point mathematically and pr
esent a novel training algorithm utilizing this insight. We show in an extensive
evaluation across different datasets (in both the wireless and vision applicati
on areas), architectures, and perturbation magnitudes that our method outperform
s SOTA methods achieving good standard performance with strong robustness guaran
tees.
****
Tanishq Kumar,Blake Bordelon,Cengiz Pehlevan,Venkatesh N Murthy,Samuel J. Gershm
an
Do Mice Grok? Glimpses of Hidden Progress in Sensory Cortex
Does learning of task-relevant representations stop when behavior stops changing
? Motivated by recent work in machine learning and the intuitive observation tha
t human experts continue to learn after mastery, we hypothesize that task-specif
ic representation learning in cortex can continue, even when behavior saturates.
In a novel reanalysis of recently published neural data, we find evidence for s
uch learning in posterior piriform cortex of mice following continued training o
n a task, long after behavior saturates at near-ceiling performance ("overtraini
ng"). We demonstrate that class representations in cortex continue to separate d
uring overtraining, so that examples that were incorrectly classified at the beg
inning of overtraining can abruptly be correctly classified later on, despite no
changes in behavior during that time. We hypothesize this hidden learning takes
the form of approximate margin maximization; we validate this and other predict
ions in the neural data, as well as build and interpret a simple synthetic model
that recapitulates these phenomena. We conclude by demonstrating how this model
of late-time feature learning implies an explanation for the empirical puzzle o
f overtraining reversal in animal learning, where task-specific representations
are more robust to particular task changes because the learned features can be r
eused.
****
Limei Wang,Kaveh Hassani,Si Zhang,Dongqi Fu,Baichuan Yuan,Weilin Cong,Zhigang Hu
a,Hao Wu,Ning Yao,Bo Long
Learning Graph Quantized Tokenizers
Transformers serve as the backbone architectures of Foundational Models, where d
omain-specific tokenizers allow them to adapt to various domains. Graph Transfor
mers (GTs) have recently emerged as leading models in geometric deep learning, o
utperforming Graph Neural Networks (GNNs) in various graph learning tasks. Howev
er, the development of tokenizers for graphs has lagged behind other modalities,

with existing approaches relying on heuristics or GNNs co-trained with Transformers. To address this, we introduce GQT (\textbf{G}raph \textbf{Q}uantized \textbf{T}okenizer), which decouples tokenizer training from Transformer training by leveraging multi-task graph self-supervised learning, yielding robust and generalizable graph tokens. Furthermore, the GQT utilizes Residual Vector Quantization (RVQ) to learn hierarchical discrete tokens, resulting in significantly reduced memory requirements and improved generalization capabilities. By combining the GQT with token modulation, a Transformer encoder achieves state-of-the-art performance on 20 out of 22 benchmarks, including large-scale homophilic and heterophilic datasets. The implementation is publicly available at \href{https://github.com/limei0307/GQT}{https://github.com/limei0307/GQT}.
****

Weizhong Huang,Yuxin Zhang,Xiawu Zheng,Liuyang,Jing Lin,Yiwu Yao,Rongrong Ji
Dynamic Low-Rank Sparse Adaptation for Large Language Models
Despite the efficacy of network sparsity in alleviating the deployment strain of Large Language Models (LLMs), it endures significant performance degradation. Applying Low-Rank Adaptation (LoRA) to fine-tune the sparse LLMs offers an intuitive approach to counter this predicament, while it holds shortcomings include: 1) The inability to integrate LoRA weights into sparse LLMs post-training, and 2) Insufficient performance recovery at high sparsity ratios. In this paper, we introduces dynamic $\textbf{Lo}$w-rank $\textbf{S}$parse $\textbf{A}$daptation $\textbf{(LoSA)}$, a novel method that seamlessly integrates low-rank adaptation into LLM sparsity within a unified framework, thereby enhancing the performance of sparse LLMs without increasing the inference latency. In particular, LoSA dynamically sparsifies the LoRA outcomes based on the corresponding sparse weights during fine-tuning, thus guaranteeing that the LoRA module can be integrated into the sparse LLMs post-training. Besides, to achieve the optimal sparse model architecture, LoSA leverages Representation Mutual Information (RMI) as an indicator to determine the importance of layers, thereby dynamically determining the optimal layer-wise sparsity rates during fine-tuning. Predicated on this, LoSA adjusts the rank of the LoRA module based on the variability in layer-wise reconstruction errors, allocating an appropriate fine-tuning for each layer to reduce the output discrepancies between dense and sparse LLMs. Extensive experiments tell that LoSA can efficiently boost the efficacy of sparse LLMs within a few hours, without introducing any additional inferential burden. For example, LoSA reduced the perplexity of sparse LLaMA-2-7B by $\textbf{68.73}$$\downarrow$ and increased zero-shot accuracy by $\textbf{16.32}$%$\uparrow$, achieving a $\textbf{2.60$\times$}$ speedup on CPU and $\textbf{2.23$\times$}$ speedup on GPU, requiring only $\textbf{45 minutes}$ of fine-tuning on $\textbf{a single}$ NVIDIA A100 80GB GPU. Code is available at https://github.com/wzhuang-xmu/LoSA.
****

Ke Yang,Yao Liu,Sapana Chaudhary,Rasool Fakoor,Pratik Chaudhari,George Karypis,Huzefa Rangwala
AgentOccam: A Simple Yet Strong Baseline for LLM-Based Web Agents
Autonomy via agents based on large language models (LLMs) that can carry out personalized yet standardized tasks presents a significant opportunity to drive human efficiency. There is an emerging need and interest in automating web tasks (e.g., booking a hotel for a given date within a budget). Being a practical use case itself, the web agent also serves as an important proof-of-concept example for various agent grounding scenarios, with its success promising advancements in many future applications. Meanwhile, much prior research focuses on handcrafting their web agent strategies (e.g., agent's prompting templates, reflective workflow, role-play and multi-agent systems, search or sampling methods, etc.) and the corresponding in-context examples. However, these custom strategies often struggle with generalizability across all potential real-world applications. On the other hand, there has been limited study on the misalignment between a web agent's observation and action representation, and the data on which the agent's underlying LLM has been pre-trained. This discrepancy is especially notable when LLMs are primarily trained for language completion rather than tasks involving embodied navigation actions and symbolic web elements. In our study, we enhance an

LLM-based web agent by simply refining its observation and action space, alignin
g these more closely with the LLM's capabilities. This approach enables our base
 agent to significantly outperform previous methods on a wide variety of web tas
ks. Specifically, on WebArena, a benchmark featuring general-purpose web interac
tion tasks, our agent AgentOccam surpasses the previous state-of-the-art and con
current work by 9.8 (+29.4%) and 5.9 (+15.8%) absolute points respectively, and
boosts the success rate by 26.6 points (+161%) over similar plain web agents wit
h its observation and action space alignment. Furthermore, on WebVoyager benchma
rk comprising tasks defined on real-world websites, AgentOccam exceeds the forme
r best agent by 2.4 points (+4.6%) on tasks with deterministic answers. We achie
ve this without using in-context examples, new agent roles, online feedback or s
earch strategies. AgentOccam's simple design highlights LLMs' impressive zero-sh
ot performance on web tasks, and underlines the critical role of carefully tunin
g observation and action spaces for LLM-based agents.
****

Zehan Qi,Xiao Liu,Iat Long Iong,Hanyu Lai,Xueqiao Sun,Jiadai Sun,Xinyue Yang,Yu
Yang,Shuntian Yao,Wei Xu,Jie Tang,Yuxiao Dong
WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement
 Learning
Large language models (LLMs) have shown remarkable potential as autonomous agent
s, particularly in web-based tasks.
However, existing LLM web agents face significant limitations: high-performing a
gents rely on expensive proprietary LLM APIs, while open LLMs lack the necessary
 decision-making capabilities.
This paper introduces WebRL, a novel self-evolving online curriculum reinforceme
nt learning framework designed to train high-performance web agents using open L
LMs.
Our approach addresses key challenges in this domain, including the scarcity of
training tasks, sparse feedback signals, and policy distribution drift in online
 learning.
WebRL incorporates a self-evolving curriculum that generates new tasks from unsu
ccessful attempts, a robust outcome-supervised reward model (ORM), and adaptive
reinforcement learning strategies to ensure consistent improvement.
We apply WebRL to transform Llama-3.1 models into proficient web agents, achievi
ng remarkable results on the WebArena-Lite benchmark.
Our Llama-3.1-8B agent improves from an initial 4.8\% success rate to 42.4\%, wh
ile the Llama-3.1-70B agent achieves a 47.3\% success rate across five diverse w
ebsites.
These results surpass the performance of GPT-4-Turbo (17.6\%) by over 160\% rela
tively and significantly outperform previous state-of-the-art web agents trained
 on open LLMs (AutoWebGLM, 18.2\%).
Our findings demonstrate WebRL's effectiveness in bridging the gap between open
and proprietary LLM-based web agents, paving the way for more accessible and pow
erful autonomous web interaction systems.
****

Bowen Jin,Jinsung Yoon,Jiawei Han,Sercan O Arik
Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG
Retrieval-augmented generation (RAG) empowers large language models (LLMs) to ut
ilize external knowledge sources. The increasing capacity of LLMs to process lon
ger input sequences opens up avenues for providing more retrieved information, t
o potentially enhance the quality of generated outputs. From a long-context LLM
 perspective, it assumes that a larger retrieval set would contain more relevant
 information (higher recall), that might result in improved performance. However
, our empirical findings demonstrate that for many long-context LLMs, the qualit
y of generated output initially improves first, but then subsequently declines a
s the number of retrieved passages increases. This paper investigates this pheno
menon, identifying the detrimental impact of retrieved "hard negatives" as a key
 contributor. To mitigate this and enhance the robustness of long-context LLM-ba
sed RAG, we propose both training-free and training-based approaches. We first s
howcase the effectiveness of retrieval reordering as a simple yet powerful train

ing-free optimization. Furthermore, we explore training-based methods, specifica
lly RAG-specific implicit LLM fine-tuning and RAG-oriented fine-tuning with inte
rmediate reasoning, demonstrating their capacity for substantial performance gai
ns. Finally, we conduct a systematic analysis of design choices for these traini
ng-based methods, including data distribution, retriever selection, and training
 context length.
****
Leander Girrbach,Stephan Alaniz,Yiran Huang,Trevor Darrell,Zeynep Akata
Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs)
Pre-trained large language models (LLMs) have been reliably integrated with visu
al input for multimodal tasks. The widespread adoption of instruction-tuned imag
e-to-text vision-language assistants (VLAs) like LLaVA and InternVL necessitates
 evaluating gender biases. We study gender bias in 22 popular open-source VLAs w
ith respect to personality traits, skills, and occupations. Our results show tha
t VLAs replicate human biases likely present in the data, such as real-world occ
upational imbalances. Similarly, they tend to attribute more skills and positive
 personality traits to women than to men, and we see a consistent tendency to as
sociate negative personality traits with men. To eliminate the gender bias in th
ese models, we find that fine-tuning-based debiasing methods achieve the best tr
ade-off between debiasing and retaining performance on downstream tasks. We argu
e for pre-deploying gender bias assessment in VLAs and motivate further developm
ent of debiasing strategies to ensure equitable societal outcomes. Code is avail
able at https://github.com/ExplainableML/vla-gender-bias.
****
Feng Li,Renrui Zhang,Hao Zhang,Yuanhan Zhang,Bo Li,Wei Li,Zejun MA,Chunyuan Li
LLaVA-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models
Visual instruction tuning has made considerable strides in enhancing the capabil
ities of Large Multimodal Models (LMMs). However, existing open LMMs largely foc
us on single-image tasks, their applications to multi-image scenarios remains le
ss explored. Additionally, prior LMM research separately tackles different scena
rios, leaving it impossible to generalize cross scenarios with new
emerging capabilities. To this end, we introduce LLaVA-Interleave, which simulta
neously tackles Multi-image, Multi-frame (video), Multi-view (3D), and Multi-pat
ch (single-image) scenarios in LMMs. To enable these capabilities, we regard the
 interleaved data format as a general template and compile the M4-Instruct datas
et with 1,177.6k samples, spanning 4 primary domains with 14
tasks and 41 datasets. We also curate the LLaVA-Interleave Bench to comprehensiv
ely evaluate the multi-image performance of LMMs. Through extensive
experiments, LLaVA-Interleave achieves leading results in multi-image, video,
and 3D benchmarks, while maintaining the performance of single-image tasks.
Besides, our model also exhibits several emerging capabilities, e.g., transferri
ng tasks across different settings and modalities.
****
Görkay Aydemir,Xiongyi Cai,Weidi Xie,Fatma Guney
Track-On: Transformer-based Online Point Tracking with Memory
In this paper, we consider the problem of long-term point tracking, which requir
es consistent identification of points across multiple frames in a video, despit
e changes in appearance, lighting, perspective, and occlusions. We target online
 tracking on a frame-by-frame basis, making it suitable for real-world, streamin
g scenarios. Specifically, we introduce Track-On, a simple transformer-based mod
el designed for online long-term point tracking. Unlike prior methods that depen
d on full temporal modeling, our model processes video frames causally without a
ccess to future frames, leveraging two memory modules —spatial memory and contex
t memory— to capture temporal information and maintain reliable point tracking o
ver long time horizons. At inference time, it employs patch classification and r
efinement to identify correspondences and track points with high accuracy. Throu
gh extensive experiments, we demonstrate that Track-On sets a new state-of-the-a
rt for online models and delivers superior or competitive results compared to of
fline approaches on seven datasets, including the TAP-Vid benchmark. Our method
offers a robust and scalable solution for real-time tracking in diverse applicat

ions. Project page: https://kuis-ai.github.io/track_on
****
Brian Nlong Zhao,Yuhang Xiao,Jiashu Xu,XINYANG JIANG,Yifan Yang,Dongsheng Li,Laurent Itti,Vibhav Vineet,Yunhao Ge
DreamDistribution: Learning Prompt Distribution for Diverse In-distribution Generation
The popularization of Text-to-Image (T2I) diffusion models enables the generation of high-quality images from text descriptions. However, generating diverse customized images with reference visual attributes remains challenging. This work focuses on personalizing T2I diffusion models at a more abstract concept or category level, adapting commonalities from a set of reference images while creating new instances with sufficient variations. We introduce a solution that allows a pretrained T2I diffusion model to learn a set of soft prompts, enabling the generation of novel images by sampling prompts from the learned distribution. These prompts offer text-guided editing capabilities and additional flexibility in controlling variation and mixing between multiple distributions. We also show the adaptability of the learned prompt distribution to other tasks, such as text-to-3D. Finally we demonstrate effectiveness of our approach through quantitative analysis including automatic evaluation and human assessment.
****
Haiyang Wang,Yue Fan,Muhammad Ferjad Naeem,Yongqin Xian,Jan Eric Lenssen,Liwei Wang,Federico Tombari,Bernt Schiele
TokenFormer: Rethinking Transformer Scaling with Tokenized Model Parameters
Transformers have become the predominant architecture in foundation models due to their excellent performance across various domains. However, the substantial cost of scaling these models remains a significant concern. This problem arises primarily from their dependence on a fixed number of parameters within linear projections. When architectural modifications (e.g., channel dimensions) are introduced, the entire model typically requires retraining from scratch. As model sizes continue growing, this strategy results in increasingly high computational costs and becomes unsustainable. To overcome this problem, we introduce Tokenformer, a natively scalable architecture that leverages the attention mechanism not only for computations among input tokens but also for interactions between tokens and model parameters, thereby enhancing architectural flexibility. By treating model parameters as tokens, we replace all the linear projections in Transformers with our token-parameter attention layer, where input tokens act as queries and model parameters as keys and values. This reformulation allows for progressive and efficient scaling without necessitating retraining from scratch. Our model scales from 124M to 1.4B parameters by incrementally adding new key-value parameter pairs, achieving performance comparable to Transformers trained from scratch while greatly reducing training costs. Code and models are available at {\color{red}\url{https://github.com/Haiyang-W/TokenFormer.git}}
****
Jonghyeok Lee,Chen Xu,Yao Xie
Kernel-based Optimally Weighted Conformal Time-Series Prediction
Conformal prediction has been a popular distribution-free framework for uncertainty quantification. In this work, we present a novel conformal prediction method for time-series, which we call Kernel-based Optimally Weighted Conformal Prediction Intervals ($\texttt{KOWCPI}$). Specifically, $\texttt{KOWCPI}$ adapts the classic Reweighted Nadaraya-Watson (RNW) estimator for quantile regression on dependent data and learns optimal data-adaptive weights. Theoretically, we tackle the challenge of establishing a conditional coverage guarantee for non-exchangeable data under strong mixing conditions on the non-conformity scores. We demonstrate the superior performance of $\texttt{KOWCPI}$ on real time-series against state-of-the-art methods, where $\texttt{KOWCPI}$ achieves narrower confidence intervals without losing coverage.
****
Hanzhu Chen,Xu Shen,Jie Wang,Zehao Wang,Qitan Lv,Junjie He,Rong Wu,Feng Wu,Jieping Ye
Knowledge Graph Finetuning Enhances Knowledge Manipulation in Large Language Mod

els
Despite the impressive performance of general large language models(LLMs), many of their applications in specific domains (e.g., low-data and knowledge-intensive) still confront significant challenges. Supervised fine-tuning (SFT)---where a general LLM is further trained on a small labeled dataset to adapt for specific tasks or domains---has shown great power for developing domain-specific LLMs. However, existing SFT data primarily consist of Question and Answer (Q&A) pairs, which poses a significant challenge for LLMs to comprehend the correlation and logic of knowledge underlying the Q&A. To address this challenge, we propose a conceptually flexible and general framework to boost SFT, namely Knowledge Graph-Driven Supervised Fine-Tuning (KG-SFT). The key idea of KG-SFT is to generate high-quality explanations for each Q&A pair via a structured knowledge graph to enhance the knowledge comprehension and manipulation of LLMs. Specifically, KG-SFT consists of three components: Extractor, Generator, and Detector. For a given Q&A pair, (i) Extractor first identifies entities within Q&A pairs and extracts relevant reasoning subgraphs from external KGs, (ii) Generator then produces corresponding fluent explanations utilizing these reasoning subgraphs, and (iii) finally, Detector performs sentence-level knowledge conflicts detection on these explanations to guarantee the reliability. KG-SFT focuses on generating high-quality explanations to improve the quality of Q&A pair, which reveals a promising direction for supplementing existing data augmentation methods. Extensive experiments on fifteen different domains and six different languages demonstrate the effectiveness of KG-SFT, leading to an accuracy improvement of up to 18% and an average of 8.7% in low-data scenarios.
****
Xinhe Wang,Pingbang Hu,Junwei Deng,Jiaqi W. Ma
Adversarial Attacks on Data Attribution
Data attribution aims to quantify the contribution of individual training data points to the outputs of an AI model, which has been used to measure the value of training data and compensate data providers. Given the impact on financial decisions and compensation mechanisms, a critical question arises concerning the adversarial robustness of data attribution methods. However, there has been little to no systematic research addressing this issue. In this work, we aim to bridge this gap by detailing a threat model with clear assumptions about the adversary's goal and capabilities and proposing principled adversarial attack methods on data attribution. We present two methods, *Shadow Attack* and *Outlier Attack*, which generate manipulated datasets to inflate the compensation adversarially. The Shadow Attack leverages knowledge about the data distribution in the AI applications, and derives adversarial perturbations through "shadow training", a technique commonly used in membership inference attacks. In contrast, the Outlier Attack does not assume any knowledge about the data distribution and relies solely on black-box queries to the target model's predictions. It exploits an inductive bias present in many data attribution methods - outlier data points are more likely to be influential - and employs adversarial examples to generate manipulated datasets. Empirically, in image classification and text generation tasks, the Shadow Attack can inflate the data-attribution-based compensation by at least 200%, while the Outlier Attack achieves compensation inflation ranging from 185% to as much as 643%. Our implementation is ready at https://github.com/TRAIS-Lab/adversarial-attack-data-attribution
****
Isaac Reid,Stratis Markou,Krzysztof Marcin Choromanski,Richard E. Turner,Adrian Weller
Variance-Reducing Couplings for Random Features
Random features (RFs) are a popular technique to scale up kernel methods in machine learning, replacing exact kernel evaluations with stochastic Monte Carlo estimates. They underpin models as diverse as efficient transformers (by approximating attention) to sparse spectrum Gaussian processes (by approximating the covariance function). Efficiency can be further improved by speeding up the convergence of these estimates: a variance reduction problem. We tackle this through the unifying lens of optimal transport, finding couplings to improve RFs defined on

both Euclidean and discrete input spaces. They enjoy theoretical guarantees and sometimes provide strong downstream gains, including for scalable inference on graphs. We reach surprising conclusions about the benefits and limitations of variance reduction as a paradigm, showing that other properties of the coupling should be optimised for attention estimation in efficient transformers.
****
Junyao Gao,Yanan SUN,Fei Shen,Xin Jiang,Zhening Xing,Kai Chen,Cairong Zhao
FaceShot: Bring Any Character into Life
In this paper, we present ***FaceShot***, a novel training-free portrait animation framework designed to bring any character into life from any driven video without fine-tuning or retraining.
We achieve this by offering precise and robust reposed landmark sequences from an appearance-guided landmark matching module and a coordinate-based landmark retargeting module.
Together, these components harness the robust semantic correspondences of latent diffusion models to produce facial motion sequence across a wide range of character types.
After that, we input the landmark sequences into a pre-trained landmark-driven animation model to generate animated video.
With this powerful generalization capability, FaceShot can significantly extend the application of portrait animation by breaking the limitation of realistic portrait landmark detection for any stylized character and driven video.
Also, FaceShot is compatible with any landmark-driven animation model, significantly improving overall performance.
Extensive experiments on our newly constructed character benchmark CharacBench confirm that FaceShot consistently surpasses state-of-the-art (SOTA) approaches across any character domain.
More results are available at our project website https://faceshot2024.github.io/faceshot/.
****
Hsun-Yu Kuo,Yin-Hsiang Liao,Yu-Chieh Chao,Wei-Yun Ma,Pu-Jen Cheng
Not All LLM-Generated Data Are Equal: Rethinking Data Weighting in Text Classification
Synthetic data augmentation via Large Language Models (LLMs) allows researchers to leverage additional training data, thus enhancing the performance of downstream tasks, especially when real-world data is scarce. However, the generated data can deviate from the real-world data, and this misalignment can bring about deficient results while applying the trained model to applications. Therefore, we proposed efficient weighted-loss approaches to align synthetic data with real-world distribution by emphasizing high-quality and diversified data generated by LLMs using merely a tiny amount of real-world data. We empirically assessed the effectiveness of our methods on multiple text classification tasks, and the results showed that leveraging our approaches on a BERT-level model robustly outperformed standard cross-entropy and other data weighting approaches, providing potential solutions to effectively leveraging synthetic data from any suitable data generator.
****
Shengjie Ma,Chengjin Xu,Xuhui Jiang,Muzhi Li,Huaren Qu,Cehao Yang,Jiaxin Mao,Jian Guo
Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation
Retrieval-augmented generation (RAG) has improved large language models (LLMs) by using knowledge retrieval to overcome knowledge deficiencies. However, current RAG methods often fall short of ensuring the depth and completeness of retrieved information, which is necessary for complex reasoning tasks. In this work, we introduce Think-on-Graph 2.0 (ToG-2), a hybrid RAG framework that iteratively retrieves information from both unstructured and structured knowledge sources in a tight-coupling manner. Specifically, ToG-2 leverages knowledge graphs (KGs) to link documents via entities, facilitating deep and knowledge-guided context retrieval. Simultaneously, it utilizes documents as entity contexts to achieve preci

se and efficient graph retrieval.
ToG-2 alternates between graph retrieval and context retrieval to search for in-depth clues relevant to the question, enabling LLMs to generate answers.
We conduct a series of well-designed experiments to highlight the following advantages of ToG-2: 1) ToG-2 tightly couples the processes of context retrieval and graph retrieval, deepening context retrieval via the KG while enabling reliable graph retrieval based on contexts; 2) it achieves deep and faithful reasoning in LLMs through an iterative knowledge retrieval process of collaboration between contexts and the KG; and 3) ToG-2 is training-free and plug-and-play compatible with various LLMs. Extensive experiments demonstrate that ToG-2 achieves overall state-of-the-art (SOTA) performance on 6 out of 7 knowledge-intensive datasets with GPT-3.5, and can elevate the performance of smaller models (e.g., LLAMA-2-13B) to the level of GPT-3.5’s direct reasoning. The source code is available on https://anonymous.4open.science/r/ToG2.
****
Aiwei Liu,Haoping Bai,Zhiyun Lu,Yanchao Sun,Xiang Kong,Xiaoming Simon Wang,Jiulong Shan,Albin Madappally Jose,Xiaojiang Liu,Lijie Wen,Philip S. Yu,Meng Cao
TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights
Direct Preference Optimization (DPO) has been widely adopted for preference alignment of Large Language Models (LLMs) due to its simplicity and effectiveness.
However, DPO is derived as a bandit problem in which the whole response is treated as a single arm, ignoring the importance differences between tokens, which may affect optimization efficiency and make it difficult to achieve optimal results.
In this work, we propose that the optimal data for DPO has equal expected rewards for each token in winning and losing responses, as there is no difference in token importance.
However, since the optimal dataset is unavailable in practice, we propose using the original dataset for importance sampling to achieve unbiased optimization.
Accordingly, we propose a token-level importance sampling DPO objective named TIS-DPO that assigns importance weights to each token based on its reward.
Inspired by previous works, we estimate the token importance weights using the difference in prediction probabilities from a pair of contrastive LLMs. We explore three methods to construct these contrastive LLMs: (1) guiding the original LLM with contrastive prompts, (2) training two separate LLMs using winning and losing responses, and (3) performing forward and reverse DPO training with winning and losing responses.
Experiments show that TIS-DPO significantly outperforms various baseline methods on harmlessness and helpfulness alignment and summarization tasks. We also visualize the estimated weights, demonstrating their ability to identify key token positions.
****
Zeyuan Allen-Zhu,Yuanzhi Li
Physics of Language Models: Part 3.2, Knowledge Manipulation
Language models can store vast factual knowledge, yet their ability to flexibly use this knowledge for downstream tasks (e.g., via instruction finetuning) remains questionable. This paper investigates four fundamental knowledge manipulation tasks: \textbf{retrieval} (e.g., "What is person A's attribute X?"), \textbf{classification} (e.g., "Is A's attribute X even or odd?"), \textbf{comparison} (e.g., "Is A greater than B in attribute X?"), and \textbf{inverse search} (e.g., "Which person's attribute X equals T?").

We show that language models excel in knowledge retrieval but struggle even in the simplest classification or comparison tasks unless Chain of Thoughts (CoTs) are employed during both training and inference. Moreover, their performance in inverse knowledge search is virtually 0\%, regardless of the prompts.
Our primary contribution is a \emph{controlled, synthetic experiment} that confirms these weaknesses are \emph{inherent} to language models: they cannot efficiently manipulate knowledge from pre-training data, even when such knowledge is pe

rfectly stored in the models, despite adequate training and sufficient model siz
e. Our findings also apply to modern pretrained language models such as GPT-4, t
hus giving rise to many Turing tests to distinguish Humans from contemporary AIs
.
****
Zhen Zhang,Ignavier Ng,Dong Gong,Yuhang Liu,Mingming Gong,Biwei Huang,Kun Zhang,
Anton van den Hengel,Javen Qinfeng Shi
Analytic DAG Constraints for Differentiable DAG Learning
Recovering the underlying Directed Acyclic Graph (DAG)
structures from observational data presents a formidable challenge, partly due
to the combinatorial nature of the DAG-constrained optimization
problem. Recently, researchers have identified gradient vanishing as
one of the primary obstacles in differentiable DAG learning and have
proposed several DAG constraints to mitigate this issue. By developing
the necessary theory to establish a connection between analytic
functions and DAG constraints, we demonstrate that analytic functions
from the set $\\{f(x) = c_0 + \\sum_{i=1}^{\infty}c_ix^i | \\forall i > 0, c_i
> 0; r = \\lim_{i\\rightarrow \\infty}c_{i}/c_{i+1} > 0\\}$ can be employed to
formulate effective DAG constraints. Furthermore, we establish that
this set of functions is closed under several functional operators,
including differentiation, summation, and
multiplication. Consequently, these operators can be leveraged to
create novel DAG constraints based on existing ones. Using these
properties, we design a series of DAG constraints and develop an
efficient algorithm to evaluate them. Experiments
in various settings demonstrate that our DAG constraints
outperform previous state-of-the-art comparators. Our implementation is availa
ble at https://github.com/zzhang1987/AnalyticDAGLearning.
****
Alexander Cong Li,Ananya Kumar,Deepak Pathak
Generative Classifiers Avoid Shortcut Solutions
Discriminative approaches to classification often learn shortcuts that hold in-d
istribution but fail even under minor distribution shift. This failure mode stem
s from an overreliance on features that are spuriously correlated with the label
. We show that generative classifiers, which use class-conditional generative mo
dels, can avoid this issue by modeling all features, both core and spurious, ins
tead of mainly spurious ones. These generative classifiers are simple to train,
avoiding the need for specialized augmentations, strong regularization, extra hy
perparameters, or knowledge of the specific spurious correlations to avoid. We f
ind that diffusion-based and autoregressive generative classifiers achieve state
-of-the-art performance on five standard image and text distribution shift bench
marks and reduce the impact of spurious correlations in realistic applications,
such as medical or satellite datasets. Finally, we carefully analyze a Gaussian
toy setting to understand the inductive biases of generative classifiers, as wel
l as the data properties that determine when generative classifiers outperform d
iscriminative ones.
****
Jan-Matthis Lueckmann,Alexander Immer,Alex Bo-Yuan Chen,Peter H. Li,Mariela D Pe
tkova,Nirmala A Iyer,Luuk Willem Hesselink,Aparna Dev,Gudrun Ihrke,Woohyun Park,
Alyson Petruncio,Aubrey Weigel,Wyatt Korff,Florian Engert,Jeff Lichtman,Misha Ah
rens,Michal Januszewski,Viren Jain
ZAPBench: A Benchmark for Whole-Brain Activity Prediction in Zebrafish
Data-driven benchmarks have led to significant progress in key scientific modeli
ng domains including weather and structural biology. Here, we introduce the Zebr
afish Activity Prediction Benchmark (ZAPBench) to measure progress on the proble
m of predicting cellular-resolution neural activity throughout an entire vertebr
ate brain. The benchmark is based on a novel dataset containing 4d light-sheet m
icroscopy recordings of over 70,000 neurons in a larval zebrafish brain, along w
ith motion stabilized and voxel-level cell segmentations of these data that faci
litate development of a variety of forecasting methods. Initial results from a s

election of time series and volumetric video modeling approaches achieve better performance than naive baseline methods, but also show room for further improvem ent. The specific brain used in the activity recording is also undergoing synapt ic-level anatomical mapping, which will enable future integration of detailed st ructural information into forecasting methods.
****
Hui Chen,Viet Luong,Lopamudra Mukherjee,Vikas Singh
SimpleTM: A Simple Baseline for Multivariate Time Series Forecasting
The versatility of large Transformer-based models has led to many efforts focuse d on adaptations to other modalities, including time-series data. For instance, one could start from a pre-trained checkpoint of a large language model and atta ch adapters to recast
the new modality (e.g., time-series) as ``language''. Alternatively, one can use a suitably large Transformer-based model, and make some modifications for time-series data. These ideas offer good performance across available benchmarks. But temporal data are quite heterogeneous (e.g., wearable sensors, physiological me asurements in healthcare), and unlike text/image corpus, much of it is not publi cly available. So, these models need a fair bit of domain-specific fine-tuning t o achieve good performance -- this is often expensive or difficult with limited resources. In this paper, we study and characterize the performance profile of a non-generalist approach: our SimpleTM model is specialized for multivariate tim e-series forecasting. By simple, we mean that the model is lightweight. It is re stricted to tokenization based on textbook signal processing ideas (shown to be effective in vision) which are then allowed to attend/interact: via self-attenti on but also via ways that are a bit more general than dot-product attention, acc omplished via basic geometric algebra operations. We show that even a single- or two-layer model gives results that are competitive with much bigger models, inc luding large transformer-based architectures,
on most benchmarks commonly reported in the literature.
****
Ricardo Buitrago,Tanya Marwah,Albert Gu,Andrej Risteski
On the Benefits of Memory for Modeling Time-Dependent PDEs
Data-driven techniques have emerged as a promising alternative to traditional n umerical methods for solving PDEs. For time-dependent PDEs, many approaches are Markovian---the evolution of the trained system only depends on the current stat e, and not the past states. In this work, we investigate the benefits of using m emory for modeling time-dependent PDEs: that is, when past states are explicitly used to predict the future. Motivated by the Mori-Zwanzig theory of model reduc tion, we theoretically exhibit examples of simple (even linear) PDEs, in which a solution that uses memory is arbitrarily better than a Markovian solution. Addi tionally, we introduce Memory Neural Operator (MemNO), a neural operator archite cture that combines recent state space models (specifically, S4) and Fourier Neu ral Operators (FNOs) to effectively model memory. We empirically demonstrate th at when the PDEs are supplied in low resolution or contain observation noise at train and test time, MemNO significantly outperforms the baselines without memor y---with up to $6 \times$ reduction in test error. Furthermore, we show that thi s benefit is particularly pronounced when the PDE solutions have significant hig h-frequency Fourier modes (e.g., low-viscosity fluid dynamics) and we construct a challenging benchmark dataset consisting of such PDEs.
****
Dingrong Wang,Krishna Prasad Neupane,Ervine Zheng,Qi Yu
Looking into User’s Long-term Interests through the Lens of Conservative Evident ial Learning
Reinforcement learning (RL) provides an effective means to capture users' evolvi ng preferences, leading to improved recommendation performance over time. Howeve r, existing RL approaches primarily rely on standard exploration strategies, whi ch are less effective for a large item space with sparse reward signals given th e limited interactions for most users. Therefore, they may not be able to learn the optimal policy that effectively captures user's evolving preferences and ach ieves the maximum expected reward over the long term. In this paper, we propose

a novel evidential conservative Q-learning framework (ECQL) that learns an effec
tive and conservative recommendation policy by integrating evidence-based uncert
ainty and conservative learning. ECQL conducts evidence-aware explorations to di
scover items that are located beyond current observations but reflect users' lon
g-term interests. It offers an uncertainty-aware conservative view on policy eva
luation to discourage deviating too much from users' current interests. Two cent
ral components of ECQL include a uniquely designed sequential state encoder and
a novel conservative evidential-actor-critic (CEAC) module. The former generates
 the current state of the environment by aggregating historical information and
a sliding window that contains the current user interactions as well as newly re
commended items from RL exploration that may represent short and long-term inter
ests respectively. The latter performs an evidence-based rating prediction by ma
ximizing the conservative evidential Q-value and leverages an uncertainty-aware
ranking score to explore the item space for a more diverse and valuable recommen
dation. Experiments on multiple real-world dynamic datasets demonstrate the stat
e-of-the-art performance of ECQL and its capability to capture users' long-term
interests.
****

Jinheng Xie,Weijia Mao,Zechen Bai,David Junhao Zhang,Weihao Wang,Kevin Qinghong
Lin,Yuchao Gu,Zhijie Chen,Zhenheng Yang,Mike Zheng Shou
Show-o: One Single Transformer to Unify Multimodal Understanding and Generation
We present a unified transformer, i.e., Show-o, that unifies multimodal understa
nding and generation. Unlike fully autoregressive models, Show-o unifies autoreg
ressive and (discrete) diffusion modeling to adaptively handle inputs and output
s of various and mixed modalities. The unified model flexibly supports a wide ra
nge of vision-language tasks including visual question-answering, text-to-image
generation, text-guided inpainting/extrapolation, and mixed-modality generation.
 Across various benchmarks, it demonstrates comparable or superior performance t
o existing individual models with an equivalent or larger number of parameters t
ailored for understanding or generation. This significantly highlights its poten
tial as a next-generation foundation model.
****

Safwan Hossain,Evi Micha,Yiling Chen,Ariel D. Procaccia
Strategic Classification With Externalities
We propose a new variant of the strategic classification problem: a principal re
veals a classifier, and n agents report their (possibly manipulated) features
to be classified. Motivated by real-world applications, our model crucially allo
ws the manipulation of one agent to affect another; that is, it explicitly captu
res inter-agent externalities. The principal-agent interactions are formally mod
eled as a Stackelberg game, with the resulting agent manipulation dynamics captu
red as a simultaneous game. We show that under certain assumptions, the pure Nas
h Equilibrium of this agent manipulation game is unique and can be efficiently c
omputed. Leveraging this result, PAC learning guarantees are established for the
 learner: informally, we show that it is possible to learn classifiers that mini
mize loss on the distribution, even when a random number of agents are manipulat
ing their way to a pure Nash Equilibrium. We also comment on the optimization of
 such classifiers through gradient-based approaches. This work sets the theoreti
cal foundations for a more realistic analysis of classifiers that are robust aga
inst multiple strategic actors interacting in a common environment.
****

Zhengzhuo Xu,Bowen Qu,Yiyan Qi,SiNan Du,Chengjin Xu,Chun Yuan,Jian Guo
ChartMoE: Mixture of Diversely Aligned Expert Connector for Chart Understanding
Automatic chart understanding is crucial for content comprehension and document
parsing. Multimodal Large Language Models (MLLMs) have demonstrated remarkable c
apabilities in chart understanding through domain-specific alignment and fine-tu
ning. However, current MLLMs still struggle to provide faithful data and reliabl
e analysis only based on charts. To address it, we propose ChartMoE, which emplo
ys the Mixture of Expert (MoE) architecture to replace the traditional linear pr
ojector to bridge the modality gap. Specifically, we train several linear connec
tors through distinct alignment tasks, which are utilized as the foundational in

itialization parameters for different experts. Additionally, we introduce ChartMoE-Align, a dataset with nearly 1 million chart-table-JSON-code quadruples to conduct three alignment tasks (chart-table/JSON/code). Combined with the vanilla connector, we initialize different experts diversely and adopt high-quality knowledge learning to further refine the MoE connector and LLM parameters. Extensive experiments demonstrate the effectiveness of the MoE connector and our initialization strategy, e.g., ChartMoE improves the accuracy of the previous state-of-the-art from 80.48% to 84.64% on the ChartQA benchmark.
****

Carl Qi,Dan Haramati,Tal Daniel,Aviv Tamar,Amy Zhang
EC-Diffuser: Multi-Object Manipulation via Entity-Centric Behavior Generation
Object manipulation is a common component of everyday tasks, but learning to manipulate objects from high-dimensional observations presents significant challenges. These challenges are heightened in multi-object environments due to the combinatorial complexity of the state space as well as of the desired behaviors. While recent approaches have utilized large-scale offline data to train models from pixel observations, achieving performance gains through scaling, these methods struggle with compositional generalization in unseen object configurations with constrained network and dataset sizes. To address these issues, we propose a novel behavioral cloning (BC) approach that leverages object-centric representations and an entity-centric Transformer with diffusion-based optimization, enabling efficient learning from offline image data. Our method first decomposes observations into Deep Latent Particles (DLP), which are then processed by our entity-centric Transformer that computes attention at the particle level, simultaneously predicting object dynamics and the agent's actions. Combined with the ability of diffusion models to capture multi-modal behavior distributions, this results in substantial performance improvements in multi-object tasks and, more importantly, enables compositional generalization. We present BC agents capable of zero-shot generalization to perform tasks with novel compositions of objects and goals, including larger numbers of objects than seen during training. We provide video rollouts on our webpage: https://sites.google.com/view/ec-diffuser.
****

Maxence Faldor,Antoine Cully
CAX: Cellular Automata Accelerated in JAX
Cellular automata have become a cornerstone for investigating emergence and self-organization across diverse scientific disciplines. However, the absence of a hardware-accelerated cellular automata library limits the exploration of new research directions, hinders collaboration, and impedes reproducibility. In this work, we introduce CAX (Cellular Automata Accelerated in JAX), a high-performance and flexible open-source library designed to accelerate cellular automata research. CAX delivers cutting-edge performance through hardware acceleration while maintaining flexibility through its modular architecture, intuitive API, and support for both discrete and continuous cellular automata in arbitrary dimensions. We demonstrate CAX's performance and flexibility through a wide range of benchmarks and applications. From classic models like elementary cellular automata and Conway's Game of Life to advanced applications such as growing neural cellular automata and self-classifying MNIST digits, CAX speeds up simulations up to 2,000 times faster. Furthermore, we demonstrate CAX's potential to accelerate research by presenting a collection of three novel cellular automata experiments, each implemented in just a few lines of code thanks to the library's modular architecture. Notably, we show that a simple one-dimensional cellular automaton can outperform GPT-4 on the 1D-ARC challenge.
****

Nilo Schwencke,Cyril Furtlehner
ANaGRAM: A Natural Gradient Relative to Adapted Model for efficient PINNs learning
In the recent years, Physics Informed Neural Networks (PINNs) have received strong interest as a method to solve PDE driven systems, in particular for data assimilation purpose. This method is still in its infancy, with many shortcomings and failures that remain not properly understood.

In this paper we propose a natural gradient approach to PINNs which contributes to speed-up and improve the accuracy of the training.
Based on an in depth analysis of the differential geometric structures of the problem, we come up with two distinct contributions:
(i) a new natural gradient algorithm that scales as $\min(P^2S, S^2P)$, where P is the number of parameters, and S the batch size;
(ii) a mathematically principled reformulation of the PINNs problem that allows the extension of natural gradient to it, with proved connections to Green's function theory.
****

Weize Chen,Ziming You,Ran Li,yitong guan,Chen Qian,Chenyang Zhao,Cheng Yang,Ruobing Xie,Zhiyuan Liu,Maosong Sun
Internet of Agents: Weaving a Web of Heterogeneous Agents for Collaborative Intelligence
The rapid advancement of large language models (LLMs) has paved the way for the development of highly capable autonomous agents. However, existing multi-agent frameworks often struggle with integrating diverse capable third-party agents due to reliance on agents defined within their own ecosystems. They also face challenges in simulating distributed environments, as most frameworks are limited to single-device setups. Furthermore, these frameworks often rely on hard-coded communication pipelines, limiting their adaptability to dynamic task requirements. Inspired by the concept of the Internet, we propose the Internet of Agents (IoA), a novel framework that addresses these limitations by providing a flexible and scalable platform for LLM-based multi-agent collaboration. IoA introduces an agent integration protocol, an instant-messaging-like architecture design, and dynamic mechanisms for agent teaming and conversation flow control. Through extensive experiments on general assistant tasks, embodied AI tasks, and retrieval-augmented generation benchmarks, we demonstrate that IoA consistently outperforms state-of-the-art baselines, showcasing its ability to facilitate effective collaboration among heterogeneous agents. IoA represents a step towards linking diverse agents in an Internet-like environment, where agents can seamlessly collaborate to achieve greater intelligence and capabilities. We will release our code to facilitate further research.
****

Jiawei Huang,Hu Ding
An Effective Manifold-based Optimization Method for Distributionally Robust Classification
How to promote the robustness of existing deep learning models is a challenging problem for many practical classification tasks. Recently, Distributionally Robust Optimization (DRO) methods have shown promising potential to tackle this problem. These methods aim to construct reliable models by minimizing the worst-case risk within a local region (called ''uncertainty set'') around the empirical data distribution. However, conventional DRO methods tend to be overly pessimistic, leading to certain discrepancy between the real data distribution and the uncertainty set, which can degrade the classification performance. To address this issue, we propose a manifold-based DRO method that takes the geometric structure of training data into account for constructing the uncertainty set. Specifically, our method employs a carefully designed ''game'' that integrates contrastive learning with Jacobian regularization to capture the manifold structure, enabling us to solve DRO problems constrained by the data manifold. By utilizing a novel idea for approximating geodesic distance on manifolds, we also provide the theoretical guarantees for its robustness. Moreover, our proposed method is easy to implement in practice. We conduct a set of experiments on several popular benchmark datasets, where the results demonstrate our advantages in terms of accuracy and robustness.
****

Yiwen Yuan,Zecheng Zhang,Xinwei He,Akihiro Nitta,Weihua Hu,Manan Shah,Blaž Stojanovi■,Shenyang Huang,Jan Eric Lenssen,Jure Leskovec,Matthias Fey
ContextGNN: Beyond Two-Tower Recommendation Systems
Recommendation systems predominantly utilize two-tower architectures, which eval

uate user-item rankings through the inner product of their respective embeddings. However, one key limitation of two-tower models is that they learn a pair-agnostic representation of users and items. In contrast, pair-wise representations either scale poorly due to their quadratic complexity or are too restrictive on the candidate pairs to rank. To address these issues, we introduce Context-based Graph Neural Networks (ContextGNNs), a novel deep learning architecture for link prediction in recommendation systems. The method employs a pair-wise representation technique for familiar items situated within a user's local subgraph, while leveraging two-tower representations to facilitate the recommendation of exploratory items. A final network then predicts how to fuse both pair-wise and two-tower recommendations into a single ranking of items. We demonstrate that ContextGNN is able to adapt to different data characteristics and outperforms existing methods, both traditional and GNN-based, on a diverse set of practical recommendation tasks, improving performance by 20\% on average.
****
Han Lin,Jaemin Cho,Abhay Zala,Mohit Bansal
Ctrl-Adapter: An Efficient and Versatile Framework for Adapting Diverse Controls to Any Diffusion Model
ControlNets are widely used for adding spatial control to text-to-image diffusion models. However, when it comes to controllable video generation, ControlNets cannot be directly integrated into new backbones due to feature space mismatches, and training ControlNets for new backbones can be a significant burden for many users. Furthermore, applying ControlNets independently to different frames cannot effectively maintain object temporal consistency. To address these challenges, we introduce Ctrl-Adapter, an efficient and versatile framework that adds diverse controls to any image/video diffusion models through the adaptation of pretrained ControlNets. Ctrl-Adapter offers strong and diverse capabilities, including image and video control, sparse-frame video control, fine-grained patch-level multi-condition control, zero-shot adaptation to unseen conditions, and supports a variety of downstream tasks beyond spatial control, including video editing, video style transfer, and text-guided motion control. With six diverse U-Net/DiT-based image/video diffusion models (SDXL, PixArt-α, I2VGen-XL, SVD, Latte, Hotshot-XL), Ctrl-Adapter matches the performance of pretrained ControlNets on COCO and achieves the state-of-the-art on DAVIS 2017 with significantly lower computation (< 10 GPU hours).
****
Gezheng Xu,Hui Guo,Li Yi,Charles Ling,Boyu Wang,Grace Yi
Revisiting Source-Free Domain Adaptation: a New Perspective via Uncertainty Control
Source-Free Domain Adaptation (SFDA) seeks to adapt a pre-trained source model to the target domain using only unlabeled target data, without access to the original source data. While current state-of-the-art (SOTA) methods rely on leveraging weak supervision from the source model to extract reliable information for self-supervised adaptation, they often overlook the uncertainty that arises during the transfer process. In this paper, we conduct a systematic and theoretical analysis of the uncertainty inherent in existing SFDA methods and demonstrate its impact on transfer performance through the lens of Distributionally Robust Optimization (DRO). Building upon the theoretical results, we propose a novel instance-dependent uncertainty control algorithm for SFDA. Our method is designed to quantify and exploit the uncertainty during the adaptation process, significantly improving the model performance. Extensive experiments on benchmark datasets and empirical analyses confirm the validity of our theoretical findings and the effectiveness of the proposed method.
This work offers new insights into understanding and advancing SFDA performance.
****
Takeru Miyato,Sindy Löwe,Andreas Geiger,Max Welling
Artificial Kuramoto Oscillatory Neurons
It has long been known in both neuroscience and AI that ``binding'' between neurons leads to a form of competitive learning where representations are compressed in order to represent more abstract concepts in deeper layers of the network. M

ore recently, it was also hypothesized that dynamic (spatiotemporal) representat
ions play an important role in both neuroscience and AI. Building on these ideas
, we introduce Artificial Kuramoto Oscillatory Neurons (*AKOrN*) as a dynamical
alternative to threshold units, which can be combined with arbitrary connectivit
y designs such as fully connected, convolutional, or attentive mechanisms. Our g
eneralized Kuramoto updates bind neurons together through their synchronization
dynamics. We show that this idea provides performance improvements across a wide
 spectrum of tasks such as unsupervised object discovery, adversarial robustness
, calibrated uncertainty quantification, and reasoning. We believe that these em
pirical results show the importance of rethinking our assumptions at the most ba
sic neuronal level of neural representation, and in particular show the importan
ce of dynamical representations.
****

Dimitris Oikonomou,Nicolas Loizou

Stochastic Polyak Step-sizes and Momentum: Convergence Guarantees and Practical
Performance

Stochastic gradient descent with momentum, also known as Stochastic Heavy Ball m
ethod (SHB), is one of the most popular algorithms for solving large-scale stoch
astic optimization problems in various machine learning tasks. In practical scen
arios, tuning the step-size and momentum parameters of the method is a prohibiti
vely expensive and time-consuming process. In this work, inspired by the recent
advantages of stochastic Polyak step-size in the performance of stochastic gradi
ent descent (SGD), we propose and explore new Polyak-type variants suitable for
the update rule of the SHB method. In particular, using the Iterate Moving Avera
ge (IMA) viewpoint of SHB, we propose and analyze three novel step-size selectio
ns: MomSPSmax, MomDecSPS, and MomAdaSPS. For MomSPSmax, we provide convergence g
uarantees for SHB to a neighborhood of the solution for convex and smooth proble
ms (without assuming interpolation). If interpolation is also satisfied, then us
ing MomSPSmax, SHB converges to the true solution at a fast rate matching the de
terministic HB. The other two variants, MomDecSPS and MomAdaSPS, are the first a
daptive step-size for SHB that guarantee convergence to the exact minimizer - wi
thout a priori knowledge of the problem parameters and without assuming interpol
ation. Our convergence analysis of SHB is tight and obtains the convergence guar
antees of stochastic Polyak step-size for SGD as a special case. We supplement o
ur analysis with experiments validating our theory and demonstrating the effecti
veness and robustness of our algorithms.
****

Xiongye Xiao,Heng Ping,Chenyu Zhou,Defu Cao,Yaxing Li,Yi-Zhuo Zhou,Shixuan Li,Ni
kos Kanakaris,Paul Bogdan

Neuron-based Multifractal Analysis of Neuron Interaction Dynamics in Large Model
s

In recent years, there has been increasing attention on the capabilities of larg
e-scale models, particularly in handling complex tasks that small-scale models a
re unable to perform. Notably, large language models (LLMs) have demonstrated ``
intelligent'' abilities such as complex reasoning and abstract language comprehe
nsion, reflecting cognitive-like behaviors. However, current research on emergen
t abilities in large models predominantly focuses on the relationship between mo
del performance and size, leaving a significant gap in the systematic quantitati
ve analysis of the internal structures and mechanisms driving these emergent abi
lities. Drawing inspiration from neuroscience research on brain network structur
e and self-organization, we propose (i) a general network representation of larg
e models, (ii) a new analytical framework — *Neuron-based Multifractal Analysis
(NeuroMFA)* - for structural analysis, and (iii) a novel structure-based metric
as a proxy for emergent abilities of large models. By linking structural feature
s to the capabilities of large models, *NeuroMFA* provides a quantitative framew
ork for analyzing emergent phenomena in large models. Our experiments show that
the proposed method yields a comprehensive measure of the network's evolving het
erogeneity and organization, offering theoretical foundations and a new perspect
ive for investigating emergence in large models.
****

Shurong Wang,Zhuoyang Shen,Xinbao Qiao,Tongning Zhang,Meng Zhang
DynFrs: An Efficient Framework for Machine Unlearning in Random Forest
Random Forests are widely recognized for establishing efficacy in classification and regression tasks, standing out in various domains such as medical diagnosis, finance, and personalized recommendations. These domains, however, are inherently sensitive to privacy concerns, as personal and confidential data are involved. With increasing demand for the right to be forgotten, particularly under regulations such as GDPR and CCPA, the ability to perform machine unlearning has become crucial for Random Forests. However, insufficient attention was paid to this topic, and existing approaches face difficulties in being applied to real-world scenarios. Addressing this gap, we propose the DynFrs framework designed to enable efficient machine unlearning in Random Forests while preserving predictive accuracy. Dynfrs leverages subsampling method Occ(q) and a lazy tag strategy Lzy, and is still adaptable to any Random Forest variant. In essence, Occ(q) ensures that each sample in the training set occurs only in a proportion of trees so that the impact of deleting samples is limited, and Lzy delays the reconstruction of a tree node until necessary, thereby avoiding unnecessary modifications on tree structures. In experiments, applying Dynfrs on Extremely Randomized Trees yields substantial improvements, achieving orders of magnitude faster unlearning performance and better predictive accuracy than existing machine unlearning methods for Random Forests.
****
Yanhong Li,Karen Livescu,Jiawei Zhou
Chunk-Distilled Language Modeling
We introduce Chunk-Distilled Language Modeling (CD-LM), an approach to text generation that addresses two challenges in current large language models (LLMs): the inefficiency of token-level generation, and the difficulty of adapting to new data and knowledge. Our method combines deep network-based LLMs with a straightforward retrieval module, which allows the generation of multi-token text chunks at a single decoding step. Our retrieval framework enables flexible construction of model- or domain-specific datastores, either leveraging the internal knowledge of existing models, or incorporating expert insights from human-annotated corpora. This adaptability allows for enhanced control over the language model's distribution without necessitating additional training. We present the CD-LM formulation along with performance metrics demonstrating its ability to improve language model performance and efficiency across a diverse set of downstream applications. Code and data will be made publicly available.
****
Zijian Guo,Weichao Zhou,Shengao Wang,Wenchao Li
Constraint-Conditioned Actor-Critic for Offline Safe Reinforcement Learning
Offline safe reinforcement learning (OSRL) aims to learn policies with high rewards while satisfying safety constraints solely from data collected offline. However, the learned policies often struggle to handle states and actions that are not present or out-of-distribution (OOD) from the offline dataset, which can result in violation of the safety constraints or overly conservative behaviors during their online deployment. Moreover, many existing methods are unable to learn policies that can adapt to varying constraint thresholds. To address these challenges, we propose constraint-conditioned actor-critic (CCAC), a novel OSRL method that models the relationship between state-action distributions and safety constraints, and leverages this relationship to regularize critics and policy learning. CCAC learns policies that can effectively handle OOD data and adapt to varying constraint thresholds. Empirical evaluations on the $\texttt{DSRL}$ benchmarks show that CCAC significantly outperforms existing methods for learning adaptive, safe, and high-reward policies.
****
Jiyang Zheng,Jialiang Shen,Yu Yao,Min Wang,Yang Yang,Dadong Wang,Tongliang Liu
Chain-of-Focus Prompting: Leveraging Sequential Visual Cues to Prompt Large Autoregressive Vision Models
In-context learning (ICL) has revolutionized natural language processing by enabling models to adapt to diverse tasks with only a few illustrative examples. How

ever, the exploration of ICL within the field of computer vision remains limited. Inspired by Chain-of-Thought (CoT) prompting in the language domain, we propose Chain-of-Focus (CoF) Prompting, which enhances vision models by enabling step-by-step visual comprehension. CoF Prompting addresses the challenges of absent logical structure in visual data by generating intermediate reasoning steps through visual saliency. Moreover, it provides a solution for creating tailored prompts from visual inputs by selecting contextually informative prompts based on query similarity and target richness. The significance of CoF prompting is demonstrated by the recent introduction of Large Autoregressive Vision Models (LAVMs), which predict downstream targets via in-context learning with pure visual inputs. By integrating intermediate reasoning steps into visual prompts and effectively selecting the informative ones, the LAVMs are capable of generating significantly better inferences. Extensive experiments on downstream visual understanding tasks validate the effectiveness of our proposed method for visual in-context learning.
****

Rongzhen Zhao,Vivienne Huiling Wang,Juho Kannala,Joni Pajarinen
Multi-Scale Fusion for Object Representation
Representing images or videos as object-level feature vectors, rather than pixel-level feature maps, facilitates advanced visual tasks.
Object-Centric Learning (OCL) primarily achieves this by reconstructing the input under the guidance of Variational Autoencoder (VAE) intermediate representation to drive so-called slots to aggregate as much object information as possible.
However, existing VAE guidance does not explicitly address that objects can vary in pixel sizes while models typically excel at specific pattern scales.
We propose Multi-Scale Fusion (MSF) to enhance VAE guidance for OCL training.
To ensure objects of all sizes fall within VAE's comfort zone, we adopt the image pyramid, which produces intermediate representations at multiple scales;
To foster scale-invariance/variance in object super-pixels, we devise inter/intra-scale fusion, which augments low-quality object super-pixels of one scale with corresponding high-quality super-pixels from another scale.
On standard OCL benchmarks, our technique improves mainstream methods, including state-of-the-art diffusion-based ones.
The source code is available on https://github.com/Genera1Z/MultiScaleFusion.
****

Cheng Tan,Zhenxiao Cao,Zhangyang Gao,Lirong Wu,Siyuan Li,Yufei Huang,Jun Xia,Bozhen Hu,Stan Z. Li
MeToken: Uniform Micro-environment Token Boosts Post-Translational Modification Prediction
Post-translational modifications (PTMs) profoundly expand the complexity and functionality of the proteome, regulating protein attributes and interactions that are crucial for biological processes. Accurately predicting PTM sites and their specific types is therefore essential for elucidating protein function and understanding disease mechanisms. Existing computational approaches predominantly focus on protein sequences to predict PTM sites, driven by the recognition of sequence-dependent motifs. However, these approaches often overlook protein structural contexts. In this work, we first compile a large-scale sequence-structure PTM dataset, which serves as the foundation for fair comparison. We introduce the MeToken model, which tokenizes the micro-environment of each amino acid, integrating both sequence and structural information into unified discrete tokens. This model not only captures the typical sequence motifs associated with PTMs but also leverages the spatial arrangements dictated by protein tertiary structures, thus providing a holistic view of the factors influencing PTM sites. Designed to address the long-tail distribution of PTM types, MeToken employs uniform sub-codebooks that ensure even the rarest PTMs are adequately represented and distinguished. We validate the effectiveness and generalizability of MeToken across multiple datasets, demonstrating its superior performance in accurately identifying PTM types. The results underscore the importance of incorporating structural data and highlight MeToken's potential in facilitating accurate and comprehensive PTM predictions, which could significantly impact proteomics research.

****
Jiawei Wang,Shaofei Lu,Da Cao,Dongyu Wang,Yuquan Le,Zhe Quan,Tat-Seng Chua
Neural Causal Graph for Interpretable and Intervenable Classification
Advancements in neural networks have significantly enhanced the performance of c lassification models, achieving remarkable accuracy across diverse datasets. How ever, these models often lack transparency and do not support interactive reason ing with human users, which are essential attributes for applications that requi re trust and user engagement. To overcome these limitations, we introduce an inn ovative framework, Neural Causal Graph (NCG), that integrates causal inference w ith neural networks to enable interpretable and intervenable reasoning. We then propose an intervention training method to model the intervention probability of the prediction, serving as a contextual prompt to facilitate the fine-grained r easoning and human-AI interaction abilities of NCG. Our experiments show that th e proposed framework significantly enhances the performance of traditional class ification baselines. Furthermore, NCG achieves nearly 95\% top-1 accuracy on the ImageNet dataset by employing a test-time intervention method. This framework n ot only supports sophisticated post-hoc interpretation but also enables dynamic human-AI interactions, significantly improving the model's transparency and appl icability in real-world scenarios.
****
Amandine Brunetto,Sascha Hornauer,Fabien Moutarde
NeRAF: 3D Scene Infused Neural Radiance and Acoustic Fields
Sound plays a major role in human perception. Along with vision, it provides ess ential information for understanding our surroundings. Despite advances in neura l implicit representations, learning acoustics that align with visual scenes rem ains a challenge. We propose NeRAF, a method that jointly learns acoustic and ra diance fields. NeRAF synthesizes both novel views and spatialized room impulse r esponses (RIR) at new positions by conditioning the acoustic field on 3D scene g eometric and appearance priors from the radiance field. The generated RIR can be applied to auralize any audio signal. Each modality can be rendered independent ly and at spatially distinct positions, offering greater versatility. We demonst rate that NeRAF generates high-quality audio on SoundSpaces and RAF datasets, ac hieving significant performance improvements over prior methods while being more data-efficient. Additionally, NeRAF enhances novel view synthesis of complex sc enes trained with sparse data through cross-modal learning.
NeRAF is designed as a Nerfstudio module, providing convenient access to realist ic audio-visual generation. Project page: https://amandinebtto.github.io/NeRAF
****
Berken Utku Demirel,Christian Holz
Shifting the Paradigm: A Diffeomorphism Between Time Series Data Manifolds for A chieving Shift-Invariancy in Deep Learning
Deep learning models lack shift invariance, making them sensitive to input shift s that cause changes in output. While recent techniques seek to address this for images, our findings show that these approaches fail to provide shift-invarianc e in time series, where the data generation mechanism is more challenging due to the interaction of low and high frequencies. Worse, they also decrease performa nce across several tasks. In this paper, we propose a novel differentiable bijec tive function that maps samples from their high-dimensional data manifold to ano ther manifold of the same dimension, without any dimensional reduction. Our appr oach guarantees that samples---when subjected to random shifts---are mapped to a unique point in the manifold while preserving all task-relevant information wit hout loss. We theoretically and empirically demonstrate that the proposed transf ormation guarantees shift-invariance in deep learning models without imposing an y limits to the shift. Our experiments on six time series tasks with state-of-th e-art methods show that our approach consistently improves the performance while enabling models to achieve complete shift-invariance without modifying or impos ing restrictions on the model's topology. The source code is available on GitHub .
****
Jindou Chen,Jidong Tian,Liang Wu,ChenXinWei,Xiaokang Yang,Yaohui Jin,Yanyan Xu

KinFormer: Generalizable Dynamical Symbolic Regression for Catalytic Organic Reaction Kinetics
Modeling kinetic equations is essential for understanding the mechanisms of chemical reactions, yet a complex and time-consuming task. Kinetic equation prediction is formulated as a problem of dynamical symbolic regression (DSR) subject to physical chemistry constraints. Deep learning (DL) holds the potential to capture reaction patterns and predict kinetic equations from data of chemical species, effectively avoiding empirical bias and improving efficiency compared with traditional analytical methods. Despite numerous studies focusing on DSR and the introduction of Transformers to predict ordinary differential equations, the corresponding models lack generalization abilities across diverse categories of reactions. In this study, we propose KinFormer, a generalizable kinetic equation prediction model. KinFormer utilizes a conditional Transformer to model DSR under physical constraints and employs Monte Carlo Tree Search to apply the model to new types of reactions. Experimental results on 20 types of organic reactions demonstrate that KinFormer not only outperforms classical baselines, but also exceeds Transformer baselines in out-of-domain evaluations, thereby proving its generalization ability.
****
Andrey Gromov,Kushal Tirumala,Hassan Shapourian,Paolo Glorioso,Dan Roberts
The Unreasonable Ineffectiveness of the Deeper Layers
How is knowledge stored in an LLM’s weights? We study this via layer pruning: if removing a certain layer does not affect model performance in common question-answering benchmarks, then the weights in that layer are not necessary for storing the knowledge needed to answer those questions. To find these unnecessary parameters, we identify the optimal block of layers to prune by considering similarity across layers; then, to “heal” the damage, we perform a small amount of finetuning. Surprisingly, with this method we find minimal degradation of performance until after a large fraction (up to half) of the layers are removed for some common open-weight models. From a scientific perspective, the robustness of these LLMs to the deletion of layers implies either that current pretraining methods are not properly leveraging the parameters in the deeper layers of the network or that the shallow layers play a critical role in storing knowledge. For our study, we use parameter-efficient finetuning (PEFT) methods, specifically quantization and Low Rank Adapters (QLoRA), such that each of our experiments can be performed on a single 40GB A100 GPU.
****
Andrew Estornell,Jean-Francois Ton,Yuanshun Yao,Yang Liu
ACC-Collab: An Actor-Critic Approach to Multi-Agent LLM Collaboration
Large language models (LLMs) have demonstrated a remarkable ability to serve as general-purpose tools for various language-based tasks.
Recent works have demonstrated that the efficacy of such models can be improved through iterative dialog between multiple models.
While these
paradigms show promise
in
improving model efficacy, most works in this area treat collaboration as an emergent behavior, rather than a learned behavior.
In doing so, current multi-agent frameworks rely on collaborative behaviors to have been sufficiently trained into off-the-shelf models.
To address this limitation, we propose ACC-Collab, an **A**ctor-**C**riti**c** based learning framework to produce a two-agent team (an actor-agent and a critic-agent) specialized in collaboration.
We demonstrate that ACC-Collab outperforms SotA multi-agent techniques on a wide array of benchmarks.
****
Guangyan Sun,Mingyu Jin,Zhenting Wang,Cheng-Long Wang,Siqi Ma,Qifan Wang,Tong Geng,Ying Nian Wu,Yongfeng Zhang,Dongfang Liu
Visual Agents as Fast and Slow Thinkers
Achieving human-level intelligence requires refining cognitive distinctions betw

een \textit{System 1} and \textit{System 2} thinking. While contemporary AI, dri
ven by large language models, demonstrates human-like traits, it falls short of
genuine cognition. Transitioning from structured benchmarks to real-world scenar
ios presents challenges for visual agents, often leading to inaccurate and overl
y confident responses. To address the challenge, we introduce \textbf{\textsc{Fa
ST}}, which incorporates the \textbf{Fa}st and \textbf{S}low \textbf{T}hinking m
echanism into visual agents. \textsc{FaST} employs a switch adapter to dynamical
ly select between \textit{System 1/2} modes, tailoring the problem-solving appro
ach to different task complexity. It tackles uncertain and unseen objects by adj
usting model confidence and integrating new contextual data. With this novel des
ign, we advocate a \textit{flexible system}, \textit{hierarchical reasoning} cap
abilities, and a \textit{transparent decision-making} pipeline, all of which con
tribute to its ability to emulate human-like cognitive processes in visual intel
ligence. Empirical results demonstrate that \textsc{FaST} outperforms various we
ll-known baselines, achieving 80.8\% accuracy over VQA^{v2} for visual questio
n answering and 48.7\% $GIoU$ score over ReasonSeg for reasoning segmentation, d
emonstrate \textsc{FaST}'s superior performance. Extensive testing validates the
 efficacy and robustness of \textsc{FaST}'s core components, showcasing its pote
ntial to advance the development of cognitive visual agents in AI systems.
****

Zhongyi Shui,Jianpeng Zhang,Weiwei Cao,Sinuo Wang,Ruizhe Guo,Le Lu,Lin Yang,Xian
ghua Ye,Tingbo Liang,Qi Zhang,Ling Zhang

Large-scale and Fine-grained Vision-language Pre-training for Enhanced CT Image
Understanding

Artificial intelligence (AI) shows great potential in assisting radiologists to
improve the efficiency and accuracy of medical image interpretation and diagnosi
s. However, a versatile AI model requires large-scale data and comprehensive ann
otations, which are often impractical in medical settings. Recent studies levera
ge radiology reports as a naturally high-quality supervision for medical images,
 using contrastive language-image pre-training (CLIP) to develop language-inform
ed models for radiological image interpretation. Nonetheless, these approaches t
ypically contrast entire images with reports, neglecting the local associations
between imaging regions and report sentences, which may undermine model performa
nce and interoperability. In this paper, we propose a fine-grained vision-langua
ge model (fVLM) for anatomy-level CT image interpretation. Specifically, we expl
icitly match anatomical regions of CT images with corresponding descriptions in
radiology reports and perform contrastive pre-training for each anatomy individu
ally. Fine-grained alignment, however, faces considerable false-negative challen
ges, mainly from the abundance of anatomy-level healthy samples and similarly di
seased abnormalities, leading to ambiguous patient-level pairings. To tackle thi
s issue, we propose identifying false negatives of both normal and abnormal samp
les and calibrating contrastive learning from patient-level to disease-aware pai
ring. We curated the largest CT dataset to date, comprising imaging and report d
ata from 69,086 patients, and conducted a comprehensive evaluation of 54 major a
nd important disease (including several most deadly cancers) diagnosis tasks acr
oss 15 main anatomies. Experimental results demonstrate the substantial potentia
l of fVLM in versatile medical image interpretation. In the zero-shot classifica
tion task, we achieved an average AUC of 81.3% on 54 diagnosis tasks, surpassing
 CLIP and supervised methods by 12.9% and 8.0%, respectively. Additionally, on t
he publicly available CT-RATE and Rad-ChestCT benchmarks, our fVLM outperformed
the current state-of-the-art methods with absolute AUC gains of 7.4% and 4.8%, r
espectively.
****

Ya-Wei Eileen Lin,Ronald R. Coifman,Gal Mishne,Ronen Talmon

Tree-Wasserstein Distance for High Dimensional Data with a Latent Feature Hierar
chy

Finding meaningful distances between high-dimensional data samples is an importa
nt scientific task. To this end, we propose a new tree-Wasserstein distance (TWD
) for high-dimensional data with two key aspects. First, our TWD is specifically
 designed for data with a latent feature hierarchy, i.e., the features lie in a

hierarchical space, in contrast to the usual focus on embedding samples in hyper
bolic space. Second, while the conventional use of TWD is to speed up the comput
ation of the Wasserstein distance, we use its inherent tree as a means to learn
the latent feature hierarchy. The key idea of our method is to embed the feature
s into a multi-scale hyperbolic space using diffusion geometry and then present
a new tree decoding method by establishing analogies between the hyperbolic embe
dding and trees. We show that our TWD computed based on data observations provab
ly recovers the TWD defined with the latent feature hierarchy and that its compu
tation is efficient and scalable. We showcase the usefulness of the proposed TWD
 in applications to word-document and single-cell RNA-sequencing datasets, demon
strating its advantages over existing TWDs and methods based on pre-trained mode
ls.
****
Yifan Niu,Ziqi Gao,Tingyang Xu,Yang Liu,Yatao Bian,Yu Rong,Junzhou Huang,Jia Li
InversionGNN: A Dual Path Network for Multi-Property Molecular Optimization
Exploring chemical space to find novel molecules that simultaneously satisfy mul
tiple properties is crucial in drug discovery. However, existing methods often s
truggle with trading off multiple properties due to the conflicting or correlate
d nature of chemical properties. To tackle this issue, we introduce InversionGN
N framework, an effective yet sample-efficient dual-path graph neural network (G
NN) for multi-objective drug discovery. In the direct prediction path of Invers
ionGNN, we train the model for multi-property prediction to acquire knowledge of
 the optimal combination of functional groups.
Then the learned chemical knowledge helps the inversion generation path to gener
ate molecules with required properties.
In order to decode the complex knowledge of multiple properties in the inversion
 path, we propose a gradient-based Pareto search method to balance conflicting p
roperties and generate Pareto optimal molecules.
Additionally, InversionGNN is able to search the full Pareto front approximately
 in discrete chemical space. Comprehensive experimental evaluations show that In
versionGNN is both effective and sample-efficient in various discrete multi-obje
ctive settings including drug discovery.
****
Yushun Dong,Patrick Soga,Yinhan He,Song Wang,Jundong Li
Graph Neural Networks Are More Than Filters: Revisiting and Benchmarking from A
Spectral Perspective
Graph Neural Networks (GNNs) have achieved remarkable success in various graph-b
ased learning tasks. While their performance is often attributed to the powerful
 neighborhood aggregation mechanism, recent studies suggest that other component
s such as non-linear layers may also significantly affecting how GNNs process th
e input graph data in the spectral domain. Such evidence challenges the prevalen
t opinion that neighborhood aggregation mechanisms dominate the behavioral chara
cteristics of GNNs in the spectral domain. To demystify such a conflict, this pa
per introduces a comprehensive benchmark to measure and evaluate GNNs' capabilit
y in capturing and leveraging the information encoded in different frequency com
ponents of the input graph data. Specifically, we first conduct an exploratory s
tudy demonstrating that GNNs can flexibly yield outputs with diverse frequency c
omponents even when certain frequencies are absent or filtered out from the inpu
t graph data. We then formulate a novel research problem of measuring and benchm
arking the performance of GNNs from a spectral perspective. To take an initial s
tep towards a comprehensive benchmark, we design an evaluation protocol supporte
d by comprehensive theoretical analysis. Finally, we introduce a comprehensive b
enchmark on real-world datasets, revealing insights that challenge prevalent opi
nions from a spectral perspective. We believe that our findings will open new av
enues for future advancements in this area. Our implementations can be found at:
 https://github.com/yushundong/Spectral-benchmark.
****
Yuchen Liang,Peizhong Ju,Yingbin Liang,Ness Shroff
Theory on Score-Mismatched Diffusion Models and Zero-Shot Conditional Samplers
The denoising diffusion model has recently emerged as a powerful generative tech

nique, capable of transforming noise into meaningful data. While theoretical con
vergence guarantees for diffusion models are well established when the target di
stribution aligns with the training distribution, practical scenarios often pres
ent mismatches. One common case is in the zero-shot conditional diffusion sampli
ng, where the target conditional distribution is different from the (uncondition
al) training distribution. These score-mismatched diffusion models remain largel
y unexplored from a theoretical perspective. In this paper, we present the first
 performance guarantee with explicit dimensional dependencies for general score-
mismatched diffusion samplers, focusing on target distributions with finite seco
nd moments. We show that score mismatches result in an asymptotic distributional
 bias between the target and sampling distributions, proportional to the accumul
ated mismatch between the target and training distributions. This result can be
directly applied to zero-shot conditional samplers for any conditional model, ir
respective of measurement noise. Interestingly, the derived convergence upper bo
und offers useful guidance for designing a novel bias-optimal zero-shot sampler
in linear conditional models that minimizes the asymptotic bias. For such bias-o
ptimal samplers, we further establish convergence guarantees with explicit depen
dencies on dimension and conditioning, applied to several interesting target dis
tributions, including those with bounded support and Gaussian mixtures. Our find
ings are supported by numerical studies.
****

Viet Dung Nguyen,Minh Nguyen Hoang,Luc Nguyen,Rachel Teo,Tan Minh Nguyen,Linh Du
y Tran

CAMEx: Curvature-aware Merging of Experts

Existing methods for merging experts during model training and fine-tuning predo
minantly rely on Euclidean geometry, which assumes a flat parameter space. This
assumption can limit the model's generalization ability, especially during the p
re-training phase, where the parameter manifold might exhibit more complex curva
ture. Curvature-aware merging methods typically require additional information a
nd computational resources to approximate the Fisher Information Matrix, adding
memory overhead. In this paper, we introduce CAMEx (Curvature-Aware Merging of E
xperts), a novel expert merging protocol that incorporates natural gradients to
account for the non-Euclidean curvature of the parameter manifold. By leveraging
 natural gradients, CAMEx adapts more effectively to the structure of the parame
ter space, improving alignment between model updates and the manifold's geometry
. This approach enhances both pre-training and fine-tuning, resulting in better
optimization trajectories and improved generalization without the substantial me
mory overhead typically associated with curvature-aware methods. Our contributio
ns are threefold: (1) CAMEx significantly outperforms traditional Euclidean-base
d expert merging techniques across various natural language processing tasks, le
ading to enhanced performance during pre-training and fine-tuning; (2) we introd
uce a dynamic merging architecture that optimizes resource utilization, achievin
g high performance while reducing computational costs, facilitating efficient sc
aling of large language models; and (3) we provide both theoretical and empirica
l evidence to demonstrate the efficiency of our proposed method. The code is pub
licly available at: https://github.com/kpup1710/CAMEx.
****

Sujoy Bhore,Devdan Dey,Satyam Singh

ONLINE EPSILON NET & PIERCING SET FOR GEOMETRIC CONCEPTS

VC-dimension (Vapnik & Chervonenkis (1971)) and ε-nets (Haussler &
Welzl (1987)) are key concepts in Statistical Learning Theory. Intuitively, VC-d
imension is a measure of the size of a class of sets. The famous ε-n
et theorem, a fundamental result in Discrete Geometry, asserts that if the VC-di
mension of a set system is bounded, then a small sample exists that intersects a
ll sufficiently large sets.

In online learning scenarios where data arrives sequentially, the VC-dimensi
on helps to bound the complexity of the set system, and ε-nets ensur
e the selection of a small representative set. This sampling framework is crucia
l in various domains, including spatial data analysis, motion planning in dynami

c environments, optimization of sensor networks, and feature extraction in compu
ter vision, among others. Motivated by these applications, we study the online $
\varepsilon$-net problem for geometric concepts with bounded VC-dimension. While
 the offline version of this problem has been extensively studied, surprisingly,
 there are no known theoretical results for the online version to date. We prese
nt the first deterministic online algorithm with an optimal competitive ratio fo
r intervals in $\mathbb{R}$. Next, we give a randomized online algorithm with a
near-optimal competitive ratio for axis-aligned boxes in $\mathbb{R}^d$, for $d\
le 3$. Furthermore, we introduce a novel technique to analyze similar-sized obje
cts of constant description complexity in $\mathbb{R}^d$, which may be of indepe
ndent interest.

Next, we focus on the continuous version of this problem (called online pier
cing set), where ranges of the set system are geometric concepts in $\mathbb{R}^
d$ arriving in an online manner, but the universe is the entire ambient space, a
nd the objective is to choose a small sample that intersects all the ranges. Alt
hough online piercing set is a very well-studied problem in the literature, to o
ur surprise, very few works have addressed generic geometric concepts without an
y assumption about the sizes. We advance this field by proposing asymptotically
optimal competitive deterministic algorithms for boxes and ellipsoids in $\mathb
b{R}^d$, for any $d\in\mathbb{N}$.
****
Serin Yang,Taesung Kwon,Jong Chul Ye
ViBiDSampler: Enhancing Video Interpolation Using Bidirectional Diffusion Sample
r
Recent progress in large-scale text-to-video (T2V) and image-to-video (I2V) diff
usion models has greatly enhanced video generation, especially in terms of keyfr
ame interpolation. However, current image-to-video diffusion models, while power
ful in generating videos from a single conditioning frame, need adaptation for t
wo-frame (start \& end) conditioned generation, which is essential for effective
 bounded interpolation. Unfortunately, existing approaches that fuse temporally
forward and backward paths in parallel often suffer from off-manifold issues, le
ading to artifacts or requiring multiple iterative re-noising steps. In this wor
k, we introduce a novel, bidirectional sampling strategy to address these off-ma
nifold issues without requiring extensive re-noising or fine-tuning. Our method
employs sequential sampling along both forward and backward paths, conditioned o
n the start and end frames, respectively, ensuring more coherent and on-manifold
 generation of intermediate frames. Additionally, we incorporate advanced guidan
ce techniques, CFG++ and DDS, to further enhance the interpolation process. By i
ntegrating these, our method achieves state-of-the-art performance, efficiently
generating high-quality, smooth videos between keyframes. On a single 3090 GPU,
our method can interpolate 25 frames at 1024$\times$576 resolution in just 195 s
econds, establishing it as a leading solution for keyframe interpolation.
Project page: https://vibidsampler.github.io/
****
Ally Yalei Du,Lin Yang,Ruosong Wang
Misspecified Q-Learning with Sparse Linear Function Approximation: Tight Boun
ds on Approximation Error
The recent work by Dong and Yang (2023) showed for misspecified sparse linear ba
ndits, one can obtain an $O(\epsilon)$-optimal policy using a polynomial number
of samples when the sparsity is a constant, where ϵ is the misspecifica
tion error. This result is in sharp contrast to misspecified linear bandits with
out sparsity, which require an exponential number of samples to get the same gua
rantee. In order to study whether the analog result is possible in the reinforce
ment learning setting, we consider the following problem: assuming the optimal $
Q$-function is a d-dimensional linear function with sparsity k and misspecif
ication error ϵ, whether we can obtain an $O(\epsilon)$-optimal policy
using number of samples polynomially in the feature dimension d. We first demo
nstrate why the standard approach based on Bellman backup or the existing optimi
stic value function elimination approach such as OLIVE (Jiang et al., 2017) achi

eves suboptimal guarantees for this problem. We then design a novel elimination-based algorithm to show one can obtain an
$O(H\epsilon)$-optimal policy with sample complexity polynomially in the feature dimension d and planning horizon H. Lastly, we complement our upper bound with an $\tilde \Omega(H\epsilon)$ suboptimality lower bound, giving a complete picture of this problem.
****
Yuchen Duan,Weiyun Wang,Zhe Chen,Xizhou Zhu,Lewei Lu,Tong Lu,Yu Qiao,Hongsheng Li,Jifeng Dai,Wenhai Wang
Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures
Transformers have revolutionized computer vision and natural language processing, but their high computational complexity limits their application in high-resolution image processing and long-context analysis. This paper introduces Vision-RWKV (VRWKV), a model that builds upon the RWKV architecture from the NLP field with key modifications tailored specifically for vision tasks. Similar to the Vision Transformer (ViT), our model demonstrates robust global processing capabilities, efficiently handles sparse inputs like masked images, and can scale up to accommodate both large-scale parameters and extensive datasets. Its distinctive advantage is its reduced spatial aggregation complexity, enabling seamless processing of high-resolution images without the need for window operations. Our evaluations demonstrate that VRWKV surpasses ViT's performance in image classification and has significantly faster speeds and lower memory usage processing high-resolution inputs. In dense prediction tasks, it outperforms window-based models, maintaining comparable speeds. These results highlight VRWKV's potential as a more efficient alternative for visual perception tasks. Code and models are available at~\url{https://github.com/OpenGVLab/Vision-RWKV}.
****
Jianglin Lu,Yixuan Liu,Yitian Zhang,Yun Fu
Scale-Free Graph-Language Models
Graph-language models (GLMs) have demonstrated great potential in graph-based semi-supervised learning. A typical GLM consists of two key stages: graph generation and text embedding, which are usually implemented by inferring a latent graph and finetuning a language model (LM), respectively. However, the former often relies on artificial assumptions about the underlying edge distribution, while the latter requires extensive data annotations. To tackle these challenges, this paper introduces a novel GLM that integrates graph generation and text embedding within a unified framework. Specifically, for graph generation, we leverage an inherent characteristic of real edge distribution—the scale-free property—as a structural prior. We unexpectedly find that this natural property can be effectively approximated by a simple k-nearest neighbor (KNN) graph. For text embedding, we develop a graph-based pseudo-labeler that utilizes scale-free graphs to provide complementary supervision for improved LM finetuning. Extensive experiments on representative datasets validate our findings on the scale-free structural approximation of KNN graphs and demonstrate the effectiveness of integrating graph generation and text embedding with a real structural prior. Our code is available at https://github.com/Jianglin954/SFGL.
****
Garrett Tanzer,Biao Zhang
YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus
Even for better-studied sign languages like American Sign Language (ASL), data is the bottleneck for machine learning research. The situation is worse yet for the many other sign languages used by Deaf/Hard of Hearing communities around the world. In this paper, we present YouTube-SL-25, a large-scale, open-domain multilingual corpus of sign language videos with seemingly well-aligned captions drawn from YouTube. With >3000 hours of videos across >25 sign languages, YouTube-SL-25 is a) >3x the size of YouTube-ASL, b) the largest parallel sign language dataset to date, and c) the first or largest parallel dataset for many of its component languages. We provide baselines for sign-to-text tasks using a unified mul

tilingual multitask model based on T5 and report scores on benchmarks across 4 s
ign languages. The results demonstrate that multilingual transfer benefits both
higher- and lower-resource sign languages within YouTube-SL-25.
****

Fu-Yun Wang,Ling Yang,Zhaoyang Huang,Mengdi Wang,Hongsheng Li
Rectified Diffusion: Straightness Is Not Your Need in Rectified Flow
Diffusion models have greatly improved visual generation but are hindered by slo
w generation speed due to the computationally intensive nature of solving genera
tive ODEs. Rectified flow, a widely recognized solution, improves generation spe
ed by straightening the ODE path. Its
key components include: 1) using the linear interpolating diffusion form of flow
-matching, 2) employing $\boldsymbol v$-prediction, and 3) performing rectificat
ion (a.k.a. reflow). In this paper, we argue that the success of rectification p
rimarily lies in using a pretrained diffusion model to obtain matched pairs of n
oise and samples, followed by retraining with these matched noise-sample pairs.
Based on this, components 1) and 2) are unnecessary. Furthermore, we highlight t
hat straightness is not an essential training target for rectification; rather,
it is a specific case of flow-matching models. The more critical training target
 is to achieve a first-order approximate ODE path, which is inherently curved fo
r models like DDPM and Sub-VP. Building on this insight, we propose Rectified Di
ffusion, which generalizes the design space and application scope of rectificati
on to encompass the broader category of diffusion models, rather than being rest
ricted to flow-matching models. We validate our method on Stable Diffusion v1-5
and Stable Diffusion XL. Our method not only greatly simplifies the training pro
cedure of rectified flow-based previous works (e.g., InstaFlow) but also achieve
s superior performance with even lower training cost.
Our code is available at https://github.com/G-U-N/Rectified-Diffusion.
****

Zihao Zhou,Shudong Liu,Maizhen Ning,Wei Liu,Jindong Wang,Derek F. Wong,Xiaowei H
uang,Qiufeng Wang,Kaizhu Huang
Is Your Model Really A Good Math Reasoner? Evaluating Mathematical Reasoning wit
h Checklist
Exceptional mathematical reasoning ability is one of the key features that demon
strate the power of large language models (LLMs). How to comprehensively define
and evaluate the mathematical abilities of LLMs, and even reflect the user exper
ience in real-world scenarios, has emerged as a critical issue. Current benchmar
ks predominantly concentrate on problem-solving capabilities, presenting a subst
antial risk of model overfitting and fails to accurately measure the genuine mat
hematical reasoning abilities. In this paper, we argue that if a model really un
derstands a problem, it should be robustly and readily applied across a diverse
array of tasks. To this end, we introduce MathCheck, a well-designed checklist f
or testing task generalization and reasoning robustness, as well as an automatic
 tool to generate checklists efficiently. MathCheck includes multiple mathematic
al reasoning tasks and robustness tests to facilitate a comprehensive evaluation
 of both mathematical reasoning ability and behavior testing. Utilizing MathChec
k, we develop MathCheck-GSM and MathCheck-GEO to assess mathematical textual rea
soning and multi-modal reasoning capabilities, respectively, serving as upgraded
 versions of benchmarks including GSM8k, GeoQA, UniGeo, and Geometry3K. We adopt
 MathCheck-GSM and MathCheck-GEO to evaluate over 26 LLMs and 17 multi-modal LLM
s, assessing their comprehensive mathematical reasoning abilities. Our results d
emonstrate that while frontier LLMs like GPT-4o continue to excel in various abi
lities on the checklist, many other model families exhibit a significant decline
. Further experiments indicate that, compared to traditional math benchmarks, Ma
thCheck better reflects true mathematical abilities and represents mathematical
intelligence more linearly, thereby supporting our design. Using MathCheck, we c
an also efficiently conduct informative behavior analysis to deeply investigate
models. Finally, we show that our proposed checklist paradigm can easily extend
to other reasoning tasks for their comprehensive evaluation.
****

Moritz Reuss,Jyothish Pari,Pulkit Agrawal,Rudolf Lioutikov

Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning
Diffusion Policies have become widely used in Imitation Learning, offering several appealing properties, such as generating multimodal and discontinuous behavior.
As models are becoming larger to capture more complex capabilities, their computational demands increase, as shown by recent scaling laws.
Therefore, continuing with the current architectures will present a computational roadblock.
To address this gap, we propose Mixture-of-Denoising Experts (MoDE) as a novel policy for Imitation Learning.
MoDE surpasses current state-of-the-art Transformer-based Diffusion Policies while enabling parameter-efficient scaling through sparse experts and noise-conditioned routing, reducing both active parameters by 40\% and inference costs by 90\% via expert caching.
Our architecture combines this efficient scaling with noise-conditioned self-attention mechanism, enabling more effective denoising across different noise levels.
MoDE achieves state-of-the-art performance on 134 tasks in four established imitation learning benchmarks (CALVIN and LIBERO).
Notably, by pretraining MoDE on diverse robotics data, we achieve 4.01 on CALVIN ABC and 0.95 on LIBERO-90.
It surpasses both CNN-based and Transformer Diffusion Policies by an average of 57% across 4 benchmarks, while using 90\% fewer FLOPs and fewer active parameters compared to default Diffusion Transformer architectures. Furthermore, we conduct comprehensive ablations on MoDE's components, providing insights for designing efficient and scalable Transformer architectures for Diffusion Policies.
Code and demonstrations are available at https://mbreuss.github.io/MoDE_Diffusion_Policy.
****
Hongye Cao,Fan Feng,Tianpei Yang,Jing Huo,Yang Gao
Causal Information Prioritization for Efficient Reinforcement Learning
Current Reinforcement Learning (RL) methods often suffer from sample-inefficiency, resulting from blind exploration strategies that neglect causal relationships among states, actions, and rewards. Although recent causal approaches aim to address this problem, they lack grounded modeling of reward-guided causal understanding of states and actions for goal-orientation, thus impairing learning efficiency. To tackle this issue, we propose a novel method named Causal Information Prioritization (CIP) that improves sample efficiency by leveraging factored MDPs to infer causal relationships between different dimensions of states and actions with respect to rewards, enabling the prioritization of causal information. Specifically, CIP identifies and leverages causal relationships between states and rewards to execute counterfactual data augmentation to prioritize high-impact state features under the causal understanding of the environments. Moreover, CIP integrates a causality-aware empowerment learning objective, which significantly enhances the agent's execution of reward-guided actions for more efficient exploration in complex environments.
To fully assess the effectiveness of CIP, we conduct extensive experiments across 39 tasks in 5 diverse continuous control environments, encompassing both locomotion and manipulation skills learning with pixel-based and sparse reward settings. Experimental results demonstrate that CIP consistently outperforms existing RL methods across a wide range of scenarios.
****
Jingxuan Fan,Sarah Martinson,Erik Y. Wang,Kaylie Hausknecht,Jonah Brenner,Danxian Liu,Nianli Peng,Corey Wang,Michael Brenner
HARDMath: A Benchmark Dataset for Challenging Problems in Applied Mathematics
Advanced applied mathematics problems are underrepresented in existing Large Language Model (LLM) benchmark datasets. To address this, we introduce $\textbf{HARDMath}$, a dataset inspired by a graduate course on asymptotic methods, featuring challenging applied mathematics problems that require analytical approximation

techniques. These problems demand a combination of mathematical reasoning, comp
utational tools, and subjective judgment, making them difficult for LLMs. Our fr
amework auto-generates a large number of problems with solutions validated again
st numerical ground truths. We evaluate both open- and closed-source LLMs on $\t
extbf{HARDMath-mini}$, a sub-sampled test set of 366 problems, as well as on 40
word problems formulated in applied science contexts. Even leading closed-source
models like GPT-4 achieve only 43.8% overall accuracy with few-shot Chain-of-Th
ought prompting, and all models demonstrate significantly lower performance comp
ared to results on existing mathematics benchmark datasets. We additionally cond
uct a detailed error analysis to gain insights into the failure cases of LLMs. T
hese results demonstrate the limitations of current LLM performance on advanced
graduate-level applied math problems and underscore the importance of datasets l
ike $\textbf{HARDMath}$ to advance mathematical abilities of LLMs.
****

Donghyeok Shin,HeeSun Bae,Gyuwon Sim,Wanmo Kang,Il-chul Moon
Distilling Dataset into Neural Field
Utilizing a large-scale dataset is essential for training high-performance deep
learning models, but it also comes with substantial computation and storage cost
s. To overcome these challenges, dataset distillation has emerged as a promising
solution by compressing the large-scale dataset into a smaller synthetic datase
t that retains the essential information needed for training. This paper propose
s a novel parameterization framework for dataset distillation, coined Distilling
Dataset into Neural Field (DDiF), which leverages the neural field to store the
necessary information of the large-scale dataset. Due to the unique nature of t
he neural field, which takes coordinates as input and output quantity, DDiF effe
ctively preserves the information and easily generates various shapes of data. W
e theoretically confirm that DDiF exhibits greater expressiveness than some prev
ious literature when the utilized budget for a single synthetic instance is the
same. Through extensive experiments, we demonstrate that DDiF achieves superior
performance on several benchmark datasets, extending beyond the image domain to
include video, audio, and 3D voxel. We release the code at \url{https://github.c
om/aailab-kaist/DDiF}.
****

Xiangyu Zeng,Kunchang Li,Chenting Wang,Xinhao Li,Tianxiang Jiang,Ziang Yan,Songz
e Li,Yansong Shi,Zhengrong Yue,Yi Wang,Yali Wang,Yu Qiao,Limin Wang
TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning
Multimodal Large Language Models (MLLMs) have demonstrated impressive performanc
e in short video understanding. However, understanding long-form videos still re
mains challenging for MLLMs. This paper proposes TimeSuite, a collection of new
designs to adapt the existing short-form video MLLMs for long video understandin
g, including a simple yet efficient framework to process long video sequence, a
high-quality video dataset for grounded tuning of MLLMs, and a carefully-designe
d instruction tuning task to explicitly incorporate the grounding supervision in
the traditional QA format. Specifically, based on VideoChat, we propose our lon
g-video MLLM, coined as VideoChat-T, by implementing a token shuffling to compre
ss long video tokens and introducing Temporal Adaptive Position Encoding (TAPE)
to enhance the temporal awareness of visual representation. Meanwhile, we introd
uce the TimePro, a comprehensive grounding-centric instruction tuning dataset co
mposed of 9 tasks and 349k high-quality grounded annotations. Notably, we design
a new instruction tuning task type, called Temporal Grounded Caption, to peform
detailed video descriptions with the corresponding time stamps prediction. This
explicit temporal location prediction will guide MLLM to correctly attend on th
e visual content when generating description, and thus reduce the hallucination
risk caused by the LLMs. Experimental results demonstrate that our TimeSuite pro
vides a successful solution to enhance the long video understanding capability o
f short-form MLLM, achieving improvement of 5.6% and 6.8% on the benchmarks of E
goschema and VideoMME, respectively. In addition, VideoChat-T exhibits robust ze
ro-shot temporal grounding capabilities, significantly outperforming the existin
g state-of-the-art MLLMs. After fine-tuning, it performs on par with the traditi
onal supervised expert models.

****
Julius Aka,Johannes Brunnemann,Jörg Eiden,Arne Speerforck,Lars Mikelsons
Balanced Neural ODEs: nonlinear model order reduction and Koopman operator approximations
Variational Autoencoders (VAEs) are a powerful framework for learning latent representations of reduced dimensionality, while Neural ODEs excel in learning transient system dynamics. This work combines the strengths of both to generate fast surrogate models with adjustable complexity reacting on time-varying inputs signals. By leveraging the VAE’s dimensionality reduction using a non-hierarchical prior, our method adaptively assigns stochastic noise, naturally complementing known NeuralODE training enhancements and enabling probabilistic
time series modeling. We show that standard Latent ODEs struggle with dimensionality reduction in systems with time-varying inputs. Our approach mitigates this by continuously propagating variational parameters through time, establishing fixed information channels in latent space. This results in a flexible and robust method that can learn different system complexities, e.g. deep neural networks or
linear matrices. Hereby, it enables efficient approximation of the Koopman operator without the need for predefining its dimensionality. As our method balances dimensionality reduction and reconstruction accuracy, we call it Balanced Neural ODE (B-NODE). We demonstrate the effectiveness of this methods on several academic and real-world test cases, e.g. a power plant or MuJoCo data.
****
Zhiyong Wu,Zhenyu Wu,Fangzhi Xu,Yian Wang,Qiushi Sun,Chengyou Jia,Kanzhi Cheng,Zichen Ding,Liheng Chen,Paul Pu Liang,Yu Qiao
OS-ATLAS: Foundation Action Model for Generalist GUI Agents
Existing efforts in building GUI agents heavily rely on the availability of robust commercial Vision-Language Models (VLMs) such as GPT-4o and GeminiProVision. Practitioners are often reluctant to use open-source VLMs due to their significant performance lag compared to their closed-source counterparts, particularly in GUI grounding and Out-Of-Distribution (OOD) scenarios. To facilitate future research in this area, we developed OS-Atlas—a foundational GUI action model that excels at GUI grounding and OOD agentic tasks through innovations in both data and modeling.
We have invested significant engineering effort in developing an open-source toolkit for synthesizing GUI grounding data across multiple platforms, including Windows, Linux, MacOS, Android, and the web. Leveraging this toolkit, we are releasing the largest open-source cross-platform GUI grounding corpus to date, which contains over 13 million GUI elements. This dataset, combined with innovations in model training, provides a solid foundation for OS-Atlas to understand GUI screenshots and generalize to unseen interfaces.
Through extensive evaluation across six benchmarks spanning three different platforms (mobile, desktop, and web), OS-Atlas demonstrates significant performance improvements over previous state-of-the-art models. Our evaluation also uncovers valuable insights into continuously improving and scaling the agentic capabilities of open-source VLMs.
****
Shihong Song,Guanlin Mo,Hu Ding
Relax and Merge: A Simple Yet Effective Framework for Solving Fair k-Means and k-sparse Wasserstein Barycenter Problems
The fairness of clustering algorithms has gained widespread attention across various areas, including machine learning, In this paper, we study fair k-means clustering in Euclidean space.
  Given a dataset comprising several groups, the fairness constraint requires that each cluster should contain a proportion of points from each group within specified lower and upper bounds.
  Due to these fairness constraints, determining the optimal locations of k centers is a quite challenging task.
  We propose a novel ``Relax and Merge'' framework that returns a $(1+4\rho + O(\epsilon))$-approximate solution, where ρ is the approximate ratio of an o

ff-the-shelf vanilla k-means algorithm and $O(\epsilon)$ can be an arbitrarily small positive number. If equipped with a PTAS of k-means, our solution can achieve an approximation ratio of $(5+O(\epsilon))$ with only a slight violation of the fairness constraints, which improves the current state-of-the-art approximation guarantee. Furthermore, using our framework, we can also obtain a $(1+4\rho +O(\epsilon))$-approximate solution for the k-sparse Wasserstein Barycenter problem, which is a fundamental optimization problem in the field of optimal transport, and a $(2+6\rho)$-approximate solution for the strictly fair k-means clustering with no violation, both of which are better than the current state-of-the-art methods. In addition, the empirical results demonstrate that our proposed algorithm can significantly outperform baseline approaches in terms of clustering cost.
****
Gabriel Ducrocq,Lukas Grunewald,Sebastian Westenhoff,Fredrik Lindsten
cryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM
The three-dimensional structure of proteins plays a crucial role in determining their function. Protein structure prediction methods, like AlphaFold, offer rapid access to a protein’s structure. However, large protein complexes cannot be reliably predicted, and proteins are dynamic, making it important to resolve their full conformational distribution. Single-particle cryo-electron microscopy (cryo-EM) is a powerful tool for determining the structures of large protein complexes. Importantly, the numerous images of a given protein contain underutilized information about conformational heterogeneity. These images are very noisy projections of the protein, and traditional methods for cryo-EM reconstruction are limited to recovering only one or a few consensus conformations.

In this paper, we introduce cryoSPHERE, which is a deep learning method that uses a nominal protein structure (e.g., from AlphaFold) as input, learns how to divide it into segments, and moves these segments as approximately rigid bodies to fit the different conformations present in the cryo-EM dataset. This approach provides enough constraints to enable meaningful reconstructions of single protein structural ensembles. We demonstrate this with two synthetic datasets featuring varying levels of noise, as well as two real dataset. We show that cryoSPHERE is very resilient to the high levels of noise typically encountered in experiments, where we see consistent improvements over the current state-of-the-art for heterogeneous reconstruction.
****
Pablo Lemos,Sammy Nasser Sharief,Nikolay Malkin,Salma Salhi,Connor Stone,Laurence Perreault-Levasseur,Yashar Hezaveh
PQMass: Probabilistic Assessment of the Quality of Generative Models using Probability Mass Estimation
We propose a likelihood-free method for comparing two distributions given samples from each, with the goal of assessing the quality of generative models. The proposed approach, PQMass, provides a statistically rigorous method for assessing the performance of a single generative model or the comparison of multiple competing models. PQMass divides the sample space into non-overlapping regions and applies chi-squared tests to the number of data samples that fall within each region, giving a p-value that measures the probability that the bin counts derived from two sets of samples are drawn from the same multinomial distribution. PQMass does not depend on assumptions regarding the density of the true distribution, nor does it rely on training or fitting any auxiliary models. We evaluate PQMass on data of various modalities and dimensions, demonstrating its effectiveness in assessing the quality, novelty, and diversity of generated samples. We further show that PQMass scales well to moderately high-dimensional data and thus obviates the need for feature extraction in practical applications.
****
Hongkang Li,Songtao Lu,Pin-Yu Chen,Xiaodong Cui,Meng Wang
Training Nonlinear Transformers for Chain-of-Thought Inference: A Theoretical Generalization Analysis
Chain-of-Thought (CoT) is an efficient prompting method that enables the reasoni

ng ability of large language models by augmenting the query using multiple examp
les with multiple intermediate steps. Despite the empirical success, the theoret
ical understanding of how to train a Transformer to achieve the CoT ability rema
ins less explored. This is primarily due to the technical challenges involved in
 analyzing the nonconvex optimization on nonlinear attention models. To the best
 of our knowledge, this work provides the first theoretical study of training Tr
ansformers with nonlinear attention to obtain the CoT generalization capability
so that the resulting model can inference on unseen tasks when the input is augm
ented by examples of the new task. We first quantify the required training sampl
es and iterations to train a Transformer model towards CoT ability. We then pro
ve the success of its CoT generalization on unseen tasks with distribution-shift
ed testing data. Moreover, we theoretically characterize the conditions for an a
ccurate reasoning output by CoT even when the provided reasoning examples contai
n noises and are not always accurate. In contrast, in-context learning (ICL), wh
ich can be viewed as one-step CoT without intermediate steps, may fail to provid
e an accurate output when CoT does. These theoretical findings are justified thr
ough experiments.
****

Shengqiong Wu,Hao Fei,Xiangtai Li,Jiayi Ji,Hanwang Zhang,Tat-Seng Chua,Shuicheng
 YAN

Towards Semantic Equivalence of Tokenization in Multimodal LLM

Multimodal Large Language Models (MLLMs) have demonstrated exceptional capabilit
ies in processing vision-language tasks. One of the crux of MLLMs lies in vision
 tokenization, which involves efficiently transforming input visual signals into
 feature representations that are most beneficial for LLMs. However, existing vi
sion tokenizers, essential for semantic alignment between vision and language, r
emain problematic. Existing methods aggressively fragment visual input, corrupti
ng the visual semantic integrity. To address this, this paper proposes a novel d
ynamic Semantic-Equivalent Vision Tokenizer (SeTok), which groups visual feature
s into semantic units via a dynamic clustering algorithm, flexibly determining t
he number of tokens based on image complexity. The resulting vision tokens effec
tively preserve semantic integrity and capture both low-frequency and high-frequ
ency visual features. The proposed MLLM (Setokim) equipped with SeTok significan
tly demonstrates superior performance across various tasks, as evidenced by our
experimental results.
****

Erik Bodin,Alexandru I. Stere,Dragos D Margineantu,Carl Henrik Ek,Henry Moss

Linear combinations of latents in generative models: subspaces and beyond

Sampling from generative models has become a crucial tool for applications like
data synthesis and augmentation. Diffusion, Flow Matching and Continuous Normali
zing Flows have shown effectiveness across various modalities, and rely on laten
t variables for generation. For experimental design or creative applications tha
t require more control over the generation process, it has become common to mani
pulate the latent variable directly. However, existing approaches for performing
 such manipulations (e.g. interpolation or forming low-dimensional representatio
ns) only work well in special cases or are network or data-modality specific.
We propose Linear combinations of Latent variables (LOL) as a general-purpose me
thod to form linear combinations of latent variables that adhere to the assumpti
ons of the generative model. As LOL is easy to implement and naturally addresses
 the broader task of forming any linear combinations, e.g. the construction of s
ubspaces of the latent space, LOL dramatically simplifies the creation of expres
sive low-dimensional representations of high-dimensional objects.
****

Carlo Alfano,Sebastian Rene Towers,Silvia Sapora,Chris Lu,Patrick Rebeschini

Learning mirror maps in policy mirror descent

Policy Mirror Descent (PMD) is a popular framework in reinforcement learning, se
rving as a unifying perspective that encompasses numerous algorithms. These algo
rithms are derived through the selection of a mirror map and enjoy finite-time c
onvergence guarantees. Despite its popularity, the exploration of PMD's full pot
ential is limited, with the majority of research focusing on a particular mirror

map---namely, the negative entropy---which gives rise to the renowned Natural Policy Gradient (NPG) method. It remains uncertain from existing theoretical studies whether the choice of mirror map significantly influences PMD's efficacy. In our work, we conduct empirical investigations to show that the conventional mirror map choice (NPG) often yields less-than-optimal outcomes across several standard benchmark environments. Using evolutionary strategies, we identify more efficient mirror maps that enhance the performance of PMD. We first focus on a tabular environment, i.e.\ Grid-World, where we relate existing theoretical bounds with the performance of PMD for a few standard mirror maps and the learned one. We then show that it is possible to learn a mirror map that outperforms the negative entropy in more complex environments, such as the MinAtar suite. Additionally, we demonstrate that the learned mirror maps generalize effectively to different tasks by testing each map across various other environments.
****

Shulan Zhu,Chenglong Bao,Defeng Sun,Yancheng Yuan

A Tight Convergence Analysis of Inexact Stochastic Proximal Point Algorithm for Stochastic Composite Optimization Problems

The \textbf{i}nexact \textbf{s}tochastic \textbf{p}roximal \textbf{p}oint \textbf{a}lgorithm (isPPA) is popular for solving stochastic composite optimization problems with many applications in machine learning. While the convergence theory of the (inexact) PPA has been well established, the known convergence guarantees of isPPA require restrictive assumptions. In this paper, we establish the stability and almost sure convergence of isPPA under mild assumptions, where smoothness and (restrictive) strong convexity of the objective function are not required. Imposing a local Lipschitz condition on component functions and a quadratic growth condition on the objective function, we establish last-iterate iteration complexity bounds of isPPA regarding the distance to the solution set and the Karush–Kuhn–Tucker (KKT) residual. Moreover, we show that the established iteration complexity bounds are tight up to a constant by explicitly analyzing the bounds for the regularized Fr\'echet mean problem. We further validate the established convergence guarantees of isPPA by numerical experiments.
****

Yunshi Wen,Tengfei Ma,Ronny Luss,Debarun Bhattacharjya,Achille Fokoue,Anak Agung Julius

Shedding Light on Time Series Classification using Interpretability Gated Networks

In time-series classification, interpretable models can bring additional insights but be outperformed by deep models since human-understandable features have limited expressivity and flexibility. In this work, we present InterpGN, a framework that integrates an interpretable model and a deep neural network. Within this framework, we introduce a novel gating function design based on the confidence of the interpretable expert, preserving interpretability for samples where interpretable features are significant while also identifying samples that require additional expertise. For the interpretable expert, we incorporate shapelets to effectively model shape-level features for time-series data. We introduce a variant of Shapelet Transforms to build logical predicates using shapelets. Our proposed model achieves comparable performance with state-of-the-art deep learning models while additionally providing interpretable classifiers for various benchmark datasets. We further show that our models improve on quantitative shapelet quality and interpretability metrics over existing shapelet-learning formulations. Finally, we show that our models can integrate additional advanced architectures and be applied to real-world tasks beyond standard benchmarks such as the MIMIC-III and time series extrinsic regression datasets.
****

Juno Kim,Taiji Suzuki

Transformers Provably Solve Parity Efficiently with Chain of Thought

This work provides the first theoretical analysis of training transformers to solve complex problems by recursively generating intermediate states, analogous to fine-tuning for chain-of-thought (CoT) reasoning. We consider training a one-layer transformer to solve the fundamental k-parity problem, extending the work

on RNNs by \citet{Wies23}. We establish three key results: (1) any finite-precis
ion gradient-based algorithm, without intermediate supervision, requires substan
tial iterations to solve parity with finite samples. (2) In contrast, when inter
mediate parities are incorporated into the loss function, our model can learn pa
rity in one gradient update when aided by \emph{teacher forcing}, where ground-t
ruth labels of the reasoning chain are provided at each generation step. (3) Eve
n without teacher forcing, where the model must generate CoT chains end-to-end,
parity can be learned efficiently if augmented data is employed to internally ve
rify the soundness of intermediate steps. Our findings, supported by numerical e
xperiments, show that task decomposition and stepwise reasoning naturally arise
from optimizing transformers with CoT; moreover, self-consistency checking can i
mprove multi-step reasoning ability, aligning with empirical studies of CoT.
****

Behrooz Tahmasebi,Stefanie Jegelka

Generalization Bounds for Canonicalization: A Comparative Study with Group Avera
ging

Canonicalization, a popular method for generating invariant or equivariant funct
ion classes from arbitrary function sets, involves initial data projection onto
a reduced input space subset, followed by applying any learning method to the pr
ojected dataset. Despite recent research on the expressive power and continuity
of functions represented by canonicalization, its generalization capabilities re
main less explored. This paper addresses this gap by theoretically examining the
 generalization benefits and sample complexity of canonicalization, comparing th
em with group averaging, another popular technique for creating invariant or equ
ivariant function classes.

Our findings reveal two distinct regimes where canonicalization may outperform o
r underperform compared to group averaging, with precise quantification of this
phase transition in terms of sample size, group action characteristics, and a ne
wly introduced concept of alignment.

To the best of our knowledge, this study represents the first theoretical explor
ation of such behavior, offering insights into the relative effectiveness of can
onicalization and group averaging under varying conditions.
****

Yuhui Xu,Zhanming Jie,Hanze Dong,Lei Wang,Xudong Lu,Aojun Zhou,Amrita Saha,Caimi
ng Xiong,Doyen Sahoo

ThinK: Thinner Key Cache by Query-Driven Pruning

Large Language Models (LLMs) have revolutionized the field of natural language p
rocessing, achieving unprecedented performance across a variety of applications.

However, their increased computational and memory demands present significant ch
allenges, especially when handling long sequences.

This paper focuses on the long-context scenario, addressing the inefficiencies i
n KV cache memory consumption during inference.

Unlike existing approaches that optimize the memory based on the sequence length
, we identify substantial redundancy in the channel dimension of the KV cache, a
s indicated by an uneven magnitude distribution and a low-rank structure in the
attention weights.

In response, we propose ThinK, a novel query-dependent KV cache pruning method d
esigned to minimize attention weight loss while selectively pruning the least si
gnificant channels. Our approach not only maintains or enhances model accuracy b
ut also achieves a reduction in KV cache memory costs by over 20\% compared with
 vanilla KV cache eviction and quantization methods. For instance, ThinK integra
ted with KIVI can achieve $2.8\times$ peak memory reduction while maintaining ne
arly the same quality, enabling a batch size increase from 4$\times$ (with KIVI
alone) to 5$\times$ when using a single GPU. Extensive evaluations on the LLaMA
and Mistral models across various long-sequence datasets verified the efficiency
 of ThinK. Our code has been made available at https://github.com/SalesforceAIR
esearch/ThinK.
****

Donggon Jang,Yucheol Cho,Suin Lee,Taehyeon Kim,Daeshik Kim

MMR: A Large-scale Benchmark Dataset for Multi-target and Multi-granularity Reasoning Segmentation

The fusion of Large Language Models (LLMs) with vision models is pioneering new possibilities in user-interactive vision-language tasks. A notable application is reasoning segmentation, where models generate pixel-level segmentation masks by comprehending implicit meanings in human instructions. However, seamless human-AI interaction demands more than just object-level recognition; it requires understanding both objects and the functions of their detailed parts, particularly in multi-target scenarios. For example, when instructing a robot to \textit{“turn on the TV"}, there could be various ways to accomplish this command. Recognizing multiple objects capable of turning on the TV, such as the TV itself or a remote control (multi-target), provides more flexible options and aids in finding the optimized scenario. Furthermore, understanding specific parts of these objects, like the TV's button or the remote's button (part-level), is important for completing the action. Unfortunately, current reasoning segmentation datasets predominantly focus on a single target object-level reasoning, which limits the detailed recognition of an object's parts in multi-target contexts. To address this gap, we construct a large-scale dataset called Multi-target and Multi-granularity Reasoning (MMR). MMR comprises 194K complex and implicit instructions that consider multi-target, object-level, and part-level aspects, based on pre-existing image-mask sets. This dataset supports diverse and context-aware interactions by hierarchically providing object and part information. Moreover, we propose a straightforward yet effective framework for multi-target, object-level, and part-level reasoning segmentation. Experimental results on MMR show that the proposed method can reason effectively in multi-target and multi-granularity scenarios, while the existing reasoning segmentation model still has room for improvement. The dataset is available at \url{https://github.com/jdg900/MMR}.

****

Junwei Zhou,Xueting Li,Lu Qi,Ming-Hsuan Yang

Layout-your-3D: Controllable and Precise 3D Generation with 2D Blueprint

We present Layout-Your-3D, a framework that allows controllable and compositional 3D generation from text prompts. Existing text-to-3D methods often struggle to generate assets with plausible object interactions or require tedious optimization processes. To address these challenges, our approach leverages 2D layouts as a blueprint to facilitate precise and plausible control over 3D generation. Starting with a 2D layout provided by a user or generated from a text description, we first create a coarse 3D scene using a carefully designed initialization process based on efficient reconstruction models. To enforce coherent global 3D layouts and enhance the quality of instance appearances, we propose a collision-aware layout optimization process followed by instance-wise refinement. Experimental results demonstrate that Layout-Your-3D yields more reasonable and visually appealing compositional 3D assets while significantly reducing the time required for each prompt. Additionally, Layout-Your-3D can be easily applicable to downstream tasks, such as 3D editing and object insertion.

****

Wentao Guo,Jikai Long,Yimeng Zeng,Zirui Liu,Xinyu Yang,Yide Ran,Jacob R. Gardner,Osbert Bastani,Christopher De Sa,Xiaodong Yu,Beidi Chen,Zhaozhuo Xu

Zeroth-Order Fine-Tuning of LLMs with Transferable Static Sparsity

Zeroth-order optimization (ZO) is a memory-efficient strategy for fine-tuning Large Language Models using only forward passes. However, applying ZO fine-tuning in memory-constrained settings such as mobile phones and laptops remains challenging since these settings often involve weight quantization, while ZO requires full-precision perturbation and update. In this study, we address this limitation by combining static sparse ZO fine-tuning with quantization. Our approach transfers a small, static subset (0.1%) of "sensitive" parameters from pre-training to downstream tasks, focusing fine-tuning on this sparse set of parameters. The remaining untuned parameters are quantized, reducing memory demands. Our proposed workflow enables efficient ZO fine-tuning of an Llama2-7B model on a GPU device with less than 8GB of memory while outperforming full model ZO fine-tuning performance and in-context learning.

****
Yikun Zhang,Geyan Ye,Chaohao Yuan,Bo Han,Long-Kai Huang,Jianhua Yao,Wei Liu,Yu Rong
Atomas: Hierarchical Adaptive Alignment on Molecule-Text for Unified Molecule Understanding and Generation
Molecule-and-text cross-modal representation learning has emerged as a promising direction for enhancing the quality of molecular representation, thereby improving performance in various scientific fields. However, most approaches employ a global alignment approach to learn the knowledge from different modalities that may fail to capture fine-grained information, such as molecule-and-text fragments and stereoisomeric nuances, which is crucial for downstream tasks. Furthermore, it is incapable of modeling such information using a similar global alignment strategy due to the lack of annotations about the fine-grained fragments in the existing dataset.
In this paper, we propose Atomas, a hierarchical molecular representation learning framework that jointly learns representations from SMILES strings and text. We design a Hierarchical Adaptive Alignment model to automatically learn the fine-grained fragment correspondence between two modalities and align these representations at three semantic levels.
Atomas's end-to-end training framework supports understanding and generating molecules, enabling a wider range of downstream tasks. Atomas achieves superior performance across 12 tasks on 11 datasets, outperforming 11 baseline models thus highlighting the effectiveness and versatility of our method. Scaling experiments further demonstrate Atomas’s robustness and scalability. Moreover, visualization and qualitative analysis, validated by human experts, confirm the chemical relevance of our approach. Codes are released on ~\url{https://github.com/yikunpku/Atomas}.
****
Gregor Bachmann,Sotiris Anagnostidis,Albert Pumarola,Markos Georgopoulos,Artsiom Sanakoyeu,Yuming Du,Edgar Schönfeld,Ali Thabet,Jonas K Kohler
Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment
The performance of large language models (LLMs) is closely linked to their underlying size, leading to ever-growing networks and hence slower inference. Speculative decoding has been proposed as a technique to accelerate autoregressive generation, leveraging a fast draft model to propose candidate tokens, which are then verified in parallel based on their likelihood under the target model. While this approach guarantees to reproduce the target output, it incurs a substantial penalty: many high-quality draft tokens are rejected, even when they represent objectively valid continuations. Indeed, we show that even powerful draft models such as GPT-4o, as well as human text cannot achieve high acceptance rates under the standard verification scheme. This severely limits the speedup potential of current speculative decoding methods, as an early rejection becomes overwhelmingly likely when solely relying on alignment of draft and target.
We thus ask the following question: Can we adapt verification to recognize correct, but non-aligned replies? To this end, we draw inspiration from the LLM-as-a-judge framework, which demonstrated that LLMs are able to rate answers in a versatile way. We carefully design a dataset coined TokenCourt to elicit the same capability in the target model by training a compact module on top of the embeddings to produce ``judgements" of the current continuation. We showcase our strategy on the Llama-3.1 family, where our 8B/405B-Judge achieves a speedup of $9\times$ over Llama-405B, while maintaining its quality on a large range of benchmarks. These benefits remain present even in optimized inference frameworks, where our method reaches up to 141 tokens/s for 8B/70B-Judge and 129 tokens/s for 8B/405B on 2 and 8 H100s respectively.
****
Yuda Song,Hanlin Zhang,Carson Eisenach,Sham M. Kakade,Dean Foster,Udaya Ghai
Mind the Gap: Examining the Self-Improvement Capabilities of Large Language Models
Self-improvement is a mechanism in Large Language Model (LLM) pre-training, post

-training and test-time inference. We explore a framework where the model verifi
es its own outputs, filters or reweights data based on this verification, and di
stills the filtered data. Despite several empirical successes, a fundamental un
derstanding is still lacking. In this work, we initiate a comprehensive, modular
 and controlled study on LLM self-improvement. We provide a mathematical formula
tion for self-improvement, which is largely governed by a quantity which we form
alize as the **generation-verification gap**. Through experiments with various m
odel families and tasks, we discover a scaling phenomenon of self-improvement --
 a variant of the generation-verification gap scales monotonically with the mode
l pre-training flops. We also examine when self-improvement is possible, an iter
ative self-improvement procedure, and ways to improve its performance. Our findi
ngs not only advance understanding of LLM self-improvement with practical implic
ations, but also open numerous avenues for future research into its capabilities
 and boundaries.
****
Cassidy Laidlaw,Shivam Singhal,Anca Dragan
Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking
Because it is difficult to precisely specify complex objectives, reinforcement l
earning policies are often optimized using proxy reward functions that only appr
oximate the true goal. However, optimizing proxy rewards frequently leads to rew
ard hacking: the optimized reward function ceases to be a good proxy and the res
ulting policy performs poorly with respect to the unspecified true reward. Princ
ipled solutions to reward hacking have been impeded by the lack of a good defini
tion for the problem. To address this gap, we introduce a definition of reward h
acking based on the correlation between proxy and true rewards for states and ac
tions seen by a “reference policy” that breaks down under optimization. We show
that this definition captures reward hacking behavior across several realistic s
ettings, including in reinforcement learning from human feedback (RLHF). Using o
ur formulation, we show theoretically that regularization to the reference polic
y can effectively prevent reward hacking. While the current practice in RLHF app
lies a KL penalty between action distributions for this purpose, our theory sugg
ests regularizing the $\chi 2$ divergence between the policies’ occupancy measures can
 be more effective. We intuitively show the benefits of this type of regularizat
ion and demonstrate that it better mitigates reward hacking in practice across f
our realistic settings, including RLHF. Our code is available at https://github.
com/cassidylaidlaw/orpo.
****
Qinshuo Liu,Weiqin Zhao,Wei Huang,Yanwen Fang,Lequan Yu,Guodong Li
From Layers to States: A State Space Model Perspective to Deep Neural Network La
yer Dynamics
The depth of neural networks is a critical factor for their capability, with dee
per models often demonstrating superior performance. Motivated by this, signific
ant efforts have been made to enhance layer aggregation - reusing information fr
om previous layers to better extract features at the current layer, to improve t
he representational power of deep neural networks. However, previous works have
primarily addressed this problem from a discrete-state perspective which is not
suitable as the number of network layers grows. This paper novelly treats the ou
tputs from layers as states of a continuous process and considers leveraging the
 state space model (SSM) to design the aggregation of layers in very deep neural
 networks. Moreover, inspired by its advancements in modeling long sequences, t
he Selective State Space Models (S6) is employed to design a new module called S
elective State Space Model Layer Aggregation (S6LA). This module aims to combine
 traditional CNN or transformer architectures within a sequential framework, enh
ancing the representational capabilities of state-of-the-art vision networks. Ex
tensive experiments show that S6LA delivers substantial improvements in both ima
ge classification and detection tasks, highlighting the potential of integrating
 SSMs with contemporary deep learning techniques.
****
Leigang Qu,Haochuan Li,Tan Wang,Wenjie Wang,Yongqi Li,Liqiang Nie,Tat-Seng Chua
TIGeR: Unifying Text-to-Image Generation and Retrieval with Large Multimodal Mod

els
How humans can effectively and efficiently acquire images has always been a pere
nnial question. A classic solution is *text-to-image retrieval* from an existing
 database; however, the limited database typically lacks creativity. By contrast
, recent breakthroughs in *text-to-image generation* have made it possible to pr
oduce attractive and counterfactual visual content, but it faces challenges in s
ynthesizing knowledge-intensive images. In this work, we rethink the relationshi
p between text-to-image generation and retrieval, proposing a *unified* framewor
k for both tasks with one single Large Multimodal Model (LMM). Specifically, we
first explore the intrinsic discriminative abilities of LMMs and introduce an ef
ficient generative retrieval method for text-to-image retrieval in a training-fr
ee manner. Subsequently, we unify generation and retrieval autoregressively and
propose an autonomous decision mechanism to choose the best-matched one between
generated and retrieved images as the response to the text prompt. To standardiz
e the evaluation of unified text-to-image generation and retrieval, we construct
 TIGeR-Bench, a benchmark spanning both creative and knowledge-intensive domains
. Extensive experiments on TIGeR-Bench and two retrieval benchmarks, *i.e.*, Fli
ckr30K and MS-COCO, demonstrate the superiority of our proposed framework.
****
Feijie Wu,Xiaoze Liu,Haoyu Wang,Xingchen Wang,Lu Su,Jing Gao
Towards Federated RLHF with Aggregated Client Preference for LLMs
Reinforcement learning with human feedback (RLHF) fine-tunes a pretrained large
language model (LLM) using user preference data, enabling it to generate content
 aligned with human preferences. However, due to privacy concerns, users may be
reluctant to share sensitive preference data. To address this, we propose utiliz
ing Federated Learning (FL) techniques, allowing large-scale preference collecti
on from diverse real-world users without requiring them to transmit data to a ce
ntral server. Our federated RLHF methods (i.e., FedBis and FedBiscuit) encode ea
ch client’s preferences into binary selectors and aggregate them to capture comm
on preferences. In particular, FedBiscuit overcomes key challenges, such as pref
erence heterogeneity and reward hacking, through innovative solutions like group
ing clients with similar preferences to reduce heterogeneity and using multiple
binary selectors to enhance LLM output quality. To evaluate the performance of t
he proposed methods, we establish the first federated RLHF benchmark with a hete
rogeneous human preference dataset. Experimental results show that by integratin
g the LLM with aggregated client preferences, FedBis and FedBiscuit significantl
y enhance the professionalism and readability of the generated content.
****
Changyeon Kim,Minho Heo,Doohyun Lee,Honglak Lee,Jinwoo Shin,Joseph J Lim,Kimin L
ee
Subtask-Aware Visual Reward Learning from Segmented Demonstrations
Reinforcement Learning (RL) agents have demonstrated their potential across vari
ous robotic tasks. However, they still heavily rely on human-engineered reward f
unctions, requiring extensive trial-and-error and access to target behavior info
rmation, often unavailable in real-world settings. This paper introduces REDS: R
Eward learning from Demonstration with Segmentations, a novel reward learning fr
amework that leverages action-free videos with minimal supervision. Specifically
, REDS employs video demonstrations segmented into subtasks from diverse sources
 and treats these segments as ground-truth rewards. We train a dense reward func
tion conditioned on video segments and their corresponding subtasks to ensure al
ignment with ground-truth reward signals by minimizing the Equivalent-Policy Inv
ariant Comparison distance. Additionally, we employ contrastive learning objecti
ves to align video representations with subtasks, ensuring precise subtask infer
ence during online interactions. Our experiments show that REDS significantly ou
tperforms baseline methods on complex robotic manipulation tasks in Meta-World a
nd more challenging real-world tasks, such as furniture assembly in FurnitureBen
ch, with minimal human intervention. Moreover, REDS facilitates generalization t
o unseen tasks and robot embodiments, highlighting its potential for scalable de
ployment in diverse environments.
****

Alireza Rezazadeh,Zichao Li,Wei Wei,Yujia Bao
From Isolated Conversations to Hierarchical Schemas: Dynamic Tree Memory Represe
ntation for LLMs
Recent advancements in large language models have significantly improved their c
ontext windows, yet challenges in effective long-term memory management remain.
We introduce MemTree, an algorithm that leverages a dynamic, tree-structured mem
ory representation to optimize the organization, retrieval, and integration of i
nformation, akin to human cognitive schemas. MemTree organizes memory hierarchic
ally, with each node encapsulating aggregated textual content, corresponding sem
antic embeddings, and varying abstraction levels across the tree's depths. Our a
lgorithm dynamically adapts this memory structure by computing and comparing sem
antic embeddings of new and existing information to enrich the model’s context-a
wareness. This approach allows MemTree to handle complex reasoning and extended
interactions more effectively than traditional memory augmentation methods, whic
h often rely on flat lookup tables. Evaluations on benchmarks for multi-turn dia
logue understanding and document question answering show that MemTree significan
tly enhances performance in scenarios that demand structured memory management.
****
Hyungjoo Chae,Namyoung Kim,Kai Tzu-iunn Ong,Minju Gwak,Gwanwoo Song,Jihoon Kim,S
unghwan Kim,Dongha Lee,Jinyoung Yeo
Web Agents with World Models: Learning and Leveraging Environment Dynamics in We
b Navigation
Large language models (LLMs) have recently gained much attention in building aut
onomous agents. However, performance of current LLM-based web agents in long-hor
izon tasks is far from optimal, often yielding errors such as repeatedly buying
a non-refundable flight ticket. By contrast, humans can avoid such an irreversib
le mistake, as we have an awareness of the potential outcomes (e.g., losing mone
y) of our actions, also known as the "world model". Motivated by this, our study
 first starts with preliminary analyses, confirming the absence of world models
in current LLMs (e.g., GPT-4o, Claude-3.5-Sonnet, etc.). Then, we present a Worl
d-model-augmented (WMA) web agent, which simulates the outcomes of its actions f
or better decision-making. To overcome the challenges in training LLMs as world
models predicting next observations, such as repeated elements across observatio
ns and long HTML inputs, we propose a transition-focused observation abstraction
, where the prediction objectives are free-form natural language descriptions ex
clusively highlighting important state differences between time steps. Experimen
ts on WebArena and Mind2Web show that our world models improve agents' policy se
lection without training and demonstrate our agents' cost- and time-efficiency c
ompared to recent tree-search-based agents.
****
Hung-Yueh Chiang,Chi-Chih Chang,Natalia Frumkin,Kai-Chiang Wu,Diana Marculescu
Quamba: A Post-Training Quantization Recipe for Selective State Space Models
State Space Models (SSMs) have emerged as an appealing alternative to Transforme
rs for large language models, achieving state-of-the-art accuracy with constant
memory complexity which allows for holding longer context lengths than attention
-based networks. The superior computational efficiency of SSMs in long sequence
modeling positions them favorably over Transformers in many scenarios. However,
improving the efficiency of SSMs on request-intensive cloud-serving and resource
-limited edge applications is still a formidable task. SSM quantization is a pos
sible solution to this problem, making SSMs more suitable for wide deployment, w
hile still maintaining their accuracy. Quantization is a common technique to red
uce the model size and to utilize the low bit-width acceleration features on mod
ern computing units, yet existing quantization techniques are poorly suited for
SSMs. Most notably, SSMs have highly sensitive feature maps within the selective
 scan mechanism (i.e., linear recurrence) and massive outliers in the output act
ivations which are not present in the output of token-mixing in the self-attenti
on modules. To address this issue, we propose a static 8-bit per-tensor SSM quan
tization method which suppresses the maximum values of the input activations to
the selective SSM for finer quantization precision and quantizes the output acti
vations in an outlier-free space with Hadamard transform. Our 8-bit weight-activ

ation quantized Mamba 2.8B SSM benefits from hardware acceleration and achieves a 1.72 $\times$ lower generation latency on an Nvidia Orin Nano 8G, with only a 0.9\% drop in average accuracy on zero-shot tasks. When quantizing Jamba, a 52B parameter SSM-style language model, we observe only a 1% drop in accuracy, demonstrating that our SSM quantization method is both effective and scalable for large language models, which require appropriate compression techniques for deployment. The experiments demonstrate the effectiveness and practical applicability of our approach for deploying SSM-based models of all sizes on both cloud and edge platforms.
****

Jaehyun Park,Dongmin Park,Jae-Gil Lee
Active Learning for Continual Learning: Keeping the Past Alive in the Present
Continual learning (CL) enables deep neural networks to adapt to ever-changing data distributions. In practice, there may be scenarios where annotation is costly, leading to *active continual learning (ACL)*, which performs *active learning (AL)* for the CL scenarios when reducing the labeling cost by selecting the most informative subset is preferable. However, conventional AL strategies are not suitable for ACL, as they focus solely on learning the new knowledge, leading to *catastrophic forgetting* of previously learned tasks. Therefore, ACL requires a new AL strategy that can balance the prevention of catastrophic forgetting and the ability to quickly learn new tasks. In this paper, we propose **AccuACL**, **Accu**mulated informativeness-based **A**ctive **C**ontinual **L**earning, by the novel use of the Fisher information matrix as a criterion for sample selection, derived from a theoretical analysis of the Fisher-optimality preservation properties within the framework of ACL, while also addressing the scalability issue of Fisher information-based AL. Extensive experiments demonstrate that AccuACL significantly outperforms AL baselines across various CL algorithms, increasing the average accuracy and forgetting by 23.8% and 17.0%, respectively, on average.
****

Zhi Cen,Huaijin Pi,Sida Peng,Qing Shuai,Yujun Shen,Hujun Bao,Xiaowei Zhou,Ruizhen Hu
Ready-to-React: Online Reaction Policy for Two-Character Interaction Generation
This paper addresses the task of generating two-character online interactions. Previously, two main settings existed for two-character interaction generation: (1) generating one's motions based on the counterpart's complete motion sequence, and (2) jointly generating two-character motions based on specific conditions. We argue that these settings fail to model the process of real-life two-character interactions, where humans will react to their counterparts in real time and act as independent individuals. In contrast, we propose an online reaction policy, called Ready-to-React, to generate the next character pose based on past observed motions. Each character has its own reaction policy as its ``brain'', enabling them to interact like real humans in a streaming manner. Our policy is implemented by incorporating a diffusion head into an auto-regressive model, which can dynamically respond to the counterpart's motions while effectively mitigating the error accumulation throughout the generation process. We conduct comprehensive experiments using the challenging boxing task. Experimental results demonstrate that our method outperforms existing baselines and can generate extended motion sequences. Additionally, we show that our approach can be controlled by sparse signals, making it well-suited for VR and other online interactive environments. Code and data will be made publicly available.
****

Bin CAO,Yang Liu,Zinan Zheng,Ruifeng Tan,Jia Li,Tong-yi Zhang
SimXRD-4M: Big Simulated X-ray Diffraction Data and Crystal Symmetry Classification Benchmark
Powder X-ray diffraction (XRD) patterns are highly effective for crystal identification and play a pivotal role in materials discovery. While machine learning (ML) has advanced the analysis of powder XRD patterns, progress has been constrained by the limited availability of training data and established benchmarks. To address this, we introduce SimXRD, the largest open-source simulated XRD pattern

dataset to date, aimed at accelerating the development of crystallographic info
rmatics. We developed a novel XRD simulation method that incorporates comprehens
ive physical interactions, resulting in a high-fidelity database. SimXRD compris
es 4,065,346 simulated powder XRD patterns, representing 119,569 unique crystal
structures under 33 simulated conditions that reflect real-world variations. We
benchmark 21 sequence models in both in-library and out-of-library scenarios and
analyze the impact of class imbalance in long-tailed crystal label distribution
s. Remarkably, we find that: (1) current neural networks struggle with classifyi
ng low-frequency crystals, particularly in out-of-library situations; (2) models
trained on SimXRD can generalize to real experimental data.
****

Vladimir Fanaskov,Ivan Oseledets

Associative memory and dead neurons

In ``Large Associative Memory Problem in Neurobiology and Machine Learning,'' Dm
itry Krotov and John Hopfield introduced a general technique for the systematic
construction of neural ordinary differential equations with non-increasing energ
y or Lyapunov function. We study this energy function and identify that it is vu
lnerable to the problem of dead neurons. Each point in the state space where the
neuron dies is contained in a non-compact region with constant energy. In these
flat regions, energy function alone does not completely determine all degrees o
f freedom and, as a consequence, can not be used to analyze stability or find st
eady states or basins of attraction. We perform a direct analysis of the dynamic
al system and show how to resolve problems caused by flat directions correspondi
ng to dead neurons: (i) all information about the state vector at a fixed point
can be extracted from the energy and Hessian matrix (of Lagrange function), (ii)
it is enough to analyze stability in the range of Hessian matrix, (iii) if stea
dy state touching flat region is stable the whole flat region is the basin of at
traction. The analysis of the Hessian matrix can be complicated for realistic ar
chitectures, so we show that for a slightly altered dynamical system (with the s
ame structure of steady states), one can derive a diverse family of Lyapunov fun
ctions that do not have flat regions corresponding to dead neurons. In addition,
these energy functions allow one to use Lagrange functions with Hessian matrice
s that are not necessarily positive definite and even consider architectures wit
h non-symmetric feedforward and feedback connections.
****

Han-Lin Hsieh,Maryam Shanechi

Probabilistic Geometric Principal Component Analysis with application to neural
data

Dimensionality reduction is critical across various domains of science including
neuroscience. Probabilistic Principal Component Analysis (PPCA) is a prominent
dimensionality reduction method that provides a probabilistic approach unlike t
he deterministic approach of PCA and serves as a connection between PCA and Fact
or Analysis (FA). Despite their power, PPCA and its extensions are mainly based
on linear models and can only describe the data in a Euclidean coordinate system
around the mean of data. However, in many neuroscience applications, data may b
e distributed around a nonlinear geometry (i.e., manifold) rather than lying in
the Euclidean space around the mean. We develop Probabilistic Geometric Principa
l Component Analysis (PGPCA) for such datasets as a new dimensionality reduction
algorithm that can explicitly incorporate knowledge about a given nonlinear man
ifold that is first fitted from these data. Further, we show how in addition to
the Euclidean coordinate system, a geometric coordinate system can be derived fo
r the manifold to capture the deviations of data from the manifold and noise. We
also derive a data-driven EM algorithm for learning the PGPCA model parameters.
As such, PGPCA generalizes PPCA to better describe data distributions by incorp
orating a nonlinear manifold geometry. In simulations and brain data analyses, w
e show that PGPCA can effectively model the data distribution around various giv
en manifolds and outperforms PPCA for such data. Moreover, PGPCA provides the ca
pability to test whether the new geometric coordinate system better describes th
e data than the Euclidean one. Finally, PGPCA can perform dimensionality reducti
on and learn the data distribution both around and on the manifold. These capabi

lities make PGPCA valuable for enhancing the efficacy of dimensionality reductio
n for analysis of high-dimensional data that exhibit noise and are distributed a
round a nonlinear manifold, especially for neural data.
****
Yang Tian,Sizhe Yang,Jia Zeng,Ping Wang,Dahua Lin,Hao Dong,Jiangmiao Pang
Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulatio
n
Current efforts to learn scalable policies in robotic manipulation primarily fal
l into two categories: one focuses on "action," which involves behavior cloning
from extensive collections of robotic data, while the other emphasizes "vision,"
 enhancing model generalization by pre-training representations or generative mo
dels, also referred to as world models, using large-scale visual datasets. This
paper presents an end-to-end paradigm that predicts actions using inverse dynami
cs models conditioned on the robot's forecasted visual states, named Predictive
Inverse Dynamics Models (PIDM). By closing the loop between vision and action, t
he end-to-end PIDM can be a better scalable action learner. In practice, we use
Transformers to process both visual states and actions, naming the model Seer. I
t is initially pre-trained on large-scale robotic datasets, such as DROID, and c
an be adapted to real-world scenarios with a little fine-tuning data. Thanks to
large-scale, end-to-end training and the continuous synergy between vision and a
ction at each execution step, Seer significantly outperforms state-of-the-art me
thods across both simulation and real-world experiments. It achieves improvement
s of 13% on the LIBERO-LONG benchmark, 22% on CALVIN ABC-D, and 43% in real-worl
d tasks. Notably, it demonstrates superior generalization for novel objects, lig
hting conditions, and environments under high-intensity disturbances. Code and m
odels will be publicly available.
****
Vikranth Srivatsa,Zijian He,Reyna Abhyankar,Dongming Li,Yiying Zhang
Preble: Efficient Distributed Prompt Scheduling for LLM Serving
Prompts to large language models (LLMs) have evolved beyond simple user question
s.
For LLMs to solve complex problems, today’s practices are to include domain-spec
ific
instructions, illustration of tool usages, and/or long context such as textbook
chapters in
prompts. As such, many parts of prompts are repetitive across requests. Recent w
orks
propose to cache and reuse KV state of prompts. However, they are all confined t
o a single-
GPU optimization, while production LLM serving systems are distributed by nature
.

This paper proposes Preble, the first distributed LLM serving platform that targ
ets and op-
timizes for prompt sharing. We designed a distributed scheduling system that co-
optimizes
KV state reuse and computation load-balancing with a new scheduling algorithm an
d a
hierarchical scheduling mechanism. Our evaluation of Preble with real workloads
and re-
quest arrival patterns on two open-source LLMs shows that Preble outperforms the
 SOTA
serving systems by 1.5× to 14.5× on average latency and 2× to 10× on p99 latency
.
****
Bo Xiong,Steffen Staab
From Tokens to Lattices: Emergent Lattice Structures in Language Models
Pretrained masked language models (MLMs) have demonstrated an impressive capabil
ity to comprehend and encode conceptual knowledge, revealing a lattice structure
 among concepts. This raises a critical question: how does this conceptualizatio

n emerge from MLM pretraining? In this paper, we explore this problem from the p
erspective of Formal Concept Analysis (FCA), a mathematical framework that deriv
es concept lattices from the observations of object-attribute relationships. We
show that the MLM's objective implicitly learns a formal context that describes
objects, attributes, and their dependencies, which enables the reconstruction of
 a concept lattice through FCA. We propose a novel framework for concept lattice
 construction from pretrained MLMs and investigate the origin of the inductive b
iases of MLMs in lattice structure learning. Our framework differs from previous
 work because it does not rely on human-defined concepts and allows for discover
ing "latent" concepts that extend beyond human definitions. We create three data
sets for evaluation, and the empirical results verify our hypothesis.
****
Shivam Duggal,Phillip Isola,Antonio Torralba,William T. Freeman
Adaptive Length Image Tokenization via Recurrent Allocation
Current vision systems typically assign fixed-length representations to images,
regardless of the information content. This contrasts with human intelligence —a
nd even large language models—which allocate varying representational capacities
 based on entropy, context and familiarity. Inspired by this, we propose an appr
oach to learn variable-length token representations for 2D images. Our encoder-d
ecoder architecture recursively processes 2D image tokens, distilling them into
1D latent tokens over multiple iterations of recurrent rollouts. Each iteration
refines the 2D tokens, updates the existing 1D latent tokens, and adaptively inc
reases representational capacity by adding new tokens. This enables compression
of images into a variable number of tokens, ranging from 32 to 256. We validate
our tokenizer using reconstruction loss and FID metrics, demonstrating that toke
n count aligns with image entropy, familiarity and downstream task requirements.
 Recurrent token processing with increasing representational capacity in each it
eration shows signs of token specialization, revealing potential for object / pa
rt discovery.
****
Mude Hui,Siwei Yang,Bingchen Zhao,Yichun Shi,Heng Wang,Peng Wang,Cihang Xie,Yuyi
n Zhou
HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing
This study introduces HQ-Edit, a high-quality instruction-based image editing da
taset with around 200,000 edits. Unlike prior approaches relying on attribute gu
idance or human feedback on building datasets, we devise a scalable data collect
ion pipeline leveraging advanced foundation models, namely GPT-4V and DALL-E 3.
To ensure its high quality, diverse examples are first collected online, expande
d, and then used to create high-quality diptychs featuring input and output imag
es with detailed text prompts, followed by precise alignment ensured through pos
t-processing. In addition, we propose two evaluation metrics, Alignment and Cohe
rence, to quantitatively assess the quality of image edit pairs using GPT-4V. HQ
-Edits high-resolution images, rich in detail and accompanied by comprehensive e
diting prompts, substantially enhance the capabilities of existing image editing
 models. For example, an HQ-Edit finetuned InstructPix2Pix can attain state-of-t
he-art image editing performance, even surpassing those models fine-tuned with h
uman-annotated data.
****
Yuzhou Chen,Yulia Gel
Topological Zigzag Spaghetti for Diffusion-based Generation and Prediction on Gr
aphs
Diffusion models have recently emerged as a new powerful machinery for generativ
e artificial intelligence on graphs, with applications ranging from drug design
to knowledge discovery. However, despite their high potential, most, if not all,
 existing graph diffusion models are limited in their ability to holistically de
scribe the intrinsic higher-order topological graph properties, which obstructs
model generalizability and adoption for downstream tasks. We address this fundam
ental challenge and extract the latent salient topological graph descriptors at
different resolutions by leveraging zigzag persistence. We develop a new computa
tionally efficient topological summary,

zigzag spaghetti (ZS), which delivers the most inherent topological properties s
imultaneously over a sequence of graphs at multiple resolutions. We derive theor
etical stability guarantees of ZS and present the first attempt to integrate
dynamic topological information into graph diffusion models. Our extensive exper
iments on graph classification and prediction tasks suggest that ZS has a high p
romise not only to enhance performance of graph diffusion models, with gains up
10\%, but also to substantially booster model robustness.
****
Alexandru Dumitrescu,Dani Korpela,Markus Heinonen,Yogesh Verma,Valerii Iakovlev,
Vikas Garg,Harri Lähdesmäki
E(3)-equivariant models cannot learn chirality: Field-based molecular generation
Obtaining the desired effect of drugs is highly dependent on their molecular geo
metries. Thus, the current prevailing paradigm focuses on 3D point-cloud atom re
presentations, utilizing graph neural network (GNN) parametrizations, with rotat
ional symmetries baked in via E(3) invariant layers. We prove that such models m
ust necessarily disregard chirality, a geometric property of the molecules that
cannot be superimposed on their mirror image by rotation and translation. Chiral
ity plays a key role in determining drug safety and potency. To address this gla
ring issue, we introduce a novel field-based representation, proposing reference
 rotations that replace rotational symmetry constraints. The proposed model capt
ures all molecular geometries including chirality, while still achieving highly
competitive performance with E(3)-based methods across standard benchmarking met
rics.
****
Rikiya Takehi,Masahiro Asami,Kosuke Kawakami,Yuta Saito
A General Framework for Off-Policy Learning with Partially-Observed Reward
Off-policy learning (OPL) in contextual bandits aims to learn a decision-making
policy that maximizes the target rewards by using only historical interaction da
ta collected under previously developed policies. Unfortunately, when rewards ar
e only partially observed, the effectiveness of OPL degrades severely. Well-know
n examples of such partial rewards include explicit ratings in content recommend
ations, conversion signals on e-commerce platforms that are partial due to delay
, and the issue of censoring in medical problems. One possible solution to deal
with such partial rewards is to use secondary rewards, such as dwelling time, cl
icks, and medical indicators, which are more densely observed. However, relying
solely on such secondary rewards can also lead to poor policy learning since the
y may not align with the target reward. Thus, this work studies a new and genera
l problem of OPL where the goal is to learn a policy that maximizes the expected
 target reward by leveraging densely observed secondary rewards as supplemental
data. We then propose a new method called Hybrid Policy Optimization for Partial
ly-Observed Reward (HyPeR), which effectively uses the secondary rewards in addi
tion to the partially observed target reward to achieve effective OPL despite th
e challenging scenario. We also discuss a case where we aim to optimize not only
 the expected target reward but also the expected secondary rewards to some exte
nt; counter-intuitively, we will show that leveraging the two objectives is in f
act advantageous also for the optimization of only the target reward. Along with
 statistical analysis of our proposed methods, empirical evaluations on both syn
thetic and real-world data show that HyPeR outperforms existing methods in vario
us scenarios.
****
Yingwen Wu,Ruiji Yu,Xinwen Cheng,Zhengbao He,Xiaolin Huang
Pursuing Feature Separation based on Neural Collapse for Out-of-Distribution Det
ection
In the open world, detecting out-of-distribution (OOD) data, whose labels are di
sjoint with those of in-distribution (ID) samples, is important for reliable dee
p neural networks (DNNs). To achieve better detection performance, one type of a
pproach proposes to fine-tune the model with auxiliary OOD datasets to amplify t
he difference between ID and OOD data through a separation loss defined on model
 outputs. However, none of these studies consider enlarging the feature disparit
y, which should be more effective compared to outputs. The main difficulty lies

in the diversity of OOD samples, which makes it hard to describe their feature d
istribution, let alone design losses to separate them from ID features. In this
paper, we neatly fence off the problem based on an aggregation property of ID fe
atures named Neural Collapse (NC). NC means that the penultimate features of ID
samples within a class are nearly identical to the last layer weight of the corr
esponding class. Based on this property, we propose a simple but effective loss
called Separation Loss, which binds the features of OOD data in a subspace ortho
gonal to the principal subspace of ID features formed by NC. In this way, the fe
atures of ID and OOD samples are separated by different dimensions. By optimizin
g the feature separation loss rather than purely enlarging output differences, o
ur detection achieves SOTA performance on CIFAR10, CIFAR100 and ImageNet benchma
rks without any additional data augmentation or sampling, demonstrating the impo
rtance of feature separation in OOD detection. Code is available
at https://github.com/Wuyingwen/Pursuing-Feature-Separation-for-OOD-Detection.
****

Yancheng Zhang,Jiaqi Xue,Mengxin Zheng,Mimi Xie,Mingzhe Zhang,Lei Jiang,Qian Lou
CipherPrune: Efficient and Scalable Private Transformer Inference

Private Transformer inference using cryptographic protocols offers promising sol
utions for privacy-preserving machine learning; however, it still faces signific
ant runtime overhead (efficiency issues) and challenges in handling long-token i
nputs (scalability issues). We observe that the Transformer's operational comple
xity scales quadratically with the number of input tokens, making it essential t
o reduce the input token length. Notably, each token varies in importance, and m
any inputs contain redundant tokens. Additionally, prior private inference metho
ds that rely on high-degree polynomial approximations for non-linear activations
 are computationally expensive. Therefore, reducing the polynomial degree for le
ss important tokens can significantly accelerate private inference. Building on
 these observations, we propose \textit{CipherPrune}, an efficient and scalable
private inference framework that includes a secure encrypted token pruning proto
col, a polynomial reduction protocol, and corresponding Transformer network opti
mizations. At the protocol level, encrypted token pruning adaptively removes uni
mportant tokens from encrypted inputs in a progressive, layer-wise manner. Addit
ionally, encrypted polynomial reduction assigns lower-degree polynomials to less
 important tokens after pruning, enhancing efficiency without decryption. At the
 network level, we introduce protocol-aware network optimization via a gradient-
based search to maximize pruning thresholds and polynomial reduction conditions
while maintaining the desired accuracy. Our experiments demonstrate that CipherP
rune reduces the execution overhead of private Transformer inference by approxim
ately $6.1\times$ for 128-token inputs and $10.6\times$ for 512-token inputs, c
ompared to previous methods, with only a marginal drop in accuracy. The code is
publicly available at https://github.com/UCF-Lou-Lab-PET/cipher-prune-inference.
****

Shubham Toshniwal,Wei Du,Ivan Moshkov,Branislav Kisacanin,Alexan Ayrapetyan,Igor
 Gitman
OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instructio
n Data

Mathematical reasoning continues to be a critical challenge in large language mo
del (LLM) development with significant interest. However, most of the cutting-ed
ge progress in mathematical reasoning with LLMs has become closed-source due to
lack of access to training data. This lack of data access limits researchers fro
m understanding the impact of different choices for synthesizing and utilizing t
he data. With the goal of creating a high-quality finetuning (SFT) dataset for m
ath reasoning, we conduct careful ablation experiments on data synthesis using t
he recently released Llama3.1 family of models. Our experiments show that: (a) s
olution format matters, with excessively verbose solutions proving detrimental t
o SFT performance, (b) data generated by a strong teacher outperforms on-policy
data generated by a weak student model, (c) SFT is robust to low-quality solutio
ns, allowing for imprecise data filtering, and (d) question diversity is crucial
 for achieving data scaling gains. Based on these insights, we create the OpenMa
thInstruct-2 dataset which consists of 14M question-solution pairs (≈ 600K uniqu

e questions), making it nearly eight times larger than the previous largest open -source math reasoning dataset. Finetuning the Llama-3.1-8B-Base using OpenMathI nstruct-2 outperforms Llama3.1-8B-Instruct on MATH by an absolute 15.9% (51.9% → 67.8%). Finally, to accelerate the open-source efforts, we release the code, th e finetuned models, and the OpenMathInstruct-2 dataset under a commercially perm issive license.
****
Xufeng Cai,Jelena Diakonikolas
Last Iterate Convergence of Incremental Methods as a Model of Forgetting
Incremental gradient and incremental proximal methods are a fundamental class of optimization algorithms used for solving finite sum problems, broadly studied i n the literature. Yet, without strong convexity, their convergence guarantees ha ve primarily been established for the ergodic (average) iterate. We establish th e first nonasymptotic convergence guarantees for the last iterate of both increm ental gradient and incremental proximal methods, in general convex smooth (for b oth) and convex Lipschitz (for the proximal variants) settings. Our oracle compl exity bounds for the last iterate nearly match (i.e., match up to a square-root-log or a log factor) the best known oracle complexity bounds for the average ite rate, for both classes of methods. We further obtain generalizations of our resu lts to weighted averaging of the iterates with increasing weights and for random ly permuted ordering of updates. We study last iterate convergence of the increm ental proximal method as a mathematical abstraction of forgetting in continual l earning and prove a lower bound that certifies that a large amount of regulariza tion is crucial to mitigating catastrophic forgetting---one of the key considera tions in continual learning. Our results generalize last iterate guarantees for incremental methods compared to state of the art, as such results were previousl y known only for overparameterized linear models, which correspond to convex qua dratic problems with infinitely many solutions.
****
Donghoon Kim,Minji Bae,Kyuhong Shim,Byonghyo Shim
Visually Guided Decoding: Gradient-Free Hard Prompt Inversion with Language Mode ls
Text-to-image generative models like DALL-E and Stable Diffusion have revolution ized visual content creation across various applications, including advertising, personalized media, and design prototyping.
However, crafting effective textual prompts to guide these models remains challe nging, often requiring extensive trial and error.
Existing prompt inversion approaches, such as soft and hard prompt techniques, a re not so effective due to the limited interpretability and incoherent prompt ge neration.
To address these issues, we propose Visually Guided Decoding (VGD), a gradient-f ree approach that leverages large language models (LLMs) and CLIP-based guidance to generate coherent and semantically aligned prompts.
In essence, VGD utilizes the robust text generation capabilities of LLMs to prod uce human-readable prompts.
Further, by employing CLIP scores to ensure alignment with user-specified visual concepts, VGD enhances the interpretability, generalization, and flexibility of prompt generation without the need for additional training.
Our experiments demonstrate that VGD outperforms existing prompt inversion techn iques in generating understandable and contextually relevant prompts, facilitati ng more intuitive and controllable interactions with text-to-image models.
****
Yu Shang,Yu Li,Keyu Zhao,Likai Ma,Jiahe Liu,Fengli Xu,Yong Li
AgentSquare: Automatic LLM Agent Search in Modular Design Space
Recent advancements in Large Language Models (LLMs) have led to a rapid growth o f agentic systems capable of handling a wide range of complex tasks. However, cu rrent research largely relies on manual, task-specific design, limiting their ad aptability to novel tasks. In this paper, we introduce a new research problem: M odularized LLM Agent Search (MoLAS). We propose a modular design space that abst racts existing LLM agent designs into four fundamental modules with uniform IO i

nterface: Planning, Reasoning, Tool Use, and Memory. Building on this design spa
ce, we present a novel LLM agent search framework called AgentSquare, which intr
oduces two core mechanisms, i.e., module evolution and recombination, to efficie
ntly search for optimized LLM agents. To further accelerate the process, we desi
gn a performance predictor that uses in-context surrogate models to skip unpromi
sing agent designs. Extensive experiments across six benchmarks, covering the di
verse scenarios of web, embodied, tool use and game applications, show that Agen
tSquare substantially outperforms hand-crafted agents, achieving an average perf
ormance gain of 17.2% against best-known human designs. Moreover, AgentSquare ca
n generate interpretable design insights, enabling a deeper understanding of age
ntic architecture and its impact on task performance. We believe that the modula
r design space and AgentSquare search framework offer a platform for fully explo
iting the potential of prior successful designs and consolidate the collective e
fforts of research community. Code repo is available at https://github.com/tsing
hua-fib-lab/AgentSquare.
****

Changkun Liu,Shuai Chen,Yash Sanjay Bhalgat,Siyan HU,Ming Cheng,Zirui Wang,Victo
r Adrian Prisacariu,Tristan Braud

GS-CPR: Efficient Camera Pose Refinement via 3D Gaussian Splatting

We leverage 3D Gaussian Splatting (3DGS) as a scene representation and propose a
 novel test-time camera pose refinement (CPR) framework, GS-CPR. This framework
enhances the localization accuracy of state-of-the-art absolute pose regression
and scene coordinate regression methods. The 3DGS model renders high-quality syn
thetic images and depth maps to facilitate the establishment of 2D-3D correspond
ences. GS-CPR obviates the need for training feature extractors or descriptors b
y operating directly on RGB images, utilizing the 3D foundation model, MASt3R, f
or precise 2D matching. To improve the robustness of our model in challenging ou
tdoor environments, we incorporate an exposure-adaptive module within the 3DGS f
ramework. Consequently, GS-CPR enables efficient one-shot pose refinement given
a single RGB query and a coarse initial pose estimation. Our proposed approach s
urpasses leading NeRF-based optimization methods in both accuracy and runtime ac
ross indoor and outdoor visual localization benchmarks, achieving new state-of-t
he-art accuracy on two indoor datasets.
****

Haitao Lin,Guojiang Zhao,Odin Zhang,Yufei Huang,Lirong Wu,Cheng Tan,Zicheng Liu,
Zhifeng Gao,Stan Z. Li

CBGBench: Fill in the Blank of Protein-Molecule Complex Binding Graph

Structure-based drug design (SBDD) aims to generate potential drugs that can bin
d to a target protein and is greatly expedited by the aid of AI techniques in ge
nerative models. However, a lack of systematic understanding persists due to the
 diverse settings, complex implementation, difficult reproducibility, and task s
ingularity. Firstly, the absence of standardization can lead to unfair compariso
ns and inconclusive insights. To address this dilemma, we propose CBGBench, a co
mprehensive benchmark for SBDD, that unifies the task as a generative graph comp
letion, analogous to fill-in-the-blank of the 3D complex binding graph. By categ
orizing existing methods based on their attributes, CBGBench facilitates a modul
ar and extensible framework that implements cutting-edge methods. Secondly, a si
ngle de novo molecule generation task can hardly reflect their capabilities. To
broaden the scope, we adapt these models to a range of tasks essential in drug d
esign, considered sub-tasks within the graph fill-in-the-blank tasks. These task
s include the generative designation of de novo molecules, linkers, fragments, s
caffolds, and sidechains, all conditioned on the structures of protein pockets.
Our evaluations are conducted with fairness, encompassing comprehensive perspect
ives on interaction, chemical properties, geometry authenticity, and substructur
e validity. We further provide insights with analysis from empirical studies. Ou
r results indicate that there is potential for further improvements on many task
s, with optimization in network architectures, and effective incorporation of ch
emical prior knowledge. Finally, to lower the barrier to entry and facilitate fu
rther developments in the field, we also provide a single [codebase](https://git
hub.com/EDAPINENUT/CBGBench) that unifies the discussed models, data pre-process

ing, training, sampling, and evaluation.
****
Xinyu Ma,Yifeng Xu,Yang Lin,Tianlong Wang,Xu Chu,Xin Gao,Junfeng Zhao,Yasha Wang
DRESSing Up LLM: Efficient Stylized Question-Answering via Style Subspace Editing
We introduce DRESS, a novel approach for generating stylized large language model (LLM) responses through representation editing. Existing methods like prompting and fine-tuning are either insufficient for complex style adaptation or computationally expensive, particularly in tasks like NPC creation or character role-playing. Our approach leverages the over-parameterized nature of LLMs to disentangle a style-relevant subspace within the model's representation space to conduct representation editing, ensuring a minimal impact on the original semantics. By applying adaptive editing strengths, we dynamically adjust the steering vectors in the style subspace to maintain both stylistic fidelity and semantic integrity. We develop two stylized QA benchmark datasets to validate the effectiveness of DRESS, and the results demonstrate significant improvements compared to baseline methods such as prompting and ITI. In short, DRESS is a lightweight, train-free solution for enhancing LLMs with flexible and effective style control, making it particularly useful for developing stylized conversational agents. Codes and benchmark datasets are available at https://github.com/ArthurLeoM/DRESS-LLM.
****
Cong Fu,Xiner Li,Blake Olson,Heng Ji,Shuiwang Ji
Fragment and Geometry Aware Tokenization of Molecules for Structure-Based Drug Design Using Language Models
Structure-based drug design (SBDD) is crucial for developing specific and effective therapeutics against protein targets but remains challenging due to complex protein-ligand interactions and vast chemical space. Although language models (LMs) have excelled in natural language processing, their application in SBDD is underexplored. To bridge this gap, we introduce a method, known as Frag2Seq, to apply LMs to SBDD by generating molecules in a fragment-based manner in which fragments correspond to functional modules. We transform 3D molecules into fragment-informed sequences using $SE(3)$-equivariant molecule and fragment local frames, extracting $SE(3)$-invariant sequences that preserve geometric information of 3D fragments. Furthermore, we incorporate protein pocket embeddings obtained from a pre-trained inverse folding model into the LMs via cross-attention to capture protein-ligand interaction, enabling effective target-aware molecule generation. Benefiting from employing LMs with fragment-based generation and effective protein context encoding, our model achieves the best performance on binding vina score and chemical properties such as QED and Lipinski, which shows our model’s efficacy in generating drug-like ligands with higher binding affinity against target proteins. Moreover, our method also exhibits higher sampling efficiency compared to atom-based autoregressive and diffusion baselines with at most $\times 300$ speedup. The code will be made publicly available at https://github.com/divelab/AIRS/tree/main/OpenMI/Frag2Seq.
****
Haipeng Luo,Qingfeng Sun,Can Xu,Pu Zhao,Jian-Guang Lou,Chongyang Tao,Xiubo Geng,Qingwei Lin,Shifeng Chen,Yansong Tang,Dongmei Zhang
WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct
Large language models (LLMs), such as GPT-4, have shown remarkable performance in natural language processing (NLP) tasks, including challenging mathematical reasoning. However, most existing open-source models are only pre-trained on large-scale internet data and without math-related optimization. In this paper, we present WizardMath, which enhances the mathematical reasoning abilities of LLMs, by applying our proposed Reinforcement Learning from Evol-Instruct Feedback (RLEIF) method to the domain of math. Through extensive experiments on two mathematical reasoning benchmarks, namely GSM8k and MATH, we reveal the extraordinary capabilities of our model. Remarkably, WizardMath-Mistral 7B surpasses all other open-source LLMs by a substantial margin. Furthermore, WizardMath 70B even outperforms ChatGPT-3.5, Claude Instant, Gemini Pro and Mistral Medium. Additionally, ou

r preliminary exploration highlights the pivotal role of instruction evolution a
nd process supervision in achieving exceptional math performance.
****

Ye Yuan,Can Chen,Christopher Pal,Xue Liu

ParetoFlow: Guided Flows in Multi-Objective Optimization

In offline multi-objective optimization (MOO), we leverage an offline dataset of
 designs and their associated labels to simultaneously minimize multiple objecti
ves. This setting more closely mirrors complex real-world problems compared to s
ingle-objective optimization. Recent works mainly employ evolutionary algorithms
 and Bayesian optimization, with limited attention given to the generative model
ing capabilities inherent in such data. In this study, we explore generative mod
eling in offline MOO through flow matching, noted for its effectiveness and effi
ciency. We introduce \textit{ParetoFlow}, specifically designed to guide flow sa
mpling to approximate the Pareto front. Traditional predictor~(classifier) guida
nce is inadequate for this purpose because it models only a single objective. In
 response, we propose a \textit{multi-objective predictor guidance} module that
assigns each sample a weight vector, representing a weighted distribution across
 multiple objective predictions. A local filtering scheme is introduced to addre
ss non-convex Pareto fronts. These weights uniformly cover the entire objective
space, effectively directing sample generation towards the Pareto front. Since d
istributions with similar weights tend to generate similar samples, we introduce
 a \textit{neighboring evolution} module to foster knowledge sharing among neigh
boring distributions. This module generates offspring from these distributions,
and selects the most promising one for the next iteration. Our method achieves s
tate-of-the-art performance across various tasks. Our code is available.
****

Lecheng Kong,Jiarui Feng,Hao Liu,Chengsong Huang,Jiaxin Huang,Yixin Chen,Muhan Z
hang

GOFA: A Generative One-For-All Model for Joint Graph Language Modeling

Foundation models, such as Large Language Models (LLMs) or Large Vision Models (
LVMs), have emerged as one of the most powerful tools in the respective fields.
However, unlike text and image data, graph data do not have a definitive structu
re, posing great challenges to developing a Graph Foundation Model (GFM). For ex
ample, current attempts at designing general graph models either transform graph
 data into a language format for LLM-based prediction or still train a GNN model
 with LLM as an assistant. The former can handle unlimited tasks, while the latt
er captures graph structure much better---yet, no existing work can achieve both
 simultaneously. In this paper, we first identify three key desirable properties
 of a GFM: self-supervised pretraining, fluidity in tasks, and graph awareness.
To account for these properties, we extend the conventional language modeling to
 the graph domain and propose a novel generative graph language model GOFA. The
model interleaves randomly initialized GNN layers into a frozen pre-trained LLM
so that the semantic and structural modeling abilities are organically combined.
 GOFA is pre-trained on newly proposed graph-level next-word prediction, questio
n-answering, structural understanding, and information retrieval tasks to obtain
 the above GFM properties. The pre-trained model is further instruction fine-tun
ed to obtain the task-solving ability. Our GOFA model is evaluated on various do
wnstream datasets unseen during the pre-training and fine-tuning phases, demonst
rating a strong ability to solve structural and contextual problems in zero-shot
 scenarios. The code is available at https://github.com/JiaruiFeng/GOFA.
****

Haoyang Liu,Jie Wang,Zijie Geng,Xijun Li,Yuxuan Zong,Fangzhou Zhu,Jianye HAO,Fen
g Wu

Apollo-MILP: An Alternating Prediction-Correction Neural Solving Framework for M
ixed-Integer Linear Programming

Leveraging machine learning (ML) to predict an initial solution for mixed-intege
r linear programming (MILP) has gained considerable popularity in recent years.
These methods predict a solution and fix a subset of variables to reduce the pro
blem dimension. Then, they solve the reduced problem to obtain the final solutio
ns. However, directly fixing variable values can lead to low-quality solutions o

r even infeasible reduced problems if the predicted solution is not accurate eno
ugh. To address this challenge, we propose an Alternating prediction-correction
neural solving framework (Apollo-MILP) that can identify and select accurate and
 reliable predicted values to fix. In each iteration, Apollo-MILP conducts a pre
diction step for the unfixed variables, followed by a correction step to obtain
an improved solution (called reference solution) through a trust-region search.
By incorporating the predicted and reference solutions, we introduce a novel Unc
ertainty-based Error upper BOund (UEBO) to evaluate the uncertainty of the predi
cted values and fix those with high confidence. A notable feature of Apollo-MILP
 is the superior ability for problem reduction while preserving optimality, lead
ing to high-quality final solutions. Experiments on commonly used benchmarks dem
onstrate that our proposed Apollo-MILP significantly outperforms other ML-based
approaches in terms of solution quality, achieving over a 50% reduction in the s
olution gap.
****
Allen Z. Ren,Justin Lidard,Lars Lien Ankile,Anthony Simeonov,Pulkit Agrawal,Anir
udha Majumdar,Benjamin Burchfiel,Hongkai Dai,Max Simchowitz
Diffusion Policy Policy Optimization
We introduce Diffusion Policy Policy Optimization, DPPO, an algorithmic framewor
k including best practices for fine-tuning diffusion-based policies (e.g. Diffus
ion Policy) in continuous control and robot learning tasks using the policy grad
ient (PG) method from reinforcement learning (RL). PG methods are ubiquitous in
training RL policies with other policy parameterizations; nevertheless, they had
 been conjectured to be less efficient for diffusion-based policies. Surprisingl
y, we show that DPPO achieves the strongest overall performance and efficiency f
or fine-tuning in common benchmarks compared to other RL methods for diffusion-b
ased policies and also compared to PG fine-tuning of other policy parameterizati
ons. Through experimental investigation, we find that DPPO takes advantage of un
ique synergies between RL fine-tuning and the diffusion parameterization, leadin
g to structured and on-manifold exploration, stable training, and strong policy
robustness. We further demonstrate the strengths of DPPO in a range of realistic
 settings, including simulated robotic tasks with pixel observations, and via ze
ro-shot deployment of simulation-trained policies on robot hardware in a long-ho
rizon, multi-stage manipulation task.
****
Yepeng Liu,Yiren Song,Hai Ci,Yu Zhang,Haofan Wang,Mike Zheng Shou,Yuheng Bu
Image Watermarks are Removable using Controllable Regeneration from Clean Noise
Image watermark techniques provide an effective way to assert ownership, deter m
isuse, and trace content sources, which has become increasingly essential in the
 era of large generative models. A critical attribute of watermark techniques is
 their robustness against various manipulations. In this paper, we introduce a w
atermark removal approach capable of effectively nullifying state-of-the-art wat
ermarking techniques. Our primary insight involves regenerating the watermarked
image starting from a \textbf{clean Gaussian noise} via a controllable diffusion
 model, utilizing the extracted semantic and spatial features from the watermark
ed image. The semantic control adapter and the spatial control network are speci
fically trained to control the denoising process towards ensuring image quality
and enhancing consistency between the cleaned image and the original watermarked
 image. To achieve a smooth trade-off between watermark removal performance and
image consistency, we further propose an adjustable and controllable regeneratio
n scheme. This scheme adds varying numbers of noise steps to the latent represen
tation of the watermarked image, followed by a controlled denoising process star
ting from this noisy latent representation. As the number of noise steps increas
es, the latent representation progressively approaches clean Gaussian noise, fac
ilitating the desired trade-off. We apply our watermark removal methods across v
arious watermarking techniques, and the results demonstrate that our methods off
er superior visual consistency/quality and enhanced watermark removal performanc
e compared to existing regeneration approaches. Our code is available at \url{ht
tps://github.com/yepengliu/CtrlRegen}.
****

Ruilin Tong,Yuhang Liu,Javen Qinfeng Shi,Dong Gong
Coreset Selection via Reducible Loss in Continual Learning
Rehearsal-based continual learning (CL) aims to mitigate catastrophic forgetting by maintaining a subset of samples from previous tasks and replaying them. The rehearsal memory can be naturally constructed as a coreset, designed to form a compact subset that enables training with performance comparable to using the full dataset. The coreset selection task can be formulated as bilevel optimization that solves for the subset to minimize the outer objective of the learning task. Existing methods primarily rely on inefficient probabilistic sampling or local gradient-based scoring to approximate sample importance through an iterative process that can be susceptible to ambiguity or noise. Specifically, non-representative samples like ambiguous or noisy samples are difficult to learn and incur high loss values even when training on the full dataset. However, existing methods relying on local gradient tend to highlight these samples in an attempt to minimize the outer loss, leading to a suboptimal coreset. To enhance coreset selection, especially in CL where high-quality samples are essential, we propose a coreset selection method that measures sample importance using reducible loss (ReL) that quantifies the impact of adding a sample to model performance. By leveraging ReL and a process derived from bilevel optimization, we identify and retain samples that yield the highest performance gain. They are shown to be informative and representative. Furthermore, ReL requires only forward computation, making it significantly more efficient than previous methods. To better apply coreset selection in CL, we extend our method to address key challenges such as task interference, streaming data, and knowledge distillation. Experiments on data summarization and continual learning demonstrate the effectiveness and efficiency of our approach.
****
Ivan Butakov,Alexander Semenenko,Alexander Tolmachev,Andrey Gladkov,Marina Munkhoeva,Alexey Frolov
Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax
Deep InfoMax (DIM) is a well-established method for self-supervised representation learning (SSRL) based on maximization of the mutual information between the input and the output of a deep neural network encoder. Despite the DIM and contrastive SSRL in general being well-explored, the task of learning representations conforming to a specific distribution (i.e., distribution matching, DM) is still under-addressed. Motivated by the importance of DM to several downstream tasks (including generative modeling, disentanglement, outliers detection and other), we enhance DIM to enable automatic matching of learned representations to a selected prior distribution. To achieve this, we propose injecting an independent noise into the normalized outputs of the encoder, while keeping the same InfoMax training objective. We show that such modification allows for learning uniformly and normally distributed representations, as well as representations of other absolutely continuous distributions. Our approach is tested on various downstream tasks. The results indicate a moderate trade-off between the performance on the downstream tasks and quality of DM.
****
Shaochen Zhong,Yifan Lu,Lize Shao,Bhargav Bhushanam,Xiaocong Du,Yixin Wan,Yuchen g Shi,Daochen Zha,Yiwei Wang,Ninghao Liu,Kaixiong Zhou,Shuai Xu,Kai-Wei Chang,Louis Feng,Vipin Chaudhary,Xia Hu
MQuAKE-Remastered: Multi-Hop Knowledge Editing Can Only Be Advanced with Reliable Evaluations
Large language models (LLMs) can give out erroneous answers to factually rooted questions either as a result of undesired training outcomes or simply because the world has moved on after a certain knowledge cutoff date. Under such scenarios, *knowledge editing* often comes to the rescue by delivering efficient patches for such erroneous answers without significantly altering the rest, where many editing methods have seen reasonable success when the editing targets are simple and direct (e.g., *``what club does Lionel Messi currently play for?''*).

However, knowledge fragments like this are often deeply intertwined in the real world, making effectively propagating the editing effect to non-directly related questions a practical challenge (to entertain an extreme example: [*"What car did the wife of the owner of the club that Messi currently plays for used to get to school in the 80s?"*](youtube.com/watch?v=DbwiHC1Fu-E\&t=132s)). Prior arts have coined this task as *multi-hop knowledge editing* with the most popular dataset being MQuAKE, serving as the sole evaluation benchmark for many later proposed editing methods due to the expensive nature of constructing knowledge editing datasets at scale.

In this work, we reveal that **up to 33\% or 76\% of \mquake{}'s questions and ground truth labels are, in fact, corrupted in various fashions due to some unintentional clerical or procedural oversights**. Our work provides a detailed audit of MQuAKE's error pattern and a comprehensive fix without sacrificing its dataset capacity. Additionally, we benchmarked almost all proposed MQuAKE-evaluated editing methods on our post-fix dataset, **MQuAKE-Remastered**. We observe that many methods try to overfit the original MQuAKE by exploiting some dataset idiosyncrasies of MQuAKE. We provide a guideline on how to approach such datasets faithfully and show that a simple, minimally invasive approach — **GWalk** — can offer beyond SOTA editing performance without such exploitation. The MQuAKE-Remastered datasets and utilities are available at huggingface.co/datasets/henryzhongsc/MQuAKE-Remastered and github.com/henryzhongsc/MQuAKE-Remastered, respectively.
****
Keda TAO,Jinjin Gu,Yulun Zhang,Xiucheng Wang,Nan Cheng
Overcoming False Illusions in Real-World Face Restoration with Multi-Modal Guided Diffusion Model
We introduce a novel Multi-modal Guided Real-World Face Restoration (MGFR) technique designed to improve the quality of facial image restoration from low-quality inputs. Leveraging a blend of attribute text prompts, high-quality reference images, and identity information, MGFR can mitigate the generation of false facial attributes and identities often associated with generative face restoration methods. By incorporating a dual-control adapter and a two-stage training strategy, our method effectively utilizes multi-modal prior information for targeted restoration tasks. We also present the Reface-HQ dataset, comprising over 23,000 high-resolution facial images across 5,000 identities, to address the need for reference face training images. Our approach achieves superior visual quality in restoring facial details under severe degradation and allows for controlled restoration processes, enhancing the accuracy of identity preservation and attribute correction. Including negative quality samples and attribute prompts in the training further refines the model's ability to generate detailed and perceptually accurate images.
****
Ming Liu,Wensheng Zhang
Is Your Video Language Model a Reliable Judge?
As video language models (VLMs) gain more applications in various scenarios, the need for robust and scalable evaluation of their performance becomes increasingly critical. The traditional human expert-based evaluation of VLMs has limitations in consistency and scalability, which sparked interest in automatic methods such as employing VLMs to evaluate VLMs. However, the reliability of VLMs as judges remains underexplored. Existing methods often rely on a single VLM as the evaluator. However, this approach can be unreliable or biased because such a model may lack the ability to fully understand the content and may have inherent biases, ultimately compromising evaluation reliability. A remedy is to apply the principle of collective thoughts, aggregating evaluations from multiple VLMs to enhance reliability. This study investigates the efficacy of such approaches, particularly when the pool of judges includes both reliable and unreliable models. Our

findings reveal that incorporating collective judgments from such a mixed pool
does not necessarily improve the accuracy of the final evaluation. The inclusion of
less reliable judges can introduce noise, undermining the overall reliability of the
outcomes. To explore the factors that impact evaluation reliability, we fine-tune
an underperforming VLM judge, Video-LLaVA, and observe that improved understanding ability alone is insufficient to make VLM judges more reliable. These
findings stress the limitations of collective thought approaches and highlight the
need for more advanced methods that can account for the reliability of individual
models. Our study promotes the development of more reliable evaluation methods for VLMs
****
Jianhong Bai,Menghan Xia,Xintao Wang,Ziyang Yuan,Zuozhu Liu,Haoji Hu,Pengfei Wan,Di ZHANG
SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints
Recent advancements in video diffusion models demonstrate remarkable capabilities in simulating real-world dynamics and 3D consistency. This progress motivates us to explore the potential of these models to maintain dynamic consistency across diverse viewpoints, a feature highly sought after in applications like virtual filming. Unlike existing methods focused on multi-view generation of single objects for 4D reconstruction, our interest lies in generating open-world videos from arbitrary viewpoints, incorporating six degrees of freedom (6 DoF) camera poses.
To achieve this, we propose a plug-and-play module that enhances a pre-trained text-to-video model for multi-camera video generation, ensuring consistent content across different viewpoints. Specifically, we introduce a multi-view synchronization module designed to maintain appearance and geometry consistency across these viewpoints. Given the scarcity of high-quality training data, we also propose a progressive training scheme that leverages multi-camera images and monocular videos as a supplement to Unreal Engine-rendered multi-camera videos. This comprehensive approach significantly benefits our model.
Experimental results demonstrate the superiority of our proposed method over existing competitors and several baselines. Furthermore, our method enables intriguing extensions, such as re-rendering a video from multiple novel viewpoints. Project webpage: https://jianhongbai.github.io/SynCamMaster/
****
Joonsung Jeon,Woo Jae Kim,Suhyeon Ha,Sooel Son,Sung-eui Yoon
AdvPaint: Protecting Images from Inpainting Manipulation via Adversarial Attention Disruption
The outstanding capability of diffusion models in generating high-quality images poses significant threats when misused by adversaries. In particular, we assume malicious adversaries exploiting diffusion models for inpainting tasks, such as replacing a specific region with a celebrity. While existing methods for protecting images from manipulation in diffusion-based generative models have primarily focused on image-to-image and text-to-image tasks, the challenge of preventing unauthorized inpainting has been rarely addressed, often resulting in suboptimal protection performance. To mitigate inpainting abuses, we propose ADVPAINT, a novel defensive framework that generates adversarial perturbations that effectively disrupt the adversary’s inpainting tasks. ADVPAINT targets the self- and cross-attention blocks in a target diffusion inpainting model to distract semantic understanding and prompt interactions during image generation. ADVPAINT also employs a two-stage perturbation strategy, dividing the perturbation region based on an enlarged bounding box around the object, enhancing robustness across diverse masks of varying shapes and sizes. Our experimental results demonstrate that ADVPAINT’s perturbations are highly effective in disrupting the adversary’s inpai

nting tasks, outperforming existing methods; ADVPAINT attains over a 100-point i
ncrease in FID and substantial decreases in precision.
****
Linus Lach,Alexander Willi Fottner,Yarema Okhrin
Bridging the Gap Between f-divergences and Bayes Hilbert Spaces
We introduce a novel framework that generalizes f-divergences by incorporating
 locally non-convex divergence-generating functions.
Using this extension, we define a new class of pseudo f-divergences, encompass
ing a wider range of distances between distributions that traditional f-diverg
ences cannot capture.
Among these, we focus on a particular pseudo divergence obtained by considering
the induced metric of Bayes Hilbert spaces.
Bayes Hilbert spaces are frequently used due to their inherent connection to Bay
es's theorem. They allow sampling from potentially intractable posterior densiti
es, which has remained challenging until now.
In the more general context, we prove that pseudo f-divergences are well-defin
ed and introduce a variational estimation framework that can be used in a statis
tical learning context.
By applying this variational estimation framework to f-GANs, we achieve improv
ed FID scores over existing f-GAN architectures and competitive results with t
he Wasserstein GAN, highlighting its potential for both theoretical research and
 practical applications in learning theory.
****
Kaiyue Wen,Zhiyuan Li,Jason S. Wang,David Leo Wright Hall,Percy Liang,Tengyu Ma
Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape
View
Training language models currently requires pre-determining a fixed compute budg
et because the typical cosine learning rate schedule depends on the total number
 of steps. In contrast, the Warmup-Stable-Decay (WSD) schedule uses a constant l
earning rate to produce a main branch of iterates that can in principle continue
 indefinitely without a pre-specified compute budget. Then, given any compute bu
dget, one can branch out from the main branch at a proper time with a rapidly de
caying learning rate to produce a strong model. Empirically, WSD generates an in
triguing, non-traditional loss curve: the loss remains elevated during the stabl
e phase but sharply declines during the decay phase. Towards explaining this phe
nomenon, we conjecture that pretraining loss exhibits a river valley landscape,
which resembles a deep valley with a river at its bottom. Under this assumption,
 we show that during the stable phase, the iterate undergoes large oscillations
due to the high learning rate, yet it progresses swiftly along the river. During
 the decay phase, the rapidly dropping learning rate minimizes the iterate’s osc
illations, moving it closer to the river and revealing true optimization progres
s. Therefore, the sustained high learning rate phase and fast decaying phase are
 responsible for progress in the river and the mountain directions, respectively
, and are both critical. Our analysis predicts phenomenons consistent with empir
ical observations and shows that this landscape can naturally emerge from pretra
ining on a simple bi-gram dataset. Inspired by the theory, we introduce WSD-S, a
 variant of WSD that reuses previous checkpoints’ decay phases and keeps only on
e main branch, where we resume from a decayed checkpoint. WSD-S empirically outp
erforms WSD and Cyclic-Cosine in obtaining multiple pretrained language model ch
eckpoints across various compute budgets in a single run for parameters scaling
from 0.1B to 1.2B.
****
Jianshuo Dong,Ziyuan Zhang,Qingjie Zhang,Tianwei Zhang,Hao Wang,Hewu Li,Qi Li,Ch
ao Zhang,Ke Xu,Han Qiu
An Engorgio Prompt Makes Large Language Model Babble on
Auto-regressive large language models (LLMs) have yielded impressive performance
 in many real-world tasks.
However, the new paradigm of these LLMs also exposes novel threats.
In this paper, we explore their vulnerability to inference cost attacks, where a
 malicious user crafts Engorgio prompts to intentionally increase the computatio

n cost and latency of the inference process. We design Engorgio, a novel methodo
logy, to efficiently generate adversarial Engorgio prompts to affect the target
LLM's service availability. Engorgio has the following two technical contributio
ns.
(1) We employ a parameterized distribution to track LLMs' prediction trajectory.
 (2) Targeting the auto-regressive nature of LLMs' inference process, we propose
 novel loss functions to stably suppress the appearance of the <EOS> token, whos
e occurrence will interrupt the LLM's generation process.
We conduct extensive experiments on 13 open-sourced LLMs with parameters ranging
 from 125M to 30B.
The results show that Engorgio prompts can successfully induce LLMs to generate
abnormally long outputs (i.e., roughly 2-13$\times$ longer to reach 90\%+ of the
 output length limit)
in a white-box scenario and our real-world experiment demonstrates Engergio's th
reat to LLM service with limited computing resources.
The code is released at https://github.com/jianshuod/Engorgio-prompt.
****
Yu-Ting Zhan,Cheng-Yuan Ho,Hebi Yang,Yi-Hsin Chen,Jui Chiu Chiang,Yu-Lun Liu,Wen
-Hsiao Peng
CAT-3DGS: A Context-Adaptive Triplane Approach to Rate-Distortion-Optimized 3DGS
 Compression
3D Gaussian Splatting (3DGS) has recently emerged as a promising 3D representati
on. Much research has been focused on reducing its storage requirements and memo
ry footprint. However, the needs to compress and transmit the 3DGS representatio
n to the remote side are overlooked. This new application calls for rate-distort
ion-optimized 3DGS compression. How to quantize and entropy encode sparse Gaussi
an primitives in the 3D space remains largely unexplored. Few early attempts res
ort to the hyperprior framework from learned image compression. But, they fail t
o utilize fully the inter and intra correlation inherent in Gaussian primitives.
 Built on ScaffoldGS, this work, termed CAT-3DGS, introduces a context-adaptive
triplane approach to their rate-distortion-optimized coding. It features multi-s
cale triplanes, oriented according to the principal axes of Gaussian primitives
in the 3D space, to capture their inter correlation (i.e. spatial correlation) f
or spatial autoregressive coding in the projected 2D planes. With these triplane
s serving as the hyperprior, we further perform channel-wise autoregressive codi
ng to leverage the intra correlation within each individual Gaussian primitive.
Our CAT-3DGS incorporates a view frequency-aware masking mechanism. It actively
skips from coding those Gaussian primitives that potentially have little impact
on the rendering quality. When trained end-to-end to strike a good rate-distorti
on trade-off, our CAT-3DGS achieves the state-of-the-art compression performance
 on the commonly used real-world datasets.
****
Parshin Shojaee,Kazem Meidani,Shashank Gupta,Amir Barati Farimani,Chandan K. Red
dy
LLM-SR: Scientific Equation Discovery via Programming with Large Language Models
Mathematical equations have been unreasonably effective in describing complex na
tural phenomena across various scientific disciplines. However, discovering such
 insightful equations from data presents significant challenges due to the neces
sity of navigating extremely large combinatorial hypothesis spaces. Current meth
ods of equation discovery, commonly known as symbolic regression techniques, lar
gely focus on extracting equations from data alone, often neglecting the domain-
specific prior knowledge that scientists typically depend on. They also employ l
imited representations such as expression trees, constraining the search space a
nd expressiveness of equations. To bridge this gap, we introduce LLM-SR, a novel
 approach that leverages the extensive scientific knowledge and robust code gene
ration capabilities of Large Language Models (LLMs) to discover scientific equat
ions from data. Specifically, LLM-SR treats equations as programs with mathemati
cal operators and combines LLMs' scientific priors with evolutionary search over
 equation programs. The LLM iteratively proposes new equation skeleton hypothese
s, drawing from its domain knowledge, which are then optimized against data to e

stimate parameters. We evaluate LLM-SR on four benchmark problems across diverse scientific domains (e.g., physics, biology), which we carefully designed to simulate the discovery process and prevent LLM recitation. Our results demonstrate that LLM-SR discovers physically accurate equations that significantly outperform state-of-the-art symbolic regression baselines, particularly in out-of-domain test settings. We also show that LLM-SR's incorporation of scientific priors enables more efficient equation space exploration than the baselines.
****
Sabine Muzellec,Drew Linsley,Alekh Karkada Ashok,Ennio Mingolla,Girik Malik,Rufin VanRullen,Thomas Serre
Tracking objects that change in appearance with phase synchrony
Objects we encounter often change appearance as we interact with them. Changes in illumination (shadows), object pose, or the movement of non-rigid objects can drastically alter available image features. How do biological visual systems track objects as they change? One plausible mechanism involves attentional mechanisms for reasoning about the locations of objects independently of their appearances --- a capability that prominent neuroscience theories have associated with computing through neural synchrony. Here, we describe a novel deep learning circuit that can learn to precisely control attention to features separately from their location in the world through neural synchrony: the complex-valued recurrent neural network (CV-RNN). Next, we compare object tracking in humans, the CV-RNN, and other deep neural networks (DNNs), using FeatureTracker: a large-scale challenge that asks observers to track objects as their locations and appearances change in precisely controlled ways. While humans effortlessly solved FeatureTracker, state-of-the-art DNNs did not. In contrast, our CV-RNN behaved similarly to humans on the challenge, providing a computational proof-of-concept for the role of phase synchronization as a neural substrate for tracking appearance-morphing objects as they move about.
****
Xingjian Wu,Xiangfei Qiu,Zhengyu Li,Yihang Wang,Jilin Hu,Chenjuan Guo,Hui Xiong,Bin Yang
CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching
Anomaly detection in multivariate time series is challenging as heterogeneous subsequence anomalies may occur. Reconstruction-based methods, which focus on learning normal patterns in the frequency domain to detect diverse abnormal subsequences, achieve promising results, while still falling short on capturing fine-grained frequency characteristics and channel correlations. To contend with the limitations, we introduce CATCH, a framework based on frequency patching. We propose to patchify the frequency domain into frequency bands, which enhances its ability to capture fine-grained frequency characteristics. To perceive appropriate channel correlations, we propose a Channel Fusion Module (CFM), which features a patch-wise mask generator and a masked-attention mechanism. Driven by a bi-level multi-objective optimization algorithm, the CFM is encouraged to iteratively discover appropriate patch-wise channel correlations, and to cluster relevant channels while isolating adverse effects from irrelevant channels. Extensive experiments on 10 real-world datasets and 12 synthetic datasets demonstrate that CATCH achieves state-of-the-art performance. We make our code and datasets available at https://github.com/decisionintelligence/CATCH.
****
Xiaoran Jiao,Weian Mao,Wengong Jin,Peiyuan Yang,Hao Chen,Chunhua Shen
Boltzmann-Aligned Inverse Folding Model as a Predictor of Mutational Effects on Protein-Protein Interactions
Predicting the change in binding free energy ($\Delta \Delta G$) is crucial for understanding and modulating protein-protein interactions, which are critical in drug design.
Due to the scarcity of experimental $\Delta\Delta G$ data,
existing methods focus on pre-training,
while neglecting the importance of alignment.
In this work, we propose Boltzmann Alignment technique to transfer knowledge fro

m pre-trained inverse folding models to prediction of $\Delta\Delta G$.
We begin by analyzing the thermodynamic definition of $\Delta\Delta G$ and intro
ducing the Boltzmann distribution to connect energy to the protein conformationa
l distribution.
However, the protein conformational distribution is intractable. Therefore, we e
mploy Bayes' theorem to circumvent direct estimation and instead utilize the log
-likelihood provided by protein inverse folding models for the estimation of $\D
elta\Delta G$.

Compared to previous methods based on inverse folding, our method explicitly acc
ounts for the unbound state of the protein complex in the $\Delta \Delta G$ ther
modynamic cycle, introducing a physical inductive bias and achieving supervised
and unsupervised state-of-the-art (SoTA) performance.
Experimental results on SKEMPI v2 indicate that our method achieves Spearman coe
fficients of 0.3201 (unsupervised) and 0.5134 (supervised) on SKEMPI v2, signifi
cantly surpassing the previously reported
%SoTA values
SoTA results
of 0.2632 and 0.4324, respectively.
Furthermore, we demonstrate the capability of our method in binding
energy prediction, protein-protein docking, and antibody optimization tasks.

Code is available at [https://github.com/aim-uofa/BA-DDG](https://github.com/aim
-uofa/BA-DDG)
****
Michael Crawshaw,Blake Woodworth,Mingrui Liu
Local Steps Speed Up Local GD for Heterogeneous Distributed Logistic Regression
We analyze two variants of Local Gradient Descent applied to distributed logisti
c regression with heterogeneous, separable data and show convergence at the rate
 $O(1/KR)$ for K local steps and sufficiently large R communication rounds.
In contrast, all existing convergence guarantees for Local GD applied to any pro
blem are at least $\Omega(1/R)$, meaning they fail to show the benefit of local
updates. The key to our improved guarantee is showing progress on the logistic r
egression objective when using a large stepsize $\eta \gg 1/K$, whereas prior an
alysis depends on $\eta \leq 1/K$.
****
Emily Jin,Joy Hsu,Jiajun Wu
Predicate Hierarchies Improve Few-Shot State Classification
State classification of objects and their relations is core to many long-horizon
 tasks, particularly in robot planning and manipulation. However, the combinator
ial explosion of possible object-predicate combinations, coupled with the need t
o adapt to novel real-world environments, makes it a desideratum for state class
ification models to generalize to novel queries with few examples. To this end,
we propose PHIER, which leverages predicate hierarchies to generalize effectivel
y in few-shot scenarios. PHIER uses an object-centric scene encoder, self-superv
ised losses that infer semantic relations between predicates, and a hyperbolic d
istance metric that captures hierarchical structure; it learns a structured late
nt space of image-predicate pairs that guides reasoning over state classificatio
n queries. We evaluate PHIER in the CALVIN and BEHAVIOR robotic environments and
 show that PHIER significantly outperforms existing methods in few-shot, out-of-
distribution state classification, and demonstrates strong zero- and few-shot ge
neralization from simulated to real-world tasks. Our results demonstrate that le
veraging predicate hierarchies improves performance on state classification task
s with limited data.
****
Chaitanya K. Joshi,Arian Rokkum Jamasb,Ramon Viñas Torné,Charles Harris,Simon V
Mathis,Alex Morehead,Rishabh Anand,Pietro Lio
gRNAde: Geometric Deep Learning for 3D RNA inverse design
Computational RNA design tasks are often posed as inverse problems, where sequen
ces are designed based on adopting a single desired secondary structure without

considering 3D conformational diversity. We introduce gRNAde, a geometric RNA de
sign pipeline operating on 3D RNA backbones to design sequences that explicitly
account for structure and dynamics. gRNAde uses a multi-state Graph Neural Netwo
rk and autoregressive decoding to generates candidate RNA sequences conditioned
on one or more 3D backbone structures where the identities of the bases are unkn
own. On a single-state fixed backbone re-design benchmark of 14 RNA structures f
rom the PDB identified by Das et al. (2010), gRNAde obtains higher native sequen
ce recovery rates (56% on average) compared to Rosetta (45% on average), taking
under a second to produce designs compared to the reported hours for Rosetta. We
 further demonstrate the utility of gRNAde on a new benchmark of multi-state des
ign for structurally flexible RNAs, as well as zero-shot ranking of mutational f
itness landscapes in a retrospective analysis of a recent ribozyme. Experimental
 wet lab validation on 10 different structured RNA backbones finds that gRNAde h
as a success rate of 50% at designing pseudoknotted RNA structures, a significan
t advance over 35% for Rosetta. Open source code and tutorials are available at:
 github.com/chaitjo/geometric-rna-design
****

Soroush H. Zargarbashi,Aleksandar Bojchevski
Robust Conformal Prediction with a Single Binary Certificate
Conformal prediction (CP) converts any model's output to prediction sets with a
guarantee to cover the true label with (adjustable) high probability. Robust CP
extends this guarantee to worst-case (adversarial) inputs. Existing baselines ac
hieve robustness by bounding randomly smoothed conformity scores. In practice, t
hey need expensive Monte-Carlo (MC) sampling (e.g. $\sim10^4$ samples per point)
 to maintain an acceptable set size. We propose a robust conformal prediction th
at produces smaller sets even with significantly lower MC samples (e.g. 150 for
CIFAR10). Our approach binarizes samples with an adjustable (or automatically ad
justed) threshold selected to preserve the coverage guarantee. Remarkably, we pr
ove that robustness can be achieved by computing only one binary certificate, un
like previous methods that certify each calibration (or test) point. Thus, our m
ethod is faster and returns smaller robust sets. We also eliminate a previous li
mitation that requires a bounded score function.
****

Sammy Khalife,Josué Tonelli-Cueto
Is uniform expressivity too restrictive? Towards efficient expressivity of GNNs
Uniform expressivity guarantees that a Graph Neural Network (GNN) can express a
 query without the parameters depending on the size of the input graphs. This pr
operty is desirable in applications in order to have number of trainable parame
ters that is independent of the size of the input graphs. Uniform expressivity
of the two variable guarded fragment (GC2) of first order logic is a well-celebr
ated result for Rectified Linear Unit (ReLU) GNNs [Barcelo &. Al, 2020]. In this
 article, we prove that uniform expressivity of GC2 queries is not possible for
GNNs with a wide class of Pfaffian activation functions (including the sigmoid a
nd $\tanh$), answering a question formulated by [Grohe, 2021]. We also show that
 despite these limitations, many of those GNNs can still efficiently express GC2
 queries in a way that the number of parameters remains logarithmic on the maxim
al degree of the input graphs. Furthermore, we demonstrate that a log-log depend
ency on the degree is achievable for a certain choice of activation function. Th
is shows that uniform expressivity can be successfully relaxed by covering large
 graphs appearing in practical applications. Our experiments illustrates that ou
r theoretical estimates hold in practice.
****

Ann-Kathrin Schuetz,A.W.P. Poon,Aobo Li
RESuM: A Rare Event Surrogate Model for Physics Detector Design
The experimental discovery of neutrinoless double-beta decay (NLDBD) would answe
r one of the most important questions in physics: Why is there more matter than
antimatter in our universe? To maximize the chances of discovery, NLDBD experime
nts must optimize their detector designs to minimize the probability of backgrou
nd events contaminating the detector. Given that this probability is inherently
low, design optimization either requires extremely costly simulations to generat

e sufficient background counts or contending with significant variance. In this work, we formalize this dilemma as a Rare Event Design (RED) problem: identifyin g optimal design parameters when the design metric to be minimized is inherently small. We then designed the Rare Event Surrogate Model (RESuM) for physics dete ctor design optimization under RED conditions. RESuM uses a pre-trained Conditio nal Neural Process (CNP) model to incorporate additional prior knowledge into a Multi-Fidelity Gaussian Process model. We applied RESuM to optimize neutron shie lding designs for the LEGEND NLDBD experiment, identifying an optimal design tha t reduces the neutron background by (66.5 ± 3.5)% while using only 3.3% of t he computational resources compared to traditional methods. Given the prevalence of RED problems in other fields of physical sciences, especially in rare-event searches, the RESuM algorithm has broad potential for accelerating simulation-in tensive applications.
****

Mohammad Mozaffari,Amir Yazdanbakhsh,Zhao Zhang,Maryam Mehri Dehnavi
SLoPe: Double-Pruned Sparse Plus Lazy Low-Rank Adapter Pretraining of LLMs
We propose SLoPe, a Double-Pruned **S**parse Plus **L**azy L**o**w-rank Adapter **P**r**e**training method for LLMs that improves the accuracy of sparse LLMs wh ile accelerating their pretraining and inference and reducing their memory footp rint. Sparse pretraining of LLMs reduces the accuracy of the model, to overcome this, prior work uses dense models during fine-tuning. SLoPe improves the accura cy of sparsely pretrained models by adding low-rank adapters in the final 1% ite rations of pretraining without adding significant overheads to the model pretrai ning and inference. In addition, SLoPe uses a double-pruned backward pass formul ation that prunes the transposed weight matrix using N:M sparsity structures to enable an accelerated sparse backward pass. SLoPe accelerates the training and i nference of models with billions of parameters up to 1.25× and 1.54× respectivel y (OPT-33B and OPT-66B) while reducing their memory usage by up to 0.63× and 0.6 1× for training and inference respectively.
****

Yam Tawachi,Bracha Laufer-Goldshtein
Multi-Dimensional Conformal Prediction
Conformal prediction has attracted significant attention as a distribution-free method for uncertainty quantification in black-box models, providing prediction sets with guaranteed coverage. However, its practical utility is often limited w hen these prediction sets become excessively large, reducing its overall effecti veness. In this paper, we introduce a novel approach to conformal prediction for classification problems, which leverages a multi-dimensional nonconformity scor e. By extending standard conformal prediction to higher dimensions, we achieve b etter separation between correct and incorrect labels. Utilizing this we can foc us on regions with low concentrations of incorrect labels, leading to smaller, m ore informative prediction sets. To efficiently generate the multi-dimensional s core, we employ a self-ensembling technique that trains multiple diverse classif ication heads on top of a backbone model. We demonstrate the advantage of our ap proach compared to baselines across different benchmarks.
****

Zesen Wang,Jiaojiao Zhang,Xuyang Wu,Mikael Johansson
From Promise to Practice: Realizing High-performance Decentralized Training
Decentralized training of deep neural networks has attracted significant attenti on for its theoretically superior scalability compared to synchronous data-paral lel methods like All-Reduce. However, realizing this potential in multi-node tra ining is challenging due to the complex design space that involves communication topologies, computation patterns, and optimization algorithms. This paper ident ifies three key factors that can lead to speedups over All-Reduce training and c onstructs a runtime model to determine when and how decentralization can shorten the per-iteration runtimes. To support the decentralized training of transforme r-based models, we introduce a decentralized Adam algorithm that overlaps commun ications with computations, prove its convergence, and propose an accumulation t echnique to mitigate the high variance caused by small local batch sizes. We dep loy our solution in clusters with up to 64 GPUs, demonstrating its practical adv

antages in both runtime and generalization performance under a fixed iteration budget.
The experiment code is open-source at https://github.com/WangZesen/Decentralized-Training-Exp, and the extension code is open-source at https://github.com/WangZesen/Decent-DP.
****
Gabriele Dominici,Pietro Barbiero,Mateo Espinosa Zarlenga,Alberto Termine,Martin Gjoreski,Giuseppe Marra,Marc Langheinrich
Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning
Causal opacity denotes the difficulty in understanding the "hidden" causal structure underlying the decisions of deep neural network (DNN) models. This leads to the inability to rely on and verify state-of-the-art DNN-based systems, especially in high-stakes scenarios. For this reason, circumventing causal opacity in DNNs represents a key open challenge at the intersection of deep learning, interpretability, and causality. This work addresses this gap by introducing Causal Concept Graph Models (Causal CGMs), a class of interpretable models whose decision-making process is causally transparent by design. Our experiments show that Causal CGMs can: (i) match the generalisation performance of causally opaque models, (ii) enable human-in-the-loop corrections to mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections but also the reliability of the explanations provided for specific instances, and (iii) support the analysis of interventional and counterfactual scenarios, thereby improving the model's causal interpretability and supporting the effective verification of its reliability and fairness.
****
Wesley Khademi,Fuxin Li
Point-based Instance Completion with Scene Constraints
Recent point-based object completion methods have demonstrated the ability to accurately recover the missing geometry of partially observed objects. However, these approaches are not well-suited for completing objects within a scene, as they do not consider known scene constraints (e.g., other observed surfaces) in their completions and further expect the partial input to be in a canonical coordinate system, which does not hold for objects within scenes. While instance scene completion methods have been proposed for completing objects within a scene, they lag behind point-based object completion methods in terms of object completion quality and still do not consider known scene constraints during completion. To overcome these limitations, we propose a point cloud-based instance completion model that can robustly complete objects at arbitrary scales and pose in the scene. To enable reasoning at the scene level, we introduce a sparse set of scene constraints represented as point clouds and integrate them into our completion model via a cross-attention mechanism. To evaluate the instance scene completion task on indoor scenes, we further build a new dataset called ScanWCF, which contains labeled partial scans as well as aligned ground truth scene completions that are watertight and collision-free. Through several experiments, we demonstrate that our method achieves improved fidelity to partial scans, higher completion quality, and greater plausibility over existing state-of-the-art methods.
****
Kasra Arabi,Benjamin Feuer,R. Teal Witter,Chinmay Hegde,Niv Cohen
Hidden in the Noise: Two-Stage Robust Watermarking for Images
As the quality of image generators continues to improve, deepfakes become a topic of considerable societal debate. Image watermarking allows responsible model owners to detect and label their AI-generated content, which can mitigate the harm. Yet, current state-of-the-art methods in image watermarking remain vulnerable to forgery and removal attacks.
This vulnerability occurs in part because watermarks distort the distribution of generated images, unintentionally revealing information about the watermarking techniques.

In this work, we first demonstrate a distortion-free watermarking method for ima

ges, based on a diffusion model's initial noise.
However, detecting the watermark requires comparing the initial noise reconstructed for an image to all previously used initial noises.
To mitigate these issues, we propose a two-stage watermarking framework for efficient detection. During generation, we augment the initial noise with generated Fourier patterns to embed information about the group of initial noises we used. For detection, we (i) retrieve the relevant group of noises, and (ii) search within the given group for an initial noise that might match our image. This watermarking approach achieves state-of-the-art robustness to forgery and removal against a large battery of attacks. The project code is available at https://github.com/Kasraarabi/Hidden-in-the-Noise.
****
Dingling Yao,Dario Rancati,Riccardo Cadei,Marco Fumero,Francesco Locatello
Unifying Causal Representation Learning with the Invariance Principle
Causal representation learning (CRL) aims at recovering latent causal variables from high-dimensional observations to solve causal downstream tasks, such as predicting the effect of new interventions or more robust classification.
A plethora of methods have been developed, each tackling carefully crafted problem settings that lead to different types of identifiability.
These different settings are widely assumed to be important because they are often linked to different rungs of Pearl's causal hierarchy, even though this correspondence is not always exact.
This work shows that instead of strictly conforming to this hierarchical mapping, *many causal representation learning approaches methodologically align their representations with inherent data symmetries.*
Identification of causal variables is guided by invariance principles that are not necessarily causal.
This result allows us to unify many existing approaches in a single method that can mix and match different assumptions, including non-causal ones, based on the invariance relevant to the problem at hand.
It also significantly benefits applicability, which we demonstrate by improving treatment effect estimation on real-world high-dimensional ecological data. Overall, this paper clarifies the role of causal assumptions in the discovery of causal variables and shifts the focus to preserving data symmetries.
****
Zihan Yu,Jingtao Ding,Yong Li,Depeng Jin
Symbolic regression via MDLformer-guided search: from minimizing prediction error to minimizing description length
Symbolic regression, a task discovering the formula best fitting the given data, is typically based on the heuristical search. These methods usually update candidate formulas to obtain new ones with lower prediction errors iteratively. However, since formulas with similar function shapes may have completely different symbolic forms, the prediction error does not decrease monotonously as the search approaches the target formula, causing the low recovery rate of existing methods. To solve this problem, we propose a novel search objective based on the minimum description length, which reflects the distance from the target and decreases monotonically as the search approaches the correct form of the target formula.
To estimate the minimum description length of any input data, we design a neural network, MDLformer, which enables robust and scalable estimation through large-scale training. With the MDLformer's output as the search objective, we implement a symbolic regression method, SR4MDL, that can effectively recover the correct mathematical form of the formula. Extensive experiments illustrate its excellent performance in recovering formulas from data. Our method successfully recovers around 50 formulas across two benchmark datasets comprising 133 problems, outperforming state-of-the-art methods by 43.92%. Experiments on 122 unseen black-box problems further demonstrate its generalization performance. We release our code at https://github.com/tsinghua-fib-lab/SR4MDL .
****
Chankyu Lee,Rajarshi Roy,Mengyao Xu,Jonathan Raiman,Mohammad Shoeybi,Bryan Catanzaro,Wei Ping

NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models
Decoder-only large language model (LLM)-based embedding models are beginning to outperform BERT or T5-based embedding models in general-purpose text embedding tasks, including dense vector-based retrieval. In this work, we introduce the NV-Embed model, incorporating architectural designs, training procedures, and curated datasets to significantly enhance the performance of LLM as a versatile embedding model, while maintaining its simplicity and reproducibility.For model architecture, we propose a latent attention layer to obtain pooled embeddings, which consistently improves retrieval and downstream task accuracy compared to mean pooling or using the last <EOS> token embedding from LLMs. To enhance representation learning, we remove the causal attention mask of LLMs during contrastive training. For training algorithm, we introduce a two-stage contrastive instruction-tuning method. It first applies contrastive training with instructions on retrieval datasets, utilizing in-batch negatives and curated hard negative examples. At stage-2, it blends various non-retrieval into instruction tuning, which not only enhances non-retrieval task accuracy but also improves retrieval performance. For training data, we utilize the hard-negative mining, synthetic data generation and existing public available datasets to boost the performance of embedding model. By combining these techniques, our NV-Embed- v1 model secured the No.1 position on the Massive Text Embedding Benchmark (MTEB) (as of May 24, 2024), across 56 embedding tasks. NV-Embed-v2 has reclaimed and maintained the top spot on MTEB since August 30, 2024, demonstrating the sustained effectiveness of the proposed methods over time. Also, it achieved the highest scores in the Long Doc section and the second-highest scores in the QA section of the AIR Benchmark, which covers a range of out-of-domain information retrieval topics beyond those in MTEB. We further provide the analysis of model compression techniques for generalist embedding models. We open-source the model at: https://huggingface.co/nvidia/NV-Embed-v2 .
****
Aidan Scannell,Mohammadreza Nakhaeinezhadfard,Kalle Kujanpää,Yi Zhao,Kevin Sebastian Luck,Arno Solin,Joni Pajarinen
Discrete Codebook World Models for Continuous Control
In reinforcement learning (RL), world models serve as internal simulators, enabling agents to predict environment dynamics and future outcomes in order to make informed decisions. While previous approaches leveraging discrete latent spaces, such as DreamerV3, have demonstrated strong performance in discrete action settings and visual control tasks, their comparative performance in state-based continuous control remains underexplored. In contrast, methods with continuous latent spaces, such as TD-MPC2, have shown notable success in state-based continuous control benchmarks. In this paper, we demonstrate that modeling discrete latent states has benefits over continuous latent states and that discrete codebook encodings are more effective representations for continuous control, compared to alternative encodings, such as one-hot and label-based encodings. Based on these insights, we introduce DCWM: Discrete Codebook World Model, a self-supervised world model with a discrete and stochastic latent space, where latent states are codes from a codebook. We combine DCWM with decision-time planning to get our model-based RL algorithm, named DC-MPC: Discrete Codebook Model Predictive Control, which performs competitively against recent state-of-the-art algorithms, including TD-MPC2 and DreamerV3, on continuous control benchmarks.
****
Ezra Karger,Houtan Bastani,Chen Yueh-Han,Zachary Jacobs,Danny Halawi,Fred Zhang,Philip Tetlock
ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities
Forecasts of future events are essential inputs into informed decision-making. Machine learning (ML) systems have the potential to deliver forecasts at scale, but there is no framework for evaluating the accuracy of ML systems on a standardized set of forecasting questions. To address this gap, we introduce ForecastBench: a dynamic benchmark that evaluates the accuracy of ML systems on an automatically generated and regularly updated set of 1,000 forecasting questions. To avoid any possibility of data leakage, ForecastBench is comprised solely of questio

ns about future events that have no known answer at the time of submission. We q
uantify the capabilities of current ML systems by collecting forecasts from expe
rt (human) forecasters, the general public, and LLMs on a random subset of quest
ions from the benchmark ($N=200$). While LLMs have achieved super-human performa
nce on many benchmarks, they perform less well here: expert forecasters outperfo
rm the top-performing LLM (p-value <0.001). We display system and human scor
es in a public leaderboard at www.forecastbench.org.
****
Théo Vincent,Fabian Wahren,Jan Peters,Boris Belousov,Carlo D'Eramo
Adaptive Q-Network: On-the-fly Target Selection for Deep Reinforcement Learnin
g
Deep Reinforcement Learning (RL) is well known for being highly sensitive to hyp
erparameters, requiring practitioners substantial efforts to optimize them for t
he problem at hand. This also limits the applicability of RL in real-world scena
rios. In recent years, the field of automated Reinforcement Learning (AutoRL) ha
s grown in popularity by trying to address this issue. However, these approaches
 typically hinge on additional samples to select well-performing hyperparameters
, hindering sample-efficiency and practicality. Furthermore, most AutoRL methods
 are heavily based on already existing AutoML methods, which were originally dev
eloped neglecting the additional challenges inherent to RL due to its non-statio
narities. In this work, we propose a new approach for AutoRL, called _Adaptive $
Q$-Network_ (AdaQN), that is tailored to RL to take into account the non-station
arity of the optimization procedure without requiring additional samples. AdaQN
learns several Q-functions, each one trained with different hyperparameters, w
hich are updated online using the Q-function with the smallest approximation e
rror as a shared target. Our selection scheme simultaneously handles different h
yperparameters while coping with the non-stationarity induced by the RL optimiza
tion procedure and being orthogonal to any critic-based RL algorithm. We demonst
rate that AdaQN is theoretically sound and empirically validate it in MuJoCo con
trol problems and Atari 2600 games, showing benefits in sample-efficiency, ove
rall performance, robustness to stochasticity and training stability. Our code i
s available at *https://github.com/theovincent/AdaDQN*.
****
Guo Chen,Yicheng Liu,Yifei Huang,Baoqi Pei,Jilan Xu,Yuping He,Tong Lu,Yali Wang,
Limin Wang
CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understandin
g
The existing video understanding benchmarks for multimodal large language models
 (MLLMs) mainly focus on short videos. The few benchmarks for long video underst
anding often rely on multiple-choice questions (MCQs). Due to the limitations of
 MCQ evaluations and the advanced reasoning abilities of MLLMs, models can often
 answer correctly by combining short video insights with elimination, without tr
uly understanding the content. To bridge this gap, we introduce CG-Bench, a benc
hmark for clue-grounded question answering in long videos. CG-Bench emphasizes t
he model's ability to retrieve relevant clues, enhancing evaluation credibility.
 It includes 1,219 manually curated videos organized into 14 primary, 171 second
ary, and 638 tertiary categories, making it the largest benchmark for long video
 analysis. The dataset features 12,129 QA pairs in three question types: percept
ion, reasoning, and hallucination. To address the limitations of MCQ-based evalu
ation, we develop two novel clue-based methods: clue-grounded white box and blac
k box evaluations, assessing whether models generate answers based on accurate v
ideo understanding. We evaluated multiple closed-source and open-source MLLMs on
 CG-Bench. The results show that current models struggle significantly with long
 videos compared to short ones, and there is a notable gap between open-source a
nd commercial models. We hope CG-Bench will drive the development of more reliab
le and capable MLLMs for long video comprehension.
****
Linjiajie Fang,Ruoxue Liu,Jing Zhang,Wenjia Wang,Bingyi Jing
Diffusion Actor-Critic: Formulating Constrained Policy Iteration as Diffusion No
ise Regression for Offline Reinforcement Learning

In offline reinforcement learning, it is necessary to manage out-of-distribution actions to prevent overestimation of value functions. One class of methods, the policy-regularized method, addresses this problem by constraining the target policy to stay close to the behavior policy. Although several approaches suggest representing the behavior policy as an expressive diffusion model to boost performance, it remains unclear how to regularize the target policy given a diffusion-modeled behavior sampler. In this paper, we propose Diffusion Actor-Critic (DAC) that formulates the Kullback-Leibler (KL) constraint policy iteration as a diffusion noise regression problem, enabling direct representation of target policies as diffusion models. Our approach follows the actor-critic learning paradigm in which we alternatively train a diffusion-modeled target policy and a critic network. The actor training loss includes a soft Q-guidance term from the Q-gradient. The soft Q-guidance is based on the theoretical solution of the KL constraint policy iteration, which prevents the learned policy from taking out-of-distribution actions. We demonstrate that such diffusion-based policy constraint, along with the coupling of the lower confidence bound of the Q-ensemble as value targets, not only preserves the multi-modality of target policies, but also contributes to stable convergence and strong performance in DAC. Our approach is evaluated on D4RL benchmarks and outperforms the state-of-the-art in nearly all environments.
****

Ruomin Huang,Rong Ge

Task Descriptors Help Transformers Learn Linear Models In-Context

Large language models (LLMs) exhibit strong in-context learning (ICL) ability, which allows the model to make predictions on new examples based on the given prompt. Recently, a line of research (Von Oswald et al., 2023; Aky¨urek et al., 2023; Ahn et al., 2023; Mahankali et al., 2023; Zhang et al., 2024) considered ICL for a simple linear regression setting and showed that the forward pass of Transformers is simulating some variants of gradient descent (GD) algorithms on the in-context examples. In practice, the input prompt usually contains a task descriptor in addition to in-context examples. We investigate how the task description helps ICL in the linear regression setting. Consider a simple setting where the task descriptor describes the mean of input in linear regression. Our results show that gradient flow converges to a global minimum for a linear Transformer. At the global minimum, the Transformer learns to use the task descriptor effectively to improve its performance. Empirically, we verify our results by showing that the weights converge to the predicted global minimum and Transformers indeed perform better with task descriptors.
****

Akira Ito,Masanori Yamada,Atsutoshi Kumagai

Analysis of Linear Mode Connectivity via Permutation-Based Weight Matching: With Insights into Other Permutation Search Methods

Recently, Ainsworth et al. (2023) showed that using weight matching (WM) to minimize the L^2 distance in a permutation search of model parameters effectively identifies permutations that satisfy linear mode connectivity (LMC), where the loss along a linear path between two independently trained models with different seeds remains nearly constant. This paper analyzes LMC using WM, which is useful for understanding stochastic gradient descent's effectiveness and its application in areas like model merging. We first empirically show that permutations found by WM do not significantly reduce the L^2 distance between two models, and the occurrence of LMC is not merely due to distance reduction by WM itself. We then demonstrate that permutations can change the directions of the singular vectors, but not the singular values, of the weight matrices in each layer. This finding shows that permutations found by WM primarily align the directions of singular vectors associated with large singular values across models. This alignment brings the singular vectors with large singular values, which determine the model's functionality, closer between the original and merged models, allowing the merged model to retain functionality similar to the original models, thereby satisfying LMC. This paper also analyzes activation matching (AM) in terms of singular vectors and finds that the principle of AM is likely the same as that of WM. F

inally, we analyze the difference between WM and the straight-through estimator (STE), a dataset-dependent permutation search method, and show that WM can be more advantageous than STE in achieving LMC among three or more models.
****

Ziyang Wu,Tianjiao Ding,Yifu Lu,Druv Pai,Jingyuan Zhang,Weida Wang,Yaodong Yu,Yi Ma,Benjamin David Haeffele

Token Statistics Transformer: Linear-Time Attention via Variational Rate Reduction

The attention operator is arguably the key distinguishing factor of transformer architectures, which have demonstrated state-of-the-art performance on a variety of tasks. However, transformer attention operators often impose a significant computational burden, with the computational complexity scaling quadratically with the number of tokens. In this work, we propose a novel transformer attention operator whose computational complexity scales linearly with the number of tokens. We derive our network architecture by extending prior work which has shown that a transformer style architecture naturally arises by "white-box" architecture design, where each layer of the network is designed to implement an incremental optimization step of a maximal coding rate reduction objective (MCR^2). Specifically, we derive a novel variational form of the MCR^2 objective and show that the architecture that results from unrolled gradient descent of this variational objective leads to a new attention module called Token Statistics Self-Attention ($\texttt{TSSA}$). $\texttt{TSSA}$ has $\textit{linear computational and memory complexity}$ and radically departs from the typical attention architecture that computes pairwise similarities between tokens. Experiments on vision, language, and long sequence tasks show that simply swapping $\texttt{TSSA}$ for standard self-attention, which we refer to as the Token Statistics Transformer ($\texttt{ToST}$), achieves competitive performance with conventional transformers while being significantly more computationally efficient and interpretable. Our results also somewhat call into question the conventional wisdom that pairwise similarity style attention mechanisms are critical to the success of transformer architectures.
****

Francisco Eiras,Aleksandar Petrov,Philip Torr,M. Pawan Kumar,Adel Bibi

Do as I do (Safely): Mitigating Task-Specific Fine-tuning Risks in Large Language Models

Recent research shows that fine-tuning on benign instruction-following data can inadvertently undo the safety alignment process and increase a model's propensity to comply with harmful queries. While instruction-following fine-tuning is important, task-specific fine-tuning-where models are trained on datasets with clear ground truth answers (e.g., multiple choice questions)-can enhance model performance on specialized downstream tasks. Understanding and mitigating safety risks in the task-specific setting remains distinct from the instruction-following context due to structural differences in the data. Our work demonstrates how malicious actors can subtly manipulate the structure of almost *any* task-specific dataset to foster significantly more dangerous model behaviors, while maintaining an appearance of innocuity and reasonable downstream task performance. To address this issue, we propose a novel mitigation strategy that mixes in safety data which *mimics* the task format and prompting style of the user data, showing this is significantly more effective and efficient than existing baselines at re-establishing safety alignment while maintaining similar task performance.
****

Gaojie Jin,Sihao Wu,Jiaxu Liu,Tianjin Huang,Ronghui Mu

Enhancing Robust Fairness via Confusional Spectral Regularization

Recent research has highlighted a critical issue known as ``robust fairness", where robust accuracy varies significantly across different classes, undermining the reliability of deep neural networks (DNNs).
A common approach to address this has been to dynamically reweight classes during training, giving more weight to those with lower empirical robust performance.

However, we find there is a divergence of class-wise robust performance between

training set and testing set, which limits the effectiveness of these explicit r
eweighting methods, indicating the need for a principled alternative.
In this work, we derive a robust generalization bound for the worst-class robust
 error within the PAC-Bayesian framework, accounting for unknown data distributi
ons.
Our analysis shows that the worst-class robust error is influenced by two main f
actors: the spectral norm of the empirical robust confusion matrix and the infor
mation embedded in the model and training set.
While the latter has been extensively studied, we propose a novel regularization
 technique targeting the spectral norm of the robust confusion matrix to improve
 worst-class robust accuracy and enhance robust fairness.
We validate our approach through comprehensive experiments on various datasets a
nd models, demonstrating its effectiveness in enhancing robust fairness.
****
Leqi Shen,Tianxiang Hao,Tao He,Sicheng Zhao,Yifeng Zhang,pengzhang liu,Yongjun B
ao,Guiguang Ding
TempMe: Video Temporal Token Merging for Efficient Text-Video Retrieval
Most text-video retrieval methods utilize the text-image pre-trained models like
 CLIP as a backbone. These methods process each sampled frame independently by t
he image encoder, resulting in high computational overhead and limiting practica
l deployment. Addressing this, we focus on efficient text-video retrieval by tac
kling two key challenges: 1. From the perspective of trainable parameters, curre
nt parameter-efficient fine-tuning methods incur high inference costs; 2. From t
he perspective of model complexity, current token compression methods are mainly
 designed for images to reduce spatial redundancy but overlook temporal redundan
cy in consecutive frames of a video. To tackle these challenges, we propose Temp
oral Token Merging (TempMe), a parameter-efficient and training-inference effici
ent text-video retrieval architecture that minimizes trainable parameters and mo
del complexity. Specifically, we introduce a progressive multi-granularity frame
work. By gradually combining neighboring clips, we reduce spatio-temporal redund
ancy and enhance temporal modeling across different frames, leading to improved
efficiency and performance. Extensive experiments validate the superiority of ou
r TempMe. Compared to previous parameter-efficient text-video retrieval methods,
 TempMe achieves superior performance with just 0.50M trainable parameters. It s
ignificantly reduces output tokens by 95% and GFLOPs by 51%, while achieving a 1
.8X speedup and a 4.4% R-Sum improvement. With full fine-tuning, TempMe achieves
 a significant 7.9% R-Sum improvement, trains 1.57X faster, and utilizes 75.2% G
PU memory usage. The code is available at https://github.com/LunarShen/TempMe.
****
Jing Wang,Ao Ma,Jiasong Feng,Dawei Leng,Yuhui Yin,Xiaodan Liang
PT-T2I/V: An Efficient Proxy-Tokenized Diffusion Transformer for Text-to-Image/V
ideo-Task
The global self-attention mechanism in diffusion transformers involves redundant
 computation due to the sparse and redundant nature of visual information, and t
he attention map of tokens within a spatial window shows significant similarity.
 To address this redundancy, we propose the Proxy-Tokenized Diffusion Transforme
r (PT-DiT), which employs sparse representative token attention (where the numbe
r of representative tokens is much smaller than the total number of tokens) to e
fficiently model global visual information. Specifically, within each transforme
r block, we compute an averaging token from each spatial-temporal window to serv
e as a proxy token for that region. The global semantics are captured through th
e self-attention of these proxy tokens and then injected into all latent tokens
via cross-attention. Simultaneously, we introduce window and shift window attent
ion to address the limitations in detail modeling caused by the sparse attention
 mechanism. Building on the well-designed PT-DiT, we further develop the PT-T2I/
V family, which includes a variety of models for T2I, T2V, and T2MV tasks. Exper
imental results show that PT-DiT achieves competitive performance while reducing
 computational complexity in image and video generation tasks (e.g., a reduction
 59\% compared to DiT and a reduction 34\% compared to PixArt-α). The vis
ual exhibition of and code are available at https://360cvgroup.github.io/Qihoo-T

2X/.
****
Mengqi Liao,Wei Chen,Junfeng Shen,Shengnan Guo,Huaiyu Wan
HMoRA: Making LLMs More Effective with Hierarchical Mixture of LoRA Experts
Recent studies have combined Mixture of Experts (MoE) and Parameter-Efficient Fine-tuning (PEFT) to fine-tune large language models (LLMs), holding excellent performance in multi-task scenarios while remaining resource-efficient. However, existing MoE approaches still exhibit the following limitations: (1) Current methods fail to consider that different LLM layers capture features at varying levels of granularity, leading to suboptimal performance. (2) Task-level routing methods lack generalizability to unseen tasks. (3) The uncertainty introduced by load imbalance loss undermines the effective specialization of the experts. To address these challenges, we propose HMoRA, a Hierarchical fine-tuning method that combines MoE and LoRA, employing hybrid routing that integrates token-level and task-level routing in a hierarchical manner. This hierarchical hybrid routing allows the model to more efficiently capture both fine-grained token information and broader task contexts. To improve the certainty of expert selection, a novel routing auxiliary loss is introduced. This auxiliary function also enhances the task router's ability to differentiate tasks and its generalization to unseen tasks. Additionally, several optional lightweight designs have been proposed to significantly reduce both the number of trainable parameters and computational costs. Experimental results demonstrate that HMoRA outperforms full fine-tuning across multiple NLP benchmarks, while fine-tuning only 3.9\% of the parameters. The code is available on: https://github.com/LiaoMengqi/HMoRA.
****
Mingyuan Zhou,Huangjie Zheng,Yi Gu,Zhendong Wang,Hai Huang
Adversarial Score identity Distillation: Rapidly Surpassing the Teacher in One Step
Score identity Distillation (SiD) is a data-free method that has achieved state-of-the-art performance in image generation by leveraging only a pretrained diffusion model, without requiring any training data. However, the ultimate performance of SiD is constrained by the accuracy with which the pretrained model captures the true data scores at different stages of the diffusion process. In this paper, we introduce SiDA (SiD with Adversarial Loss), which not only enhances generation quality but also improves distillation efficiency by incorporating real images and adversarial loss. SiDA utilizes the encoder from the generator's score network as a discriminator, allowing it to distinguish between real images and those generated by SiD. The adversarial loss is batch-normalized within each GPU and then combined with the original SiD loss. This integration effectively incorporates the average "fakeness" per GPU batch into the pixel-based SiD loss, enabling SiDA to distill a single-step generator. SiDA converges significantly faster than its predecessor when distilled from scratch, and swiftly improves upon the original model's performance during fine-tuning from a pre-distilled SiD generator. This one-step adversarial distillation method establishes new benchmarks in generation performance when distilling EDM diffusion models, achieving FID scores of **1.499** on CIFAR-10 unconditional, **1.396** on CIFAR-10 conditional, and **1.110** on ImageNet 64x64. When distilling EDM2 models trained on ImageNet 512x512, our SiDA method surpasses even the largest teacher model, EDM2-XXL, which achieved an FID of 1.81 using classifier-free guidance (CFG) and 63 generation steps. Specifically, SiDA achieves FID scores of **2.156** for size XS, **1.669** for S, **1.488** for M, **1.413** for L, **1.379** for XL, and **1.366** for XXL, all without CFG and in a single generation step. These results highlight substantial improvements across all model sizes. Our code and checkpoints are available at https://github.com/mingyuanzhou/SiD/tree/sida.
****
Arjun Roy,Kaushik Roy
DCT-CryptoNets: Scaling Private Inference in the Frequency Domain
The convergence of fully homomorphic encryption (FHE) and machine learning offers unprecedented opportunities for private inference of sensitive data. FHE enables computation directly on encrypted data, safeguarding the entire machine learn

ing pipeline, including data and model confidentiality. However, existing FHE-ba
sed implementations for deep neural networks face significant challenges in comp
utational cost, latency, and scalability, limiting their practical deployment. T
his paper introduces DCT-CryptoNets, a novel approach that operates directly in
the frequency-domain to reduce the burden of computationally expensive non-linea
r activations and homomorphic bootstrap operations during private inference. It
does so by utilizing the discrete cosine transform (DCT), commonly employed in J
PEG encoding, which has inherent compatibility with remote computing services wh
ere images are generally stored and transmitted in this encoded format. DCT-Cryp
toNets demonstrates a substantial latency reductions of up to 5.3$\times$ compar
ed to prior work on benchmark image classification tasks. Notably, it demonstrat
es inference on the ImageNet dataset within 2.5 hours (down from 12.5 hours on e
quivalent 96-thread compute resources). Furthermore, by *learning* perceptually
salient low-frequency information DCT-CryptoNets improves the reliability of enc
rypted predictions compared to RGB-based networks by reducing error accumulating
 homomorphic bootstrap operations. DCT-CryptoNets also demonstrates superior sca
lability to RGB-based networks by further reducing computational cost as image s
ize increases. This study demonstrates a promising avenue for achieving efficien
t and practical private inference of deep learning models on high resolution ima
ges seen in real-world applications.
****
Rheeya Uppaal,Apratim Dey,Yiting He,Yiqiao Zhong,Junjie Hu
Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity
Recent alignment algorithms such as direct preference optimization (DPO) have be
en developed to improve the safety of large language models (LLMs) by training t
hese models to match human behaviors exemplified by preference data. However, th
ese methods are
both computationally intensive and lacking in controllability and transparency,
inhibiting their widespread use. Furthermore, these tuning-based methods require
 large-scale preference data for training and are susceptible to noisy preferenc
e data. In this paper, we introduce a tuning-free alignment alternative, ProFS (
Projection Filter for Subspaces), and demonstrate its effectiveness under the us
e case of toxicity reduction. Grounded on theory from factor analysis, ProFS is
a sample-efficient model editing approach that identifies a toxic subspace in th
e model parameter space and reduces model toxicity by projecting away the detect
ed subspace. The toxic subspace is identified by extracting preference data embe
ddings from the language model, and removing non-toxic information from these em
beddings. We show that ProFS is more sample-efficient than DPO, further showcasi
ng greater robustness to noisy data. Finally, we attempt to connect tuning based
 alignment with editing, by establishing both theoretical and empirical connecti
ons between ProFS and DPO, showing that ProFS can be interpreted as a denoised v
ersion of a single DPO step.
****
Xiaoqiang Lin,Xinyi Xu,See-Kiong Ng,Bryan Kian Hsiang Low
Efficient Top-m Data Values Identification for Data Selection
Data valuation has found many real-world applications, e.g., data pricing and da
ta selection. However, the most adopted approach -- Shapley value (SV) -- is com
putationally expensive due to the large number of model trainings required. Fort
unately, most applications (e.g., data selection) require only knowing the m d
ata points with the highest data values (i.e., top-m data values), which impli
es the potential for fewer model trainings as exact data values are not required
. Existing work formulates top-m Shapley value identification as top-m arms
identification in multi-armed bandits (MAB). However, the proposed approach fall
s short because it does not utilize data features to predict data values, a meth
od that has been shown empirically to be effective. A recent top-m arms identi
fication work does consider the use of arm features while assuming a linear rela
tionship between arm features and rewards, which is often not satisfied in data
valuation. To this end, we propose the GPGapE algorithm that uses the Gaussian p
rocess to model the \emph{non-linear} mapping from data features to data values,
 removing the linear assumption. We theoretically analyze the correctness and st

opping iteration of GPGapE in finding an (ϵ, δ)-approximation to t
he top-m data values. We further improve the computational efficiency, by calc
ulating data values using small data subsets to reduce the computation cost of m
odel training. We empirically demonstrate that GPGapE outperforms other baseline
s in top-m data values identification, noisy data detection, and data subset s
election on real-world datasets. We also demonstrate the efficiency of our GPGap
E in data selection for large language model fine-tuning.
****
Anmol Goel,Yaxi Hu,Iryna Gurevych,Amartya Sanyal
Differentially Private Steering for Large Language Model Alignment
Aligning Large Language Models (LLMs) with human values and away from undesirabl
e behaviors (such as hallucination) has become increasingly important. Recently,
 steering LLMs towards a desired behavior via activation editing has emerged as
an effective method to mitigate harmful generations at inference-time. Activati
on editing modifies LLM representations by preserving information from positive
demonstrations (e.g., truthful) and minimising information from negative demonst
rations (e.g., hallucinations). When these demonstrations come from a private da
taset, the aligned LLM may leak private information contained in those private s
amples. In this work, we present the first study of aligning LLM behavior with p
rivate datasets. Our work proposes the \textit{\underline{P}rivate \underline{S}
teering for LLM \underline{A}lignment (PSA)} algorithm to edit LLM activations w
ith differential privacy (DP) guarantees. We conduct extensive experiments on se
ven different benchmarks with open-source LLMs of different sizes (0.5B to 7B) a
nd model families (LlaMa and Qwen). Our results show that PSA achieves DP guaran
tees for LLM alignment with minimal loss in performance, including alignment met
rics, open-ended text generation quality, and general-purpose reasoning. We also
 develop the first Membership Inference Attack (MIA) for evaluating and auditing
 the empirical privacy for the problem of LLM steering via activation editing. O
ur attack is tailored for activation editing and relies solely on the generated
texts without their associated probabilities. Our experiments support the theore
tical guarantees by showing improved guarantees for our \textit{PSA} algorithm c
ompared to several existing non-private techniques.
****
Junyi Zhang,Charles Herrmann,Junhwa Hur,Varun Jampani,Trevor Darrell,Forrester C
ole,Deqing Sun,Ming-Hsuan Yang
MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion
Estimating geometry from dynamic scenes, where objects move and deform over time
, remains a core challenge in computer vision. Current approaches often rely on
multi-stage pipelines or global optimizations that decompose the problem into su
btasks, like depth and flow, leading to complex systems prone to errors. In this
 paper, we present Motion DUSt3R (MonST3R), a novel geometry-first approach
that directly estimates per-timestep geometry from dynamic scenes. Our key insig
ht is that by simply estimating a pointmap for each timestep, we can effectively
 adapt DUSt3R’s representation, previously only used for static scenes, to dynam
ic scenes. However, this approach presents a significant challenge: the scarcity
 of suitable training data, namely dynamic, posed videos with depth labels. Desp
ite this, we show that by posing the problem as a fine-tuning task, identifying
several suitable datasets, and strategically training the model on this limited
data, we can surprisingly enable the model to handle dynamics, even without an e
xplicit motion representation. Based on this, we introduce new optimizations for
 several downstream video-specific tasks and demonstrate strong performance on v
ideo depth and camera pose estimation, outperforming prior work in terms of robu
stness and efficiency. Moreover, MonST3R shows promising results for primarily f
eed-forward 4D reconstruction. Interactive 4D results, source code, and trained
models are available at: https://monst3r-project.github.io/.
****
Geoffrey Cideron,Andrea Agostinelli,Johan Ferret,Sertan Girgin,Romuald Elie,Oliv
ier Bachem,Sarah Perrin,Alexandre Rame
Diversity-Rewarded CFG Distillation
Generative models are transforming creative domains such as music generation, wi

th inference-time strategies like Classifier-Free Guidance (CFG) playing a cruci
al role. However, CFG doubles inference cost while limiting originality and dive
rsity across generated contents. In this paper, we introduce diversity-rewarded
CFG distillation, a novel finetuning procedure that distills the strengths of CF
G while addressing its limitations. Our approach optimises two training objectiv
es: (1) a distillation objective, encouraging the model alone (without CFG) to
imitate the CFG-augmented predictions, and (2) an RL objective with a diversity
reward, promoting the generation of diverse outputs for a given prompt. By finet
uning, we learn model weights with the ability to generate high-quality and dive
rse outputs, without any inference overhead. This also unlocks the potential of
weight-based model merging strategies: by interpolating between the weights of t
wo models (the first focusing on quality, the second on diversity), we can contr
ol the quality-diversity trade-off at deployment time, and even further boost pe
rformance. We conduct extensive experiments on the MusicLM text-to-music generat
ive model, where our approach surpasses CFG in terms of quality-diversity Pareto
optimality. According to human evaluators, our finetuned-then-merged model gen
erates samples with higher quality-diversity than the base model augmented with
CFG. Explore our generations at https://musicdiversity.github.io/.
****

Saaket Agashe,Jiuzhou Han,Shuyu Gan,Jiachen Yang,Ang Li,Xin Eric Wang
Agent S: An Open Agentic Framework that Uses Computers Like a Human
We present Agent S, an open agentic framework that enables autonomous interactio
n with computers through Graphical User Interface (GUI), aimed at transforming h
uman-computer interaction by automating complex, multi-step tasks. Agent S addre
sses three key challenges in automating computer tasks: acquiring domain-specifi
c knowledge, planning over long task horizons, and handling dynamic, non-uniform
interfaces. To this end, Agent S introduces experience-augmented hierarchical p
lanning, which learns from external knowledge search and internal experience ret
rieval at multiple levels, facilitating efficient task planning and subtask exec
ution.
In addition, it employs an Agent-Computer Interface (ACI) to better elicit the r
easoning and control capabilities of GUI agents based on Multimodal Large Langua
ge Models (MLLMs). Evaluation on the OSWorld benchmark shows that Agent S outper
forms the baseline by 9.37\% on success rate (an 83.6\% relative improvement) an
d achieves a new state-of-the-art. Comprehensive analysis highlights the effecti
veness of individual components and provides insights for future improvements. F
urthermore, Agent S demonstrates broad generalizability to different operating s
ystems on a newly-released WindowsAgentArena benchmark. Code available at https:
//github.com/simular-ai/Agent-S.
****

Amin Abyaneh,Mahrokh Ghoddousi Boroujeni,Hsiu-Chin Lin,Giancarlo Ferrari-Trecate
Contractive Dynamical Imitation Policies for Efficient Out-of-Sample Recovery
Imitation learning is a data-driven approach to learning policies from expert be
havior, but it is prone to unreliable outcomes in out-of-sample (OOS) regions. W
hile previous research relying on stable dynamical systems guarantees convergenc
e to a desired state, it often overlooks transient behavior. We propose a framew
ork for learning policies modeled by contractive dynamical systems, ensuring tha
t all policy rollouts converge regardless of perturbations, and in turn, enable
efficient OOS recovery. By leveraging recurrent equilibrium networks and couplin
g layers, the policy structure guarantees contractivity for any parameter choice
, which facilitates unconstrained optimization. We also provide theoretical uppe
r bounds for worst-case and expected loss to rigorously establish the reliabilit
y of our method in deployment. Empirically, we demonstrate substantial OOS perfo
rmance improvements for simulated robotic manipulation and navigation tasks. See
[sites.google.com/view/contractive-dynamical-policies](https://sites.google.com
/view/contractive-dynamical-policies) for our codebase and highlight of the resu
lts.
****

Yucheng Shi,Quanzheng Li,Jin Sun,Xiang Li,Ninghao Liu
Enhancing Cognition and Explainability of Multimodal Foundation Models with Self

-Synthesized Data
Large Multimodal Models (LMMs), or Vision-Language Models (VLMs), have shown imp
ressive capabilities in a wide range of visual tasks. However, they often strugg
le with fine-grained visual reasoning, failing to identify domain-specific objec
tives and provide justifiable explanations for their predictions. To address the
 above challenge, we propose a novel visual rejection sampling framework to impr
ove the cognition and explainability of LMMs using self-synthesized data. Specif
ically, visual fine-tuning requires images, queries, and target answers. Our app
roach begins by synthesizing interpretable answers that include human-verifiable
 visual features. These features are based on expert-defined concepts, and caref
ully selected based on their alignment with the image content. After each round
of fine-tuning, we apply a reward model-free filtering mechanism to select the h
ighest-quality interpretable answers for the next round of tuning. This iterativ
e process of synthetic data generation and fine-tuning progressively improves th
e model's ability to generate accurate and reasonable explanations. Experimental
 results demonstrate the effectiveness of our method in improving both the accur
acy and explainability of specialized visual classification tasks.
****
Zhiwei Zhang,Fali Wang,Xiaomin Li,Zongyu Wu,Xianfeng Tang,Hui Liu,Qi He,Wenpeng
Yin,Suhang Wang
Catastrophic Failure of LLM Unlearning via Quantization
Large language models (LLMs) have shown remarkable proficiency in generating tex
t, benefiting from extensive training on vast textual corpora. However, LLMs may
 also acquire unwanted behaviors from the diverse and sensitive nature of their
training data, which can include copyrighted and private content. Machine unlear
ning has been introduced as a viable solution to remove the influence of such pr
oblematic content without the need for costly and time-consuming retraining. Thi
s process aims to erase specific knowledge from LLMs while preserving as much mo
del utility as possible. Despite the effectiveness of current unlearning methods
, little attention has been given to whether existing unlearning methods for LLM
s truly achieve forgetting or merely hide the knowledge, which current unlearnin
g benchmarks fail to detect. This paper reveals that applying quantization to mo
dels that have undergone unlearning can restore the "forgotten" information. We
conduct comprehensive experiments using various quantization techniques across m
ultiple precision levels to thoroughly evaluate this phenomenon. We find that fo
r unlearning methods with utility constraints, the unlearned model retains an av
erage of 21\% of the intended forgotten knowledge in full precision, which signi
ficantly increases to 83\% after 4-bit quantization. Based on our empirical find
ings, we provide a theoretical explanation for the observed phenomenon and propo
se a quantization-robust unlearning strategy aimed at mitigating this intricate
issue. Our results highlight a fundamental tension between preserving the utilit
y of the unlearned model and preventing knowledge recovery through quantization,
 emphasizing the challenge of balancing these two objectives. Altogether, our st
udy underscores a major failure in existing unlearning methods for LLMs, strongl
y advocating for more comprehensive and robust strategies to ensure authentic un
learning without compromising model utility. Our code is available at: https://g
ithub.com/zzwjames/FailureLLMUnlearning.
****
Rui Zhang,Ruixu Geng,Yadong Li,Ruiyuan Song,Hanqin Gong,Dongheng Zhang,Yang Hu,Y
an Chen
RFMamba: Frequency-Aware State Space Model for RF-Based Human-Centric Perception
Human-centric perception with radio frequency (RF) signals has recently entered
a new era of end-to-end processing with Transformers. Considering the long-seque
nce nature of RF signals, the State Space Model (SSM) has emerged as a superior
alternative due to its effective long-sequence modeling and linear complexity. H
owever, integrating SSM into RF-based sensing presents unique challenges includi
ng the fundamentally different signal representation, distinct frequency respons
es in different scenarios, and incomplete capture caused by specular reflection.
 To address this, we carefully devise a dual-branch SSM block that is characteri
zed by adaptively grasping the most informative frequency cues and the assistant

spatial information to fully explore the human representations from radar echoes. Based on these two branchs, we further introduce an SSM-based network for handling various downstream human perception tasks, named RFMamba. Extensive experimental results demonstrate the superior performance of our proposed RFMamba across all three downstream tasks. To the best of our knowledge, RFMamba is the first attempt to introduce SSM into RF-based human-centric perception.
****
Letitia Parcalabescu,Anette Frank
Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations?
Vision and language model (VLM) decoders are currently the best-performing architectures on multimodal tasks. Next to answers, they are able to produce natural language explanations, either in post-hoc or CoT settings. However, it is not clear to what extent they are using the input vision and text modalities when generating answers or explanations.
In this work, we investigate if VLMs rely on their input modalities differently when they produce explanations as opposed to answers.
We also evaluate the self-consistency of VLM decoders in both post-hoc and CoT explanation settings, by extending existing unimodal tests and measures to VLM decoders. We find that most tested VLMs are less self-consistent than LLMs. Text contributions in all tested VL decoders are more important than image contributions in all examined tasks. However, when comparing explanation generation to answer generation, the contributions of images are significantly stronger for generating explanations compared to answers. This difference is even larger in CoT compared to post-hoc explanations.
Lastly, we provide an up-to-date benchmarking of state-of-the-art VL decoders on the VALSE benchmark, which before was restricted to VL encoders. We find that the tested VL decoders still struggle with most phenomena tested by VALSE. We will make our code publicly available.
****
Xinghao Wang,Pengyu Wang,Bo Wang,Dong Zhang,Yunhua Zhou,Xipeng Qiu
BitStack: Any-Size Compression of Large Language Models in Variable Memory Environments
Large language models (LLMs) have revolutionized numerous applications, yet their deployment remains challenged by memory constraints on local devices. While scaling laws have enhanced LLM capabilities, the primary bottleneck has shifted from $\textit{capability}$ to $\textit{availability}$, emphasizing the need for efficient memory management. Traditional compression methods, such as quantization, often require predefined compression ratios and separate compression processes for each setting, complicating deployment in variable memory environments. In this paper, we introduce $\textbf{BitStack}$, a novel, training-free weight compression approach that enables megabyte-level trade-offs between memory usage and model performance. By leveraging weight decomposition, BitStack can dynamically adjust the model size with minimal transmission between running memory and storage devices. Our approach iteratively decomposes weight matrices while considering the significance of each parameter, resulting in an approximately 1-bit per parameter residual block in each decomposition iteration. These blocks are sorted and stacked in storage as basic transmission units, with different quantities loaded based on current memory availability. Extensive experiments across a wide range of tasks demonstrate that, despite offering fine-grained size control, BitStack consistently matches or surpasses strong quantization baselines, particularly at extreme compression ratios. To the best of our knowledge, this is the first decomposition-based method that effectively bridges the gap to practical compression techniques like quantization. Code is available at https://github.com/xinghaow99/BitStack.
****
Feng Ji,Yanan Zhao,Kai Zhao,Hanyang Meng,Jielong Yang,Wee Peng Tay
Rethinking Graph Neural Networks From A Geometric Perspective Of Node Features
Many works on graph neural networks (GNNs) focus on graph topologies and analyze graph-related operations to enhance performance on tasks such as node classific

ation. In this paper, we propose to understand GNNs based on a feature-centric a
pproach. Our main idea is to treat the features of nodes from each label class a
s a whole, from which we can identify the centroid. The convex hull of these cen
troids forms a simplex called the feature centroid simplex, where a simplex is a
 high-dimensional generalization of a triangle. We borrow ideas from coarse geom
etry to analyze the geometric properties of the feature centroid simplex by comp
aring them with basic geometric models, such as regular simplexes and degenerate
 simplexes. Such a simplex provides a simple platform to understand graph-based
feature aggregation, including phenomena such as heterophily, oversmoothing, and
 feature re-shuffling. Based on the theory, we also identify simple and useful t
ricks for the node classification task.
****

Jong-Hoon Ahn,Akshay Vashist

Gaussian Mixture Counterfactual Generator

We address the individualized treatment effect (ITE) estimation problem, focusin
g on continuous, multidimensional, and time-dependent treatments for precision m
edicine. The central challenge lies in modeling these complex treatment scenario
s while capturing dynamic patient responses and minimizing reliance on control d
ata. We propose the Gaussian Mixture Counterfactual Generator (GMCG), a generati
ve model that transforms the Gaussian mixture model—traditionally a tool for clu
stering and density estimation—into a new tool explicitly geared toward causal i
nference. This approach generates robust counterfactuals by effectively handling
 continuous and multidimensional treatment spaces. We evaluate GMCG on synthetic
 crossover trial data and simulated datasets, demonstrating its superior perform
ance over existing methods, particularly in scenarios with limited control data.
 GMCG derives its effectiveness from modeling the joint distribution of covariat
es, treatments, and outcomes using a latent state vector while employing a condi
tional distribution of the state vector to suppress confounding and isolate trea
tment-outcome relationships. GMCG shows promise for enhancing ITE estimation in
precision medicine, offering a potential unified solution for personalized thera
peutic strategies.
****

Matthew Fortier,Mats Leon Richter,Oliver Sonnentag,Christopher Pal

CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling

Terrestrial carbon fluxes provide vital information about our biosphere's health
 and its capacity to absorb anthropogenic CO$_2$
 emissions. The importance of predicting carbon fluxes has led to the emerging f
ield of data-driven carbon flux modelling (DDCFM), which uses statistical techni
ques to predict carbon fluxes from biophysical data. However, the field lacks a
standardized dataset to promote comparisons between models. To address this gap,
 we present CarbonSense, the first machine learning-ready dataset for DDCFM. Car
bonSense integrates measured carbon fluxes, meteorological predictors, and satel
lite imagery from 385 locations across the globe, offering comprehensive coverag
e and facilitating robust model training. Additionally, we provide a baseline mo
del using a current state-of-the-art DDCFM approach and a novel transformer base
d model. Our experiments illustrate the potential gains that multimodal deep lea
rning techniques can bring to this domain. By providing these resources, we aim
to lower the barrier to entry for other deep learning researchers to develop new
 models and drive new advances in carbon flux modelling.
****

Pouya Hamadanian,Arash Nasr-Esfahany,Malte Schwarzkopf,Siddhartha Sen,Mohammad A
lizadeh

Online Reinforcement Learning in Non-Stationary Context-Driven Environments

We study online reinforcement learning (RL) in non-stationary environments, wher
e a time-varying exogenous context process affects the environment dynamics. Onl
ine RL is challenging in such environments due to "catastrophic forgetting" (CF)
. The agent tends to forget prior knowledge as it trains on new experiences. Pri
or approaches to mitigate this issue assume task labels (which are often not ava
ilable in practice), employ brittle regularization heuristics or use off-policy
methods that suffer from instability and poor performance.

We present Locally Constrained Policy Optimization (LCPO), an online RL approach that combats CF by anchoring policy outputs on old experiences while optimizing the return on current experiences. To perform this anchoring, LCPO locally constrains policy optimization using samples from experiences that lie outside of the current context distribution. We evaluate LCPO in Mujoco, classic control and computer systems environments with a variety of synthetic and real context traces, and find that it outperforms a variety of baselines in the non-stationary setting, while achieving results on-par with a "prescient" agent trained offline across all context traces.

LCPO's source code is available at https://github.com/pouyahmdn/LCPO.
****
Soham Deshmukh,Shuo Han,Rita Singh,Bhiksha Raj
ADIFF: Explaining audio difference using natural language
Understanding and explaining differences between audio recordings is crucial for fields like audio forensics, quality assessment, and audio generation. This involves identifying and describing audio events, acoustic scenes, signal characteristics, and their emotional impact on listeners. This paper stands out as the first work to comprehensively study the task of explaining audio differences and then propose benchmark, baselines for the task. First, we present two new datasets for audio difference explanation derived from the AudioCaps and Clotho audio captioning datasets. Using Large Language Models (LLMs), we generate three levels of difference explanations: (1) concise descriptions of audio events and objects, (2) brief sentences about audio events, acoustic scenes, and signal properties, and (3) comprehensive explanations that include semantics and listener emotions. For the baseline, we use prefix tuning where audio embeddings from two audio files are used to prompt a frozen language model. Our empirical analysis and ablation studies reveal that the naive baseline struggles to distinguish perceptually similar sounds and generate detailed tier 3 explanations. To address these limitations, we propose ADIFF, which introduces a cross-projection module, position captioning, and a three-step training process to enhance the model’s ability to produce detailed explanations. We evaluate our model using objective metrics and human evaluation and show our model enhancements lead to significant improvements in performance over naive baseline and SoTA Audio-Language Model (ALM) Qwen Audio. Lastly, we conduct multiple ablation studies to study the effects of cross-projection, language model parameters, position captioning, third stage fine-tuning, and present our findings. Our benchmarks, findings, and strong baseline pave the way for nuanced and human-like explanations of audio differences.
****
SungUk Shin,Youngjoon Kim
Enhancing Graph Of Thought: Enhancing Prompts with LLM Rationales and Dynamic Temperature Control
We introduce Enhancing Graph of Thoughts (EGoT), a method designed to enhance the performance of large language models (LLMs) on complex reasoning tasks. EGoT automates the process of generating accurate responses using given data and a base prompt. The process consists of several steps: It obtains an initial response from the answering node using the base prompt. Evaluation node evaluates the response and generates reasoning for it, utilizing the score's probabilities to enhance evaluation accuracy. The reasoning from both the answering node and the evaluation node is aggregated to identify the problem in the response. This aggregated reasoning is incorporated into the base prompt to obtain an enhanced response. These steps are organized in a graph architecture, where the final leaf nodes are merged to produce a final response. As the graph descends, the temperature is lowered using Cosine Annealing and scoring, to explore diverse responses with earlier nodes and to focus on precise responses with later nodes. The minimum temperature in Cosine Annealing is adjusted based on scoring, ensuring that nodes with low scores continue to explore diverse responses, while those with high scores confirm accurate responses. In sorting 256 elements using GPT-4o mini, EGoT performs 88.31\% accuracy, while GoT (Graph of Thoughts) achieves 84.37\% accur

acy. In the frozen lake problem using GPT-4o, EGoT averages 0.55 jumps or falls into the hole, while ToT (Tree of Thoughts) averages 0.89.
****
Muxi Chen,Chenchen Zhao,Qiang Xu
HiBug2: Efficient and Interpretable Error Slice Discovery for Comprehensive Model Debugging
Despite the significant success of deep learning models in computer vision, they often exhibit systematic failures on specific data subsets, known as error slices. Identifying and mitigating these error slices is crucial to enhancing model robustness and reliability in real-world scenarios. In this paper, we introduce HiBug2, an automated framework for error slice discovery and model repair. HiBug2 first generates task-specific visual attributes to highlight instances prone to errors through an interpretable and structured process. It then employs an efficient slice enumeration algorithm to systematically identify error slices, overcoming the combinatorial challenges that arise during slice exploration. Additionally, HiBug2 extends its capabilities by predicting error slices beyond the validation set, addressing a key limitation of prior approaches. Extensive experiments across multiple domains — including image classification, pose estimation, and object detection — show that HiBug2 not only improves the coherence and precision of identified error slices but also significantly enhances the model repair capabilities.
****
Pau Rodriguez,Arno Blaas,Michal Klein,Luca Zappella,Nicholas Apostoloff,marco cuturi,Xavier Suau
Controlling Language and Diffusion Models by Transporting Activations
The increasing capabilities of large generative models and their ever more widespread deployment have raised concerns about their reliability, safety, and potential misuse. To address these issues, recent works have proposed to control model generation by steering model activations in order to effectively induce or prevent the emergence of concepts or behaviors in the generated output.
In this paper we introduce Activation Transport (AcT), a general framework to steer activations guided by optimal transport theory that generalizes many previous activation-steering works. AcT is modality-agnostic and provides fine-grained control over the model behavior with negligible computational overhead, while minimally impacting model abilities. We experimentally show the effectiveness and versatility of our approach by addressing key challenges in large language models (LLMs) and text-to-image diffusion models (T2Is). For LLMs, we show that AcT can effectively mitigate toxicity, induce arbitrary concepts, and increase their truthfulness. In T2Is, we show how AcT enables fine-grained style control and concept negation.
****
Zehan Wang,Ziang Zhang,Minjie Hong,Hang Zhang,Luping Liu,Rongjie Huang,Xize Cheng,Shengpeng Ji,Tao Jin,Hengshuang Zhao,Zhou Zhao
OmniBind: Large-scale Omni Multimodal Representation via Binding Spaces
Recently, human-computer interaction with various modalities has shown promising applications, like GPT-4o and Gemini. Meanwhile, multimodal representation models have emerged as the foundation for these versatile multimodal understanding and generation pipeline. Models like CLIP, CLAP and ImageBind can map their specialized modalities into respective joint spaces. To construct a high-quality omni representation space that can be shared and expert in any modality, we propose to merge these advanced models into a unified space in scale. With this insight, we present \textbf{OmniBind}, advanced multimodal joint representation models via fusing knowledge of 14 pre-trained spaces, which support 3D, audio, image, video and language inputs. To alleviate the interference between different knowledge sources in integrated space, we dynamically assign weights to different spaces by learning routers with two objectives: cross-modal overall alignment and language representation decoupling. Notably, since binding and routing spaces only require lightweight networks, OmniBind is extremely training-efficient. Extensive experiments demonstrate the versatility and superiority of OmniBind as an omni representation model, highlighting its great potential for diverse applications

, such as any-query and composable multimodal understanding.
****
Lokesh Nagalapatti,Ashutosh Srivastava,Sunita Sarawagi,Amit Sharma
Robust Root Cause Diagnosis using In-Distribution Interventions
Diagnosing the root cause of an anomaly in a complex interconnected system is a pressing problem in today's cloud services and industrial operations. We propose In-Distribution Interventions (IDI), a novel algorithm that predicts root cause
as nodes that meet two criteria: 1) Anomaly: root cause nodes should take on anomalous values; 2) Fix: had the root cause nodes assumed usual values, the target node would not have been anomalous. Prior methods of assessing the fix condition rely on counterfactuals inferred from a Structural Causal Model (SCM) trained on historical data. But since anomalies are rare and fall outside the training distribution, the fitted SCMs yield unreliable counterfactual estimates. IDI
overcomes this by relying on interventional estimates obtained by solely probing the fitted SCM at in-distribution inputs. We present a theoretical analysis comparing and bounding the errors in assessing the fix condition using interventional and counterfactual estimates. We then conduct experiments by systematically varying the SCM's complexity to demonstrate the cases where IDI's interventional approach outperforms the counterfactual approach and vice versa.
Experiments on both synthetic and PetShop RCD benchmark datasets demonstrate that IDI consistently identifies true root causes more accurately and robustly than nine existing state-of-the-art RCD baselines. Code will be released
at https://github.com/nlokeshiisc/IDI_release.
****
Jixuan Leng,Chengsong Huang,Banghua Zhu,Jiaxin Huang
Taming Overconfidence in LLMs: Reward Calibration in RLHF
Language model calibration refers to the alignment between the confidence of the model and the actual performance of its responses.
While previous studies point out the overconfidence phenomenon in Large Language Models (LLMs) and show that LLMs trained with Reinforcement Learning from Human Feedback (RLHF) are overconfident with a more sharpened output probability, in this study, we reveal that RLHF tends to lead models to express verbalized overconfidence in their own responses. We investigate the underlying cause of this overconfidence and demonstrate that reward models used for Proximal Policy Optimization (PPO) exhibit inherent biases towards high-confidence scores regardless of the actual quality of responses. Building upon this insight, we propose two PPO variants: PPO-M: $\underline{PPO}$ with Calibrated Reward $\underline{M}$odeling and PPO-C: $\underline{PPO}$ with Calibrated Reward $\underline{C}$alculation. PPO-M integrates explicit confidence scores in reward model training, which calibrates reward models
to better capture the alignment between response quality and verbalized confidence. PPO-C adjusts the reward score during PPO based on the difference between the current reward and the moving average of past rewards. Both PPO-M and PPO-C can be seamlessly integrated into the current PPO pipeline and do not require additional golden labels. We evaluate our methods on both $\texttt{Llama3-8B}$ and $\texttt{Mistral-7B}$ across six diverse datasets including multiple-choice and open-ended generation. Experiment results demonstrate that both of our methods can reduce calibration error and maintain performance comparable to standard PPO. We further show that they do not compromise model capabilities in open-ended conversation settings.
****
Yuguang Yang,Tongfei Chen,Haoyu Huang,Linlin Yang,Chunyu Xie,Dawei Leng,Xianbin Cao,Baochang Zhang
Prompt as Knowledge Bank: Boost Vision-language model via Structural Representation for zero-shot medical detection
Zero-shot medical detection can further improve detection performance without relying on annotated medical images even upon the fine-tuned model, showing great clinical value. Recent studies leverage grounded vision-language models (GLIP) t

o achieve this by using detailed disease descriptions as prompts for the target disease name during the inference phase.
However, these methods typically treat prompts as equivalent context to the targ et name, making it difficult to assign specific disease knowledge based on visua l information, leading to a coarse alignment between images and target descripti ons. In this paper, we propose StructuralGLIP, which introduces an auxiliary bra nch to encode prompts into a latent knowledge bank layer-by-layer, enabling more context-aware and fine-grained alignment. Specifically, in each layer, we selec t highly similar features from both the image representation and the knowledge b ank, forming structural representations that capture nuanced relationships betwe en image patches and target descriptions. These features are then fused across m odalities to further enhance detection performance.
Extensive experiments demonstrate that StructuralGLIP achieves a +4.1\% AP impro vement over prior state-of-the-art methods across seven zero-shot medical detect ion benchmarks, and consistently improves fine-tuned models by +3.2\% AP on endo scopy image datasets.
****

Yinghao Li,Vianne R. Gao,Chao Zhang,MohamadAli Torkamani
Ensembles of Low-Rank Expert Adapters
The training and fine-tuning of large language models (LLMs) often involve diver se textual data from multiple sources, which poses challenges due to conflicting gradient directions, hindering optimization and specialization. These challeng es can undermine model generalization across tasks, resulting in reduced downstr eam performance. Recent research suggests that fine-tuning LLMs on carefully se lected, task-specific subsets of data can match or even surpass the performance of using the entire dataset. Building on these insights, we propose the Ensembl es of Low-Rank Expert Adapters (ELREA) framework to improve the model's capabili ty to handle diverse tasks. ELREA clusters the training instructions based on t heir gradient directions, representing different areas of expertise and thereby reducing conflicts during optimization. Expert adapters are then trained on the se clusters, utilizing the low-rank adaptation (LoRA) technique to ensure traini ng efficiency and model scalability. During inference, ELREA combines predictio ns from the most relevant expert adapters based on the input data's gradient sim ilarity to the training clusters, ensuring optimal adapter selection for each ta sk. Experiments show that our method outperforms baseline LoRA adapters trained on the full dataset and other ensemble approaches with similar training and inf erence complexity across a range of domain-specific tasks.
****

Jiawei Zhou,Li Dong,Furu Wei,Lei Chen
Semi-Parametric Retrieval via Binary Bag-of-Tokens Index
Information retrieval has transitioned from standalone systems into essential co mponents across broader applications, with indexing efficiency, cost-effectivene ss, and freshness becoming increasingly critical yet often overlooked. In this p aper, we introduce SemI-parametric Disentangled Retrieval (SiDR), a bi-encoder r etrieval framework that decouples retrieval index from neural parameters to enab le efficient, low-cost, and parameter-agnostic indexing for emerging use cases. Specifically, in addition to using embeddings as indexes like existing neural re trieval methods, SiDR supports a non-parametric tokenization index for search, a chieving BM25-like indexing complexity with significantly better effectiveness. Our comprehensive evaluation across 16 retrieval benchmarks demonstrates that Si DR outperforms both neural and term-based retrieval baselines under the same ind exing workload: (i) When using an embedding-based index, SiDR exceeds the perfor mance of conventional neural retrievers while maintaining similar training compl exity; (ii) When using a tokenization-based index, SiDR drastically reduces inde xing cost and time, matching the complexity of traditional term-based retrieval, while consistently outperforming BM25 on all in-domain datasets; (iii) Addition ally, we introduce a late parametric mechanism that matches BM25 index preparati on time while outperforming other neural retrieval baselines in effectiveness.
****

Korbinian Pöppel,Maximilian Beck,Sepp Hochreiter

FlashRNN: I/O-Aware Optimization of Traditional RNNs on modern hardware
While Transformers and other sequence-parallelizable neural network architecture s seem like the current state of the art in sequence modeling, they specifically lack state-tracking capabilities. These are important for time-series tasks and logical reasoning. Traditional RNNs like LSTMs and GRUs, as well as modern vari ants like sLSTM do have these capabilities at the cost of strictly sequential pr ocessing. While this is often seen as a strong limitation, we show how fast thes e networks can get with our hardware-optimization FlashRNN in Triton and CUDA, o ptimizing kernels to the register level on modern GPUs. We extend traditional RN Ns with a parallelization variant that processes multiple RNNs of smaller hidden state in parallel, similar to the head-wise processing in Transformers. To enab le flexibility on different GPU variants, we introduce a new optimization framew ork for hardware-internal cache sizes, memory and compute handling. It models th e hardware in a setting using polyhedral-like constraints, including the notion of divisibility. This speeds up the solution process in our ConstrINT library fo r general integer constraint satisfaction problems (integer CSPs).
We show that our kernels can achieve 50x speed-ups over a vanilla PyTorch implem entation and allow 40x larger hidden sizes compared to our Triton implementation . We have open-sourced our kernels and the optimization library to boost researc h in the direction of state-tracking enabled RNNs and sequence modeling here: ht tps://github.com/NX-AI/flashrnn
****
Laetitia Chapel,Romain Tavenard
One for all and all for one: Efficient computation of partial Wasserstein distan ces on the line
Partial Wasserstein helps overcoming some of the limitations of Optimal Transpor t when the distributions at stake differ in mass, contain noise or outliers or e xhibit mass mismatches across distribution modes.
We introduce PAWL, a novel algorithm designed to efficiently compute exact PArti al Wasserstein distances on the Line. PAWL not only solves the partial transport ation problem for a specified amount of mass to be transported, but _for all_ ad missible mass amounts. This flexibility is valuable for machine learning tasks w here the level of noise is uncertain and needs to be determined through cross-va lidation, for example.
By achieving $O(n \log n)$ time complexity for the partial 1-Wasserstein problem on the line, it enables practical applications with large scale datasets.
Additionally, we introduce a novel slicing strategy tailored to Partial Wasserst ein, which does not permit transporting mass between outliers or noisy data poin ts. We demonstrate the advantages of PAWL in terms of computational efficiency a nd performance in downstream tasks, outperforming existing (sliced) Partial Opti mal Transport techniques.
****
Jingxuan Xu,Wuyang Chen,Linyi Li,Yao Zhao,Yunchao Wei
Collapsed Language Models Promote Fairness
To mitigate societal biases implicitly encoded in recent successful pretrained l anguage models, a diverse array of approaches have been proposed to encourage mo del fairness, focusing on prompting, data augmentation, regularized fine-tuning, and more. Despite the development, it is nontrivial to reach a principled under standing of fairness and an effective algorithm that can consistently debias lan guage models. In this work, by rigorous evaluations of Neural Collapse -- a lear ning phenomenon happen in last-layer representations and classifiers in deep net works -- on fairness-related words, we find that debiased language models exhibi t collapsed alignment between token representations and word embeddings. More im portantly, this observation inspires us to design a principled fine-tuning metho d that can effectively improve fairness in a wide range of debiasing methods, wh ile still preserving the performance of language models on standard natural lang uage understanding tasks. We attach our code at https://github.com/Xujxyang/Fair ness-NC-main
****
James Enouen,Yan Liu

InstaSHAP: Interpretable Additive Models Explain Shapley Values Instantly
In recent years, the Shapley value and SHAP explanations have emerged as one of the most dominant paradigms for providing post-hoc explanations of blackbox models. Despite their well-founded theoretical properties, many recent works have focused on the limitations in both their computational efficiency and their representation power. The underlying connection with additive models, however, is left critically under-emphasized in the current literature. In this work, we find
that a variational perspective linking GAM models and SHAP explanations is able to provide deep insights into nearly all recent developments. In light of this connection, we borrow in the other direction to develop a new method to train interpretable GAM models which are automatically purified to compute the Shapley value in a single forward pass. Finally, we provide theoretical results showing the
limited representation power of GAM models is the same Achilles' heel existing in SHAP and discuss the implications for SHAP's modern usage in CV and NLP.
****
Boyu Gou,Ruohan Wang,Boyuan Zheng,Yanan Xie,Cheng Chang,Yiheng Shu,Huan Sun,Yu Su
Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents
Multimodal large language models (MLLMs) are transforming the capabilities of graphical user interface (GUI) agents, facilitating their transition from controlled simulations to complex, real-world applications across various platforms. However, the effectiveness of these agents hinges on the robustness of their grounding capability. Current GUI agents predominantly utilize text-based representations such as HTML or accessibility trees, which, despite their utility, often introduce noise, incompleteness, and increased computational overhead. In this paper, we advocate a human-like embodiment for GUI agents that perceive the environment entirely visually and directly perform pixel-level operations on the GUI. The key is visual grounding models that can accurately map diverse referring expressions of GUI elements to their coordinates on the GUI across different platforms. We show that a simple recipe, which includes web-based synthetic data and slight adaptation of the LLaVA architecture, is surprisingly effective for training such visual grounding models. We collect the largest dataset for GUI visual grounding so far, containing 10M GUI elements and their referring expressions over 1.3M screenshots, and use it to train UGround, a strong universal visual grounding model for GUI agents. Empirical results on six benchmarks spanning three categories (grounding, offline agent, and online agent) show that 1) UGround substantially outperforms existing visual grounding models for GUI agents, by up to 20\% absolute, and 2) agents with UGround outperform state-of-the-art agents, despite the fact that existing agents use additional text-based input while ours only uses visual perception. These results provide strong support for the feasibility and promises of GUI agents that navigate the digital world as humans do.
****
Chunming He,Chengyu Fang,Yulun Zhang,Longxiang Tang,Jinfa Huang,Kai Li,Zhenhua Guo,Xiu Li,Sina Farsiu
Reti-Diff: Illumination Degradation Image Restoration with Retinex-based Latent Diffusion Model
Illumination degradation image restoration (IDIR) techniques aim to improve the visibility of degraded images and mitigate the adverse effects of deteriorated illumination. Among these algorithms, diffusion-based models (DM) have shown promising performance but are often burdened by heavy computational demands and pixel misalignment issues when predicting the image-level distribution. To tackle these problems, we propose to leverage DM within a compact latent space to generate concise guidance priors and introduce a novel solution called Reti-Diff for the IDIR task. Specifically, Reti-Diff comprises two significant components: the Retinex-based latent DM (RLDM) and the Retinex-guided transformer (RGformer). RLDM is designed to acquire Retinex knowledge, extracting reflectance and illumination priors to facilitate detailed reconstruction and illumination correction. RG

former subsequently utilizes these compact priors to guide the decomposition of image features into their respective reflectance and illumination components. Following this, RGformer further enhances and consolidates these decomposed features, resulting in the production of refined images with consistent content and robustness to handle complex degradation scenarios. Extensive experiments demonstrate that Reti-Diff outperforms existing methods on three IDIR tasks, as well as downstream applications.
****
Yuqian Fu,Yuanheng Zhu,Jian Zhao,Jiajun Chai,Dongbin Zhao
INS: Interaction-aware Synthesis to Enhance Offline Multi-agent Reinforcement Learning
Data scarcity in offline multi-agent reinforcement learning (MARL) is a key challenge for real-world applications. Recent advances in offline single-agent reinforcement learning (RL) demonstrate the potential of data synthesis to mitigate this issue.
However, in multi-agent systems, interactions between agents introduce additional challenges. These interactions complicate the synthesis of multi-agent datasets, leading to data distortion when inter-agent interactions are neglected. Furthermore, the quality of the synthetic dataset is often constrained by the original dataset. To address these challenges, we propose **INteraction-aware Synthesis (INS)**, which synthesizes high-quality multi-agent datasets using diffusion models. Recognizing the sparsity of inter-agent interactions, INS employs a sparse attention mechanism to capture these interactions, ensuring that the synthetic dataset reflects the underlying agent dynamics. To overcome the limitation of diffusion models requiring continuous variables, INS implements a bit action module, enabling compatibility with both discrete and continuous action spaces. Additionally, we incorporate a select mechanism to prioritize transitions with higher estimated values, further enhancing the dataset quality. Experimental results across multiple datasets in MPE and SMAC environments demonstrate that INS consistently outperforms existing methods, resulting in improved downstream policy performance and superior dataset metrics. Notably, INS can synthesize high-quality data using only 10% of the original dataset, highlighting its efficiency in data-limited scenarios.
****
Aadirupa Saha,Pierre Gaillard
Finally Rank-Breaking Conquers MNL Bandits: Optimal and Efficient Algorithms for MNL Assortment
We address the problem of active online assortment optimization problem with preference feedback, which is a framework for modeling user choices and subsetwise utility maximization. The framework is useful in various real-world applications including ad placement, online retail, recommender systems, and fine-tuning language models, amongst many others. The problem, although has been studied in the past, lacks an intuitive and practical solution approach with simultaneously efficient algorithm and optimal regret guarantee. E.g., popularly used assortment selection algorithms often require the presence of a ``strong reference" which is always included in the choice sets, further they are also designed to offer the same assortments repeatedly until the reference item gets selected---all such requirements are quite unrealistic for practical applications. In this paper, we designed efficient algorithms for the problem of regret minimization in assortment selection with \emph{Plackett Luce} (PL) based user choices. We designed a novel concentration guarantee for estimating the score parameters of the PL model using `\emph{Pairwise Rank-Breaking}', which builds the foundation of our proposed algorithms. Moreover, our methods are practical, provably optimal, and devoid of the aforementioned limitations of the existing methods.
****
Wang Qinsi,Jinghan Ke,Masayoshi Tomizuka,Kurt Keutzer,Chenfeng Xu
Dobi-SVD: Differentiable SVD for LLM Compression and Some New Perspectives
Large language models (LLMs) have sparked a new wave of AI applications; however, their substantial computational costs and memory demands pose significant challenges to democratizing access to LLMs for a broader audience. Singular Value De

composition (SVD), a technique studied for decades, offers a hardware-independen
t and flexibly tunable solution for LLM compression. In this paper, we present n
ew directions using SVD: we first theoretically analyze the optimality of trunca
ting weights and truncating activations, then we further identify three key issu
es on SVD-based LLM compression, including (1) How can we determine the optimal
truncation position for each weight matrix in LLMs? (2) How can we efficiently u
pdate the weight matrices based on truncation position? (3) How can we address t
he inherent "injection" nature that results in the information loss of the SVD?
We propose an effective approach, **Dobi-SVD**, to tackle the three issues.
First, we propose a **differentiable** truncation-value learning mechanism, alon
g with gradient-robust backpropagation, enabling the model to adaptively find th
e optimal truncation positions. Next, we utilize the Eckart-Young-Mirsky theorem
 to derive a theoretically **optimal** weight update formula through rigorous ma
thematical analysis. Lastly, by observing and leveraging the quantization-friend
ly nature of matrices after SVD decomposition, we reconstruct a mapping between
truncation positions and memory requirements, establishing a **bijection** from
truncation positions to memory.
Experimental results show that with a 40\% parameter-compression rate, our metho
d achieves a perplexity of 9.07 on the Wikitext2 dataset with the compressed LLa
ma-7B model, a 78.7\% improvement over the state-of-the-art SVD for LLM compress
ion method.
We emphasize that Dobi-SVD is the first to achieve such a high-ratio LLM compres
sion with minimal performance drop. We also extend our Dobi-SVD to VLM compress
ion, achieving a 20\% increase in throughput with minimal performance degradatio
n. We hope that the inference speedup—up to 12.4x on 12GB NVIDIA Titan Xp GPUs a
nd 3x on 80GB A100 GPUs for LLMs, and 1.2x on 80GB A100 GPUs for VLMs—will bring
 significant benefits to the broader community such as robotics.
****
Qingyun Li,Zhe Chen,Weiyun Wang,Wenhai Wang,Shenglong Ye,Zhenjiang Jin,Guanzhou
Chen,Yinan He,Zhangwei Gao,Erfei Cui,Jiashuo Yu,Hao Tian,Jiasheng Zhou,Chao Xu,B
in Wang,Xingjian Wei,Wei Li,Wenjian Zhang,Bo Zhang,Pinlong Cai,Licheng Wen,Xiang
chao Yan,Pei Chu,Yi Wang,Min Dou,Changyao Tian,Xizhou Zhu,Lewei Lu,Yushi Chen,Ju
njun He,Tong Lu,Yali Wang,Limin Wang,Dahua Lin,Yu Qiao,Botian Shi,Conghui He,Jif
eng Dai
OmniCorpus: A Unified Multimodal Corpus of 10 Billion-Level Images Interleaved w
ith Text
Image-text interleaved data, consisting of multiple images and texts arranged in
 a natural document format, aligns with the presentation paradigm of internet da
ta and closely resembles human reading habits. Recent studies have shown that su
ch data aids multimodal in-context learning and maintains the capabilities of la
rge language models during multimodal fine-tuning. However, the limited scale an
d diversity of current image-text interleaved data restrict the development of m
ultimodal large language models. In this paper, we introduce OmniCorpus, a 10 bi
llion-scale image-text interleaved dataset. Using an efficient data engine, we f
ilter and extract large-scale high-quality documents, which contain 8.6 billion
images and 1,696 billion text tokens. Compared to counterparts (e.g., MMC4, OBEL
ICS), our dataset 1) has 15 times larger scales while maintaining good data qual
ity; 2) features more diverse sources, including both English and non-English we
bsites as well as video-centric websites; 3) is more flexible, easily degradable
 from an image-text interleaved format to pure text corpus and image-text pairs.
 Through comprehensive analysis and experiments, we validate the quality, usabil
ity, and effectiveness of the proposed dataset. We hope this could provide a sol
id data foundation for future multimodal model research.
****
Nicolas Atienza,Johanne Cohen,Christophe Labreuche,Michele Sebag
Provably Safeguarding a Classifier from OOD and Adversarial Samples
This paper aims to transform a trained classifier into an abstaining classifier,
 such
that the latter is provably protected from out-of-distribution and adversarial s
amples. The proposed Sample-efficient Probabilistic Detection using Extreme Valu

e
Theory (SPADE) approach relies on a Generalized Extreme Value (GEV) model
of the training distribution in the latent space of the classifier. Under mild a
ssumptions, this GEV model allows for formally characterizing out-of-distributio
n
and adversarial samples and rejecting them. Empirical validation of the approach
is conducted on various neural architectures (ResNet, VGG, and Vision Transforme
r) and considers medium and large-sized datasets (CIFAR-10, CIFAR-100,
and ImageNet). The results show the stability and frugality of the GEV model and
demonstrate SPADE’s efficiency compared to the state-of-the-art methods.
****
Murdock Aubry,Haoming Meng,Anton Sugolov,Vardan Papyan
Transformer Block Coupling and its Correlation with Generalization in LLMs
Large Language Models (LLMs) have made significant strides in natural language p
rocessing, and a precise understanding of the internal mechanisms driving their
success is essential. In this work, we analyze the trajectories of token embeddi
ngs as they pass through transformer blocks, linearizing the system along these
trajectories through their Jacobian matrices. By examining the relationships bet
ween these block Jacobians, we uncover the phenomenon of **transformer block cou
pling** in a multitude of LLMs, characterized by the coupling of their top singu
lar vectors across tokens and depth. Our findings reveal that coupling *positive
ly correlates* with model performance, and that this relationship is stronger th
an with other hyperparameters such as parameter count, model depth, and embeddin
g dimension. We further investigate how these properties emerge during training,
 observing a progressive development of coupling, increased linearity, and layer
-wise exponential growth in token trajectories. Additionally, experiments with V
ision Transformers (ViTs) corroborate the emergence of coupling and its relation
ship with generalization, reinforcing our findings in LLMs. Collectively, these
insights offer a novel perspective on token interactions in transformers, openin
g new directions for studying their mechanisms as well as improving training and
 generalization.
****
Jinxu Lin,Linwei Tao,Minjing Dong,Chang Xu
Diffusion Attribution Score: Evaluating Training Data Influence in Diffusion Mod
els
As diffusion models become increasingly popular, the misuse of copyrighted and p
rivate images has emerged as a major concern. One promising solution to mitigate
 this issue is identifying the contribution of specific training samples in gene
rative models, a process known as data attribution. Existing data attribution me
thods for diffusion models typically quantify the contribution of a training sam
ple by evaluating the change in diffusion loss when the sample is included or ex
cluded from the training process.
However, we argue that the direct usage of diffusion loss cannot represent such
a contribution accurately due to the calculation of diffusion loss.
Specifically, these approaches measure the divergence between predicted and grou
nd truth distributions, which leads to an indirect comparison between the predic
ted distributions and cannot represent the variances between model behaviors.
To address these issues, we aim to measure the direct comparison between predict
ed distributions with an attribution score to analyse the training sample import
ance, which is achieved by Diffusion Attribution Score (\textit{DAS}).
Underpinned by rigorous theoretical analysis, we elucidate the effectiveness of
DAS.
Additionally, we explore strategies to accelerate DAS calculations, facilitating
 its application to large-scale diffusion models.
Our extensive experiments across various datasets and diffusion models demonstra
te that DAS significantly surpasses previous benchmarks in terms of the linear d
ata-modelling score, establishing new state-of-the-art performance.
****
Xiangru Tang,Tianyu Hu,Muyang Ye,Yanjun Shao,Xunjian Yin,Siru Ouyang,Wangchunshu
 Zhou,Pan Lu,Zhuosheng Zhang,Yilun Zhao,Arman Cohan,Mark Gerstein

ChemAgent: Self-updating Memories in Large Language Models Improves Chemical Rea
soning
Chemical reasoning usually involves complex, multi-step processes that demand pr
ecise calculations, where even minor errors can lead to cascading failures. Furt
hermore, large language models (LLMs) encounter difficulties handling domain-spe
cific formulas, executing reasoning steps accurately, and integrating code ef- e
ffectively when tackling chemical reasoning tasks. To address these challenges,
we present ChemAgent, a novel framework designed to improve the performance of L
LMs through a dynamic, self-updating library. This library is developed by decom
posing chemical tasks into sub-tasks and compiling these sub-tasks into a struct
ured collection that can be referenced for future queries. Then, when presented
with a new problem, ChemAgent retrieves and refines pertinent information from t
he library, which we call memory, facilitating effective task decomposition and
the generation of solutions. Our method designs three types of memory and a libr
ary-enhanced reasoning component, enabling LLMs to improve over time through exp
erience. Experimental results on four chemical reasoning datasets from SciBench
demonstrate that ChemAgent achieves performance gains of up to 46% (GPT-4), sign
ificantly outperforming existing methods. Our findings suggest substantial poten
tial for future applications, including tasks such as drug discovery and materia
ls science. Our code can be found at https://github.com/gersteinlab/ChemAgent.
****
Zhiyuan Wu,Changkyu Choi,Xiangcheng Cao,Volkan Cevher,Ali Ramezani-Kebrya
Addressing Label Shift in Distributed Learning via Entropy Regularization■
We address the challenge of minimizing "true risk" in multi-node distributed lea
rning.\footnote{We use the term node to refer to a client, FPGA, APU, CPU, GPU,
or worker.} These systems are frequently exposed to both inter-node and intra-no
de "label shifts", which present a critical obstacle to effectively optimizing m
odel performance while ensuring that data remains confined to each node.
To tackle this, we propose the Versatile Robust Label Shift (VRLS) method, which
 enhances the maximum likelihood estimation of the test-to-train label importanc
e ratio. VRLS incorporates Shannon entropy-based regularization and adjusts the
importance ratio during training to better handle label shifts at the test time
.
In multi-node learning environments, VRLS further extends its capabilities by le
arning and adapting importance ratios across nodes, effectively mitigating label
 shifts and improving overall model performance. Experiments conducted on MNIST,
 Fashion MNIST, and CIFAR-10 demonstrate the effectiveness of VRLS, outperformin
g baselines by up to 20\% in imbalanced settings. These results highlight the si
gnificant improvements VRLS offers in addressing label shifts. Our theoretical a
nalysis further supports this by establishing high-probability bounds on estimat
ion errors.
****
Zhanghao Zhouyin,Zixi Gan,Shishir Kumar Pandey,Linfeng Zhang,Qiangqiang Gu
Learning local equivariant representations for quantum operators
Predicting quantum operator matrices such as Hamiltonian, overlap, and density m
atrices in the density functional theory (DFT) framework is crucial for material
 science. Current methods often focus on individual operators and struggle with
efficiency and scalability for large systems. Here we introduce a novel deep lea
rning model, SLEM (strictly localized equivariant message-passing), for predicti
ng multiple quantum operators that achieves state-of-the-art accuracy while dram
atically improving computational efficiency. SLEM's key innovation is its strict
 locality-based design for equivariant representations of quantum tensors while
preserving physical symmetries. This enables complex many-body dependency withou
t expanding the effective receptive field, leading to superior data efficiency a
nd transferability. Using an innovative SO(2) convolution and invariant overlap
parameterization, SLEM reduces the computational complexity of high-order tensor
 products and is, therefore, capable of handling systems requiring the f and $
g$ orbitals in their basis sets. We demonstrate SLEM's capabilities across diver
se 2D and 3D materials, achieving high accuracy even with limited training data.
 SLEM's design facilitates efficient parallelization, potentially extending DFT

simulations to systems with device-level sizes, opening new possibilities for la
rge-scale quantum simulations and high-throughput materials discovery.
****
Xihan Li,Xing Li,Lei Chen,Xing Zhang,Mingxuan Yuan,Jun Wang
Circuit Transformer: A Transformer That Preserves Logical Equivalence
Implementing Boolean functions with circuits consisting of logic gates is fundam
ental in digital computer design. However, the implemented circuit must be exact
ly equivalent, which hinders generative neural approaches on this task due to th
eir occasionally wrong predictions. In this study, we introduce a generative neu
ral model, the “Circuit Transformer”, which eliminates such wrong predictions an
d produces logic circuits strictly equivalent to given Boolean functions. The ma
in idea is a carefully designed decoding mechanism that builds a circuit step-by
-step by generating tokens, which has beneficial “cutoff properties” that block
a candidate token once it invalidate equivalence. In such a way, the proposed mo
del works similar to typical LLMs while logical equivalence is strictly preserve
d. A Markov decision process formulation is also proposed for optimizing certain
 objectives of circuits. Experimentally, we trained an 88-million-parameter Circ
uit Transformer to generate equivalent yet more compact forms of input circuits,
 outperforming existing neural approaches on both synthetic and real world bench
marks, without any violation of equivalence constraints.

Code: https://github.com/snowkylin/circuit-transformer
****
Enyu Zhou,Guodong Zheng,Binghai Wang,Zhiheng Xi,Shihan Dou,Rong Bao,Wei Shen,Lim
ao Xiong,Jessica Fan,Yurong Mou,Rui Zheng,Tao Gui,Qi Zhang,Xuanjing Huang
RMB: Comprehensively benchmarking reward models in LLM alignment
Reward models (RMs) guide the alignment of large language models (LLMs), steerin
g them toward behaviors preferred by humans. Evaluating RMs is the key to better
 aligning LLMs. However, the current evaluation of RMs may not directly correspo
nd to their alignment performance due to the limited distribution of evaluation
data and evaluation methods that are not closely related to alignment objectives
. To address these limitations, we propose RMB, a comprehensive RM benchmark tha
t covers over 49 real-world scenarios and includes both pairwise and Best-of-N (
BoN) evaluations to better reflect the effectiveness of RMs in guiding alignment
 optimization.
We demonstrate a positive correlation between our benchmark and the downstream a
lignment task performance. Based on our benchmark, we conduct extensive analysis
 on the state-of-the-art RMs, revealing their generalization defects that were n
ot discovered by previous benchmarks, and highlighting the potential of generati
ve RMs. Furthermore, we delve into open questions in reward models, specificall
y examining the effectiveness of majority voting for the evaluation of reward mo
dels and analyzing the impact factors of generative RMs, including the influence
 of evaluation criteria and instructing methods. We will release our evaluation
code and datasets upon publication.
****
Sehyeok Kang,Yongsik Lee,Gahee Kim,Song Chong,Se-Young Yun
MA^{2}E: Addressing Partial Observability in Multi-Agent Reinforcement Learning
with Masked Auto-Encoder
Centralized Training and Decentralized Execution (CTDE) is a widely adopted para
digm to solve cooperative multi-agent reinforcement learning (MARL) problems. De
spite the successes achieved with CTDE, partial observability still limits coope
ration among agents. While previous studies have attempted to overcome this chal
lenge through communication, direct information exchanges could be restricted an
d introduce additional constraints. Alternatively, if an agent can infer the glo
bal information solely from local observations, it can obtain a global view with
out the need for communication. To this end, we propose the Multi-Agent Masked A
uto-Encoder (MA^{2}E), which utilizes the masked auto-encoder architecture to in
fer the information of other agents from partial observations. By employing mask
ing to learn to reconstruct global information, MA^{2}E serves as an inference m
odule for individual agents within the CTDE framework. MA^{2}E can be easily int

egrated into existing MARL algorithms and has been experimentally proven to be e
ffective across a wide range of environments and algorithms.
****

Eli Chien,Pan Li

Convergent Privacy Loss of Noisy-SGD without Convexity and Smoothness

We study the Differential Privacy (DP) guarantee of hidden-state Noisy-SGD algor
ithms over a bounded domain. Standard privacy analysis for Noisy-SGD assumes all
 internal states are revealed, which leads to a divergent R\'enyi DP bound with
respect to the number of iterations. Ye & Shokri (2022) and Altschuler & Talwar
(2022) proved convergent bounds for smooth (strongly) convex losses, and raise o
pen questions about whether these assumptions can be relaxed. We provide positiv
e answers by proving convergent R\'enyi DP bound for non-convex non-smooth losse
s, where we show that requiring losses to have H\"older continuous gradient is s
ufficient. We also provide a strictly better privacy bound compared to state-of-
the-art results for smooth strongly convex losses. Our analysis relies on the im
provement of shifted divergence analysis in multiple aspects, including forward
Wasserstein distance tracking, identifying the optimal shifts allocation, and th
e H\"older reduction lemma. Our results further elucidate the benefit of hidden
-state analysis for DP and its applicability.
****

Steve Azzolin,Antonio Longa,Stefano Teso,Andrea Passerini

Reconsidering Faithfulness in Regular, Self-Explainable and Domain Invariant GNN
s

As Graph Neural Networks (GNNs) become more pervasive, it becomes paramount to b
uild reliable tools for explaining their predictions.
A core desideratum is that explanations are *faithful*, i.e., that they portray
an accurate picture of the GNN's reasoning process.
However, a number of different faithfulness metrics exist, begging the question
of what is faithfulness exactly and how to achieve it.
We make three key contributions.
We begin by showing that *existing metrics are not interchangeable* -- i.e., exp
lanations attaining high faithfulness according to one metric may be unfaithful
according to others -- and can *systematically ignore important properties of ex
planations*.
We proceed to show that, surprisingly, *optimizing for faithfulness is not alway
s a sensible design goal*. Specifically, we prove that for injective regular GN
N architectures, perfectly faithful explanations are completely uninformative.
This does not apply to modular GNNs, such as self-explainable and domain-invaria
nt architectures, prompting us to study the relationship between architectural c
hoices and faithfulness.
Finally, we show that *faithfulness is tightly linked to out-of-distribution gen
eralization*, in that simply ensuring that a GNN can correctly recognize the dom
ain-invariant subgraph, as prescribed by the literature, does not guarantee that
 it is invariant unless this subgraph is also faithful.
All our code can be found in the supplementary material.
****

Hongyu Zhao,Ming Li,Lichao Sun,Tianyi Zhou

BenTo: Benchmark Reduction with In-Context Transferability

Evaluating large language models (LLMs) is costly: it requires the generation an
d examination of LLM outputs on a large-scale benchmark of various tasks. This p
aper investigates how to efficiently reduce the tasks used to benchmark LLMs wit
hout affecting the evaluation quality. Our study reveals that task transferabili
ty and relevance provide critical information to identify the most representativ
e subset of tasks via optimizing a facility location function. We propose a prac
tically efficient metric for estimating the transferability between two tasks vi
a in-context learning (ICL). By analyzing the pairwise transferability, we can r
educe tasks in a modern LLM benchmark (e.g., MMLU or FLAN) to 5\% while inducing
 only a <4\% difference to the evaluation on the original benchmark. Compared
to prior works, our method is training-free, gradient-free, and highly efficient
 requiring ICL only.

****
Shizhan Gong,Haoyu LEI,Qi Dou,Farzan Farnia
Boosting the visual interpretability of CLIP via adversarial fine-tuning
CLIP has achieved great success in visual representation learning and is becomin
g an important plug-in component for many large multi-modal models like LLaVA an
d DALL-E. However, the lack of interpretability caused by the intricate image en
coder architecture and training process restricts its wider use in high-stake de
cision making applications. In this work, we propose an unsupervised adversarial
 fine-tuning (AFT) with norm-regularization to enhance the visual interpretabili
ty of CLIP. We provide theoretical analysis showing that AFT has implicit regula
rization that enforces the image encoder to encode the input features sparsely,
directing the network's focus towards meaningful features. Evaluations by both f
eature attribution techniques and network dissection offer convincing evidence t
hat the visual interpretability of CLIP has significant improvements. With AFT,
the image encoder prioritizes pertinent input features, and the neuron within th
e encoder exhibits better alignment with human-understandable concepts. Moreover
, these effects are generalizable to out-of-distribution datasets and can be tra
nsferred to downstream tasks. Additionally, AFT enhances the visual interpretabi
lity of derived large vision-language models that incorporate the pre-trained CL
IP an integral component. The code of this paper is available at [the CLIP_AFT G
itHub repository](https://github.com/peterant330/CLIP_AFT).
****
Jiaxin Wen,Vivek Hebbar,Caleb Larson,Aryan Bhatt,Ansh Radhakrishnan,Mrinank Shar
ma,Henry Sleight,Shi Feng,He He,Ethan Perez,Buck Shlegeris,Akbir Khan
Adaptive Deployment of Untrusted LLMs Reduces Distributed Threats
As large language models (LLMs) grow more powerful, they also become more diffic
ult to trust. They could be either aligned with human intentions, or exhibit "su
bversive misalignment" -- introducing subtle errors that bypass safety checks. A
lthough individual errors may not immediately cause harm, each increases the ris
k of an eventual safety failure. With this uncertainty, model deployment often g
rapples with the tradeoff between ensuring safety and harnessing the capabilitie
s of untrusted models. In this work, we introduce the ``Diffuse Risk Management'
' problem, aiming to balance the average-case safety and usefulness in the deplo
yment of untrusted models over a large sequence of tasks. We approach this probl
em by developing a two-level framework: the single-task level (micro-protocol) a
nd the whole-scenario level (macro-protocol). At the single-task level, we devel
op various \textit{micro}-protocols that use a less capable, but extensively tes
ted (trusted) model to harness and monitor the untrusted model. At the whole-sce
nario level, we find an optimal \textit{macro}-protocol that uses an adaptive es
timate of the untrusted model's risk to choose between micro-protocols. To evalu
ate the robustness of our method, we follow \textit{control evaluations} in a co
de generation testbed, which involves a red team attempting to generate subtly b
ackdoored code with an LLM whose deployment is safeguarded by a blue team. Exper
iment results show that our approach retains 99.6\% usefulness of the untrusted
model while ensuring near-perfect safety, significantly outperforming existing d
eployment methods. Our approach also demonstrates robustness when the trusted an
d untrusted models have a large capability gap. Our findings demonstrate the pro
mise of managing diffuse risks in the deployment of increasingly capable but unt
rusted LLMs.
****
Krzysztof Kacprzyk,Mihaela van der Schaar
No Equations Needed: Learning System Dynamics Without Relying on Closed-Form ODE
s
Data-driven modeling of dynamical systems is a crucial area of machine learning.
 In many scenarios, a thorough understanding of the model’s behavior becomes ess
ential for practical applications. For instance, understanding the behavior of a
 pharmacokinetic model, constructed as part of drug development, may allow us to
 both verify its biological plausibility (e.g., the drug concentration curve is
non-negative and decays to zero in the long term) and to design dosing guideline
s (e.g., by looking at the peak concentration and its timing). Discovery of clos

ed-form ordinary differential equations (ODEs) can be employed to obtain such in sights by finding a compact mathematical equation and then analyzing it (a two-s tep approach). However, its widespread use is currently hindered because the ana lysis process may be time-consuming, requiring substantial mathematical expertis e, or even impossible if the equation is too complex. Moreover, if the found equ ation's behavior does not satisfy the requirements, editing it or influencing th e discovery algorithms to rectify it is challenging as the link between the symb olic form of an ODE and its behavior can be elusive. This paper proposes a conce ptual shift to modeling low-dimensional dynamical systems by departing from the traditional two-step modeling process. Instead of first discovering a closed-for m equation and then analyzing it, our approach, direct semantic modeling, predic ts the semantic representation of the dynamical system (i.e., description of its behavior) directly from data, bypassing the need for complex post-hoc analysis. This direct approach also allows the incorporation of intuitive inductive biase s into the optimization algorithm and editing the model's behavior directly, ens uring that the model meets the desired specifications. Our approach not only sim plifies the modeling pipeline but also enhances the transparency and flexibility of the resulting models compared to traditional closed-form ODEs.
****

Mohammad Bashiri,Luca Baroni,Ján Antolík,Fabian H. Sinz
Learning and aligning single-neuron invariance manifolds in visual cortex
Understanding how sensory neurons exhibit selectivity to certain features and in variance to others is central to uncovering the computational principles underly ing robustness and generalization in visual perception. Most existing methods fo r characterizing selectivity and invariance identify single or finite discrete s ets of stimuli. Since these are only isolated measurements from an underlying co ntinuous manifold, characterizing invariance properties accurately and comparing them across neurons with varying receptive field size, position, and orientatio n, becomes challenging. Consequently, a systematic analysis of invariance types at the population level remains under-explored. Building on recent advances in l earning continuous invariance manifolds, we introduce a novel method to accurate ly identify and align invariance manifolds of visual sensory neurons, overcoming these challenges. Our approach first learns the continuous invariance manifold of stimuli that maximally excite a neuron modeled by a response-predicting deep neural network. It then learns an affine transformation on the pixel coordinates such that the same manifold activates another neuron as strongly as possible, e ffectively aligning their invariance manifolds spatially. This alignment provide s a principled way to quantify and compare neuronal invariances irrespective of receptive field differences. Using simulated neurons, we demonstrate that our me thod accurately learns and aligns known invariance manifolds, robustly identifyi ng functional clusters. When applied to macaque V1 neurons, it reveals functiona l clusters of neurons, including simple and complex cells. Overall, our method e nables systematic, quantitative exploration of the neural invariance landscape, to gain new insights into the functional properties of visual sensory neurons.
****

Arkaprava Majumdar,M Anand Krishna,P. K. Srijith
Neural Wave Equation for Irregularly Sampled Sequence Data
Sequence labeling problems arise in several real-world applications such as heal thcare and robotics. In many such applications, sequence data are irregularly sa mpled and are of varying complexities. Recently, efforts have been made to devel op neural ODE-based architectures to model the evolution of hidden states contin uously in time, to address irregularly sampled sequence data. However, they assu me a fixed architectural depth and limit their flexibility to adapt to data sets with varying complexities. We propose the neural wave equation, a novel deep le arning method inspired by the wave equation, to address this through continuous modeling of depth. Neural Wave Equation models the evolution of hidden states co ntinuously across time as well as depth by using a non-homogeneous wave equation parameterized by a neural network. Through d'Alembert's analytical solution of the wave equation, we also show that the neural wave equation provides denser c onnections across the hidden states, allowing for better modeling capability. W

e conduct experiments on several sequence labeling problems involving irregularl
y sampled sequence data and demonstrate the superior performance of the proposed
 neural wave equation model.
****

Siqi Liu,Ian Gemp,Luke Marris,Georgios Piliouras,Nicolas Heess,Marc Lanctot
Re-evaluating Open-ended Evaluation of Large Language Models
Evaluation has traditionally focused on ranking candidates for a specific skill.
 Modern generalist models, such as Large Language Models (LLMs), decidedly outpa
ce this paradigm. Open-ended evaluation systems, where candidate models are comp
ared on user-submitted prompts, have emerged as a popular solution. Despite thei
r many advantages, we show that the current Elo-based rating systems can be susc
eptible to and even reinforce biases in data, intentional or accidental, due to
their sensitivity to redundancies. To address this issue, we propose evaluation
as a 3-player game, and introduce novel game-theoretic solution concepts to ensu
re robustness to redundancy. We show that our method leads to intuitive ratings
and provide insights into the competitive landscape of LLM development.
****

Mufei Li,Viraj Shitole,Eli Chien,Changhai Man,Zhaodong Wang,Srinivas,Ying Zhang,
Tushar Krishna,Pan Li
LayerDAG: A Layerwise Autoregressive Diffusion Model for Directed Acyclic Graph
Generation
Directed acyclic graphs (DAGs) serve as crucial data representations in domains
such as hardware synthesis and compiler/program optimization for computing syste
ms. DAG generative models facilitate the creation of synthetic DAGs, which can b
e used for benchmarking computing systems while preserving intellectual property
. However, generating realistic DAGs is challenging due to their inherent direct
ional and logical dependencies. This paper introduces LayerDAG, an autoregressiv
e diffusion model, to address these challenges. LayerDAG decouples the strong no
de dependencies into manageable units that can be processed sequentially. By int
erpreting the partial order of nodes as a sequence of bipartite graphs, LayerDAG
 leverages autoregressive generation to model directional dependencies and emplo
ys diffusion models to capture logical dependencies within each bipartite graph.
 Comparative analyses demonstrate that LayerDAG outperforms existing DAG generat
ive models in both expressiveness and generalization, particularly for generatin
g large-scale DAGs with up to 400 nodes—a critical scenario for system benchmark
ing. Extensive experiments on both synthetic and real-world flow graphs from var
ious computing platforms show that LayerDAG generates valid DAGs with superior s
tatistical properties and benchmarking performance. The synthetic DAGs generated
 by LayerDAG enhance the training of ML-based surrogate models, resulting in imp
roved accuracy in predicting performance metrics of real-world DAGs across diver
se computing platforms.
****

Kaan Akan,Yucel Yemez
Slot-Guided Adaptation of Pre-trained Diffusion Models for Object-Centric Learni
ng and Compositional Generation
We present SlotAdapt, an object-centric learning method that combines slot atten
tion with pretrained diffusion models by introducing adapters for slot-based con
ditioning. Our method preserves the generative power of pretrained diffusion mod
els, while avoiding their text-centric conditioning bias. We also incorporate an
 additional guidance loss into our architecture to align cross-attention from ad
apter layers with slot attention. This enhances the alignment of our model with
the objects in the input image without using external supervision. Experimental
results show that our method outperforms state-of-the-art techniques in object d
iscovery and image generation tasks across multiple datasets, including those wi
th real images. Furthermore, we demonstrate through experiments that our method
performs remarkably well on complex real-world images for compositional generati
on, in contrast to other slot-based generative methods in the literature. The pr
oject page can be found at https://kaanakan.github.io/SlotAdapt/
****

Arvind Pillai,Dimitris Spathis,Fahim Kawsar,Mohammad Malekzadeh

PaPaGei: Open Foundation Models for Optical Physiological Signals
Photoplethysmography (PPG) is the leading non-invasive technique for monitoring biosignals and cardiovascular health, with widespread adoption in both clinical settings and consumer wearable devices. While machine learning models trained on PPG signals have shown promise, they tend to be task-specific and struggle with generalization. Current research is limited by the use of single-device datasets, insufficient exploration of out-of-domain generalization, and a lack of publicly available models, which hampers reproducibility. To address these limitations, we present PaPaGei, the first open foundation model for PPG signals. The model is pre-trained on over 57,000 hours of data, comprising 20 million unlabeled PPG segments from publicly available datasets. We introduce a novel representation learning approach that leverages domain knowledge of PPG signal morphology across individuals, enabling the capture of richer representations compared to traditional contrastive learning methods. We evaluate PaPaGei against state-of-the-art time-series foundation models and self-supervised learning benchmarks across 20 tasks from 10 diverse datasets, spanning cardiovascular health, sleep disorders, pregnancy monitoring, and wellbeing assessment. Our model demonstrates superior performance, improving classification and regression metrics by 6.3% and 2.9% respectively in at least 14 tasks. Notably, PaPaGei achieves these results while being more data- and parameter-efficient, outperforming models that are 70x larger. Beyond accuracy, we examine model robustness across different skin tones, establishing a benchmark for bias evaluation in future models. PaPaGei can serve as both a feature extractor and an encoder for multimodal models, opening up new opportunities for multimodal health monitoring. Models, data, and code are available at: https://github.com/nokia-bell-labs/papagei-foundation-model
****
Zhitong Xu,Haitao Wang,Jeff M. Phillips,Shandian Zhe
Standard Gaussian Process is All You Need for High-Dimensional Bayesian Optimization
A long-standing belief holds that Bayesian Optimization (BO) with standard Gaussian processes (GP) --- referred to as standard BO --- underperforms in high-dimensional optimization problems. While this belief seems plausible, it lacks both robust empirical evidence and theoretical justification. To address this gap, we present a systematic investigation. First, through a comprehensive evaluation across twelve benchmarks, we found that while the popular Square Exponential (SE) kernel often leads to poor performance, using Mat\'ern kernels enables standard BO to consistently achieve top-tier results, frequently surpassing methods specifically designed for high-dimensional optimization. Second, our theoretical analysis reveals that the SE kernel’s failure primarily stems from improper initialization of the length-scale parameters, which are commonly used in practice but can cause gradient vanishing in training. We provide a probabilistic bound to characterize this issue, showing that Mat\'ern kernels are less susceptible and can robustly handle much higher dimensions. Third, we propose a simple robust initialization strategy that dramatically improves the performance of the SE kernel, bringing it close to state-of-the-art methods, without requiring additional priors or regularization. We prove another probabilistic bound that demonstrates how the gradient vanishing issue can be effectively mitigated with our method. Our findings advocate for a re-evaluation of standard BO’s potential in high-dimensional settings.
****
Runzhe Wu,Yiding Chen,Gokul Swamy,Kianté Brantley,Wen Sun
Diffusing States and Matching Scores: A New Framework for Imitation Learning
Adversarial Imitation Learning is traditionally framed as a two-player zero-sum game between a learner and an adversarially chosen cost function, and can therefore be thought of as the sequential generalization of a Generative Adversarial Network (GAN). However, in recent years, diffusion models have emerged as a non-adversarial alternative to GANs that merely require training a score function via regression, yet produce generations of higher quality. In response, we investigate how to lift insights from diffusion modeling to the sequential setting. We propose diffusing states and performing *score-matching* along diffused states to

measure the discrepancy between the expert's and learner's states. Thus, our approach only requires training score functions to predict noises via standard regression, making it significantly easier and more stable to train than adversarial methods. Theoretically, we prove first- and second-order instance-dependent bounds with linear scaling in the horizon, proving that our approach avoids the compounding errors that stymie offline approaches to imitation learning. Empirically, we show our approach outperforms both GAN-style imitation learning baselines and discriminator-free imitation learning baselines across various continuous control problems, including complex tasks like controlling humanoids to walk, sit, crawl, and navigate through obstacles.
****

Nikola Jovanovi■,Robin Staab,Maximilian Baader,Martin Vechev
Ward: Provable RAG Dataset Inference via LLM Watermarks
RAG enables LLMs to easily incorporate external data, raising concerns for data owners regarding unauthorized usage of their content. The challenge of detecting such unauthorized usage remains underexplored, with datasets and methods from adjacent fields being ill-suited for its study. We take several steps to bridge this gap. First, we formalize this problem as (black-box) RAG Dataset Inference (RAG-DI). We then introduce a novel dataset designed for realistic benchmarking of RAG-DI methods, alongside a set of baselines. Finally, we propose Ward, a method for RAG-DI based on LLM watermarks that equips data owners with rigorous statistical guarantees regarding their dataset's misuse in RAG corpora. Ward consistently outperforms all baselines, achieving higher accuracy, superior query efficiency and robustness. Our work provides a foundation for future studies of RAG-DI and highlights LLM watermarks as a promising approach to this problem.
****

Xingtong Yu,Zhenghao Liu,Xinming Zhang,Yuan Fang
Node-Time Conditional Prompt Learning in Dynamic Graphs
Dynamic graphs capture evolving interactions between entities, such as in social networks, online learning platforms, and crowdsourcing projects. For dynamic graph modeling, dynamic graph neural networks (DGNNs) have emerged as a mainstream technique. However, they are generally pre-trained on the link prediction task, leaving a significant gap from the objectives of downstream tasks such as node classification. To bridge the gap, prompt-based learning has gained traction on graphs, but most existing efforts focus on static graphs, neglecting the evolution of dynamic graphs. In this paper, we propose DyGPrompt, a novel pre-training and prompt learning framework for dynamic graph modeling. First, we design dual prompts to address the gap in both task objectives and temporal variations across pre-training and downstream tasks. Second, we recognize that node and time patterns often characterize each other, and propose dual condition-nets to model the evolving node-time patterns in downstream tasks. Finally, we thoroughly evaluate and analyze DyGPrompt through extensive experiments on four public datasets.
****

Shen Li,Liuyi Yao,Lan Zhang,Yaliang Li
Safety Layers in Aligned Large Language Models: The Key to LLM Security
Aligned LLMs are secure, capable of recognizing and refusing to answer malicious questions. However, the role of internal parameters in maintaining such security is not well understood yet, further these models can be vulnerable to security degradation when subjected to fine-tuning attacks. To address these challenges, our work uncovers the mechanism behind security in aligned LLMs at the parameter level, identifying a small set of contiguous layers in the middle of the model that are crucial for distinguishing malicious queries from normal ones, referred to as "safety layers". We first confirm the existence of these safety layers by analyzing variations in input vectors within the model's internal layers. Additionally, we leverage the over-rejection phenomenon and parameters scaling analysis to precisely locate the safety layers. Building on these findings, we propose a novel fine-tuning approach, Safely Partial-Parameter Fine-Tuning (SPPFT), that fixes the gradient of the safety layers during fine-tuning to address the security degradation. Our experiments demonstrate that the proposed approach can significantly preserve LLM security while maintaining performance and reducing com

putational resources compared to full fine-tuning.
****
Xilong Wang,Hao Fu,Jindong Wang,Neil Zhenqiang Gong
StringLLM: Understanding the String Processing Capability of Large Language Mode ls
String processing, which mainly involves the analysis and manipulation of string s, is a fundamental component of modern computing. Despite the significant advan cements of large language models (LLMs) in various natural language processing (NLP) tasks, their capability in string processing remains underexplored and unde rdeveloped. To bridge this gap, we present a comprehensive study of LLMs' string processing capability. In particular, we first propose StringLLM, a method to c onstruct datasets for benchmarking string processing capability of LLMs. We use StringLLM to build a series of datasets, referred to as StringBench. It encompas ses a wide range of string processing tasks, allowing us to systematically evalu ate LLMs' performance in this area. Our evaluations indicate that LLMs struggle with accurately processing strings compared to humans. To uncover the underlying reasons for this limitation, we conduct an in-depth analysis and subsequently p ropose an effective approach that significantly enhances LLMs' string processing capability via fine-tuning. This work provides a foundation for future research to understand LLMs' string processing capability. Our code and data are availab le at https://github.com/wxl-lxw/StringLLM.
****
Qizhou Wang,Guansong Pang,Mahsa Salehi,Xiaokun Xia,Christopher Leckie
Open-Set Graph Anomaly Detection via Normal Structure Regularisation
This paper considers an important Graph Anomaly Detection (GAD) task, namely ope n-set GAD, which aims to train a detection model using a small number of normal and anomaly nodes (referred to as *seen anomalies*) to detect both seen anomalie s and *unseen anomalies* (*i.e*., anomalies that cannot be illustrated the train ing anomalies). Those labelled training data provide crucial prior knowledge abo ut abnormalities for GAD models, enabling substantially reduced detection errors . However, current supervised GAD methods tend to over-emphasise fitting the see n anomalies, leading to many errors of detecting the unseen anomalies as normal nodes. Further, existing open-set AD models were introduced to handle Euclidean data, failing to effectively capture discriminative features from graph structur e and node attributes for GAD. In this work, we propose a novel open-set GAD app roach, namely $\underline{n}ormal$ $\underline{s}tructure$ $\underline{reg}ulari sation$ (**NSReg**), to achieve generalised detection ability to unseen anomalie s, while maintaining its effectiveness on detecting seen anomalies. The key idea in NSReg is to introduce a regularisation term that enforces the learning of co mpact, semantically-rich representations of normal nodes based on their structur al relations to other nodes. When being optimised with supervised anomaly detect ion losses, the regularisation term helps incorporate strong normality into the modelling, and thus, it effectively avoids over-fitting the seen anomalies and l earns a better normality decision boundary, largely reducing the false negatives of detecting unseen anomalies as normal. Extensive empirical results on seven r eal-world datasets show that NSReg significantly outperforms state-of-the-art co mpeting methods by at least 14% AUC-ROC on the unseen anomaly classes and by 10% AUC-ROC on all anomaly classes. Code and datasets are available at https://gith ub.com/mala-lab/NSReg.
****
Ruchika Chavhan,Da Li,Timothy Hospedales
ConceptPrune: Concept Editing in Diffusion Models via Skilled Neuron Pruning
While large-scale text-to-image diffusion models have demonstrated impressive im age-generation capabilities, there are significant concerns about their potentia l misuse for generating unsafe content, violating copyright, and perpetuating so cietal biases. Recently, the text-to-image generation community has begun addres sing these concerns by editing or unlearning undesired concepts from pre-trained models. However, these methods often involve data-intensive and inefficient fin e-tuning or utilize various forms of token remapping, rendering them susceptible to adversarial jailbreaks. In this paper, we present a simple and effective tra

ining-free approach, ConceptPrune, wherein we first identify critical regions within pre-trained models responsible for generating undesirable concepts, thereby facilitating straightforward concept unlearning via weight pruning. Experiments across a range of concepts including artistic styles, nudity, and object erasure demonstrate that target concepts can be efficiently erased by pruning a tiny fraction, approximately 0.12% of total weights, enabling multi-concept erasure and robustness against various white-box and black-box adversarial attacks.
****
Dario Liscai,Emanuele Luconi,Alessandro Marin Vargas,Alessandro Sanzeni
Beyond single neurons: population response geometry in digital twins of mouse visual cortex
Hierarchical visual processing is essential for cognitive functions like object recognition and spatial localization. Traditional studies of the neural basis of these computations have focused on single-neuron activity, but recent advances in large-scale neural recordings emphasize the growing need to understand computations at the population level. Digital twins-computational models trained on neural data-have successfully replicated single-neuron behavior, but their effectiveness in capturing the joint activity of neurons remains unclear. In this study, we investigate how well digital twins describe population responses in mouse visual cortex. We show that these models fail to accurately represent the geometry of population activity, particularly its differentiability and how this geometry evolves across the visual hierarchy. To address this, we explore how dataset, network architecture, loss function, and training method affect the ability of digital twins to recapitulate population properties. We demonstrate that improving model alignment with experiments requires training strategies that enhance robustness and generalization, reflecting principles observed in biological systems. These findings underscore the need to evaluate digital twins from multiple perspectives, identify key areas for refinement, and establish a foundation for using these models to explore neural computations at the population level.
****
Javier Abad,Konstantin Donhauser,Francesco Pinto,Fanny Yang
Copyright-Protected Language Generation via Adaptive Model Fusion
The risk of language models reproducing copyrighted material from their training data has led to the development of various protective measures. Among these, inference-time strategies that impose constraints via post-processing have shown promise in addressing the complexities of copyright regulation. However, they often incur prohibitive computational costs or suffer from performance trade-offs. To overcome these limitations, we introduce Copyright-Protecting Model Fusion (CP-Fuse), a novel approach that combines models trained on disjoint sets of copyrighted material during inference. In particular, CP-Fuse adaptively aggregates the model outputs to minimize the reproduction of copyrighted content, adhering to a crucial balancing property to prevent the regurgitation of memorized data. Through extensive experiments, we show that CP-Fuse significantly reduces the reproduction of protected material without compromising the quality of text and code generation. Moreover, its post-hoc nature allows seamless integration with other protective measures, further enhancing copyright safeguards. Lastly, we show that CP-Fuse is robust against common techniques for extracting training data.
****
Rayhan Zirvi,Bahareh Tolooshams,Anima Anandkumar
Diffusion State-Guided Projected Gradient for Inverse Problems
Recent advancements in diffusion models have been effective in learning data priors for solving inverse problems. They leverage diffusion sampling steps for inducing a data prior while using a measurement guidance gradient at each step to impose data consistency. For general inverse problems, approximations are needed when an unconditionally trained diffusion model is used since the measurement likelihood is intractable, leading to inaccurate posterior sampling. In other words, due to their approximations, these methods fail to preserve the generation process on the data manifold defined by the diffusion prior, leading to artifacts in applications such as image restoration. To enhance the performance and robustness of diffusion models in solving inverse problems, we propose Diffusion State

-Guided Projected Gradient (DiffStateGrad), which projects the measurement gradi
ent onto a subspace that is a low-rank approximation of an intermediate state of
 the diffusion process. DiffStateGrad, as a module, can be added to a wide range
 of diffusion-based inverse solvers to improve the preservation of the diffusion
 process on the prior manifold and filter out artifact-inducing components. We h
ighlight that DiffStateGrad improves the robustness of diffusion models in terms
 of the choice of measurement guidance step size and noise while improving the w
orst-case performance. Finally, we demonstrate that DiffStateGrad improves upon
the state-of-the-art on linear and nonlinear image restoration inverse problems.
 Our code is available at https://github.com/Anima-Lab/DiffStateGrad.
****

Tuan Truong,Rithwik Sudharsan,Yibo Yang,Peter Xiangyuan Ma,Ruihan Yang,Stephan M
andt,Joshua S. Bloom

AstroCompress: A benchmark dataset for multi-purpose compression of astronomical
 data

The site conditions that make astronomical observatories in space and on the gro
und so desirable---cold and dark---demand a physical remoteness that leads to li
mited data transmission capabilities. Such transmission limitations directly bot
tleneck the amount of data acquired and in an era of costly modern observatories
, any improvements in lossless data compression has the potential scale to billi
ons of dollars worth of additional science that can be accomplished on the same
instrument. Traditional lossless methods for compressing astrophysical data are
manually designed. Neural data compression, on the other hand, holds the promise
 of learning compression algorithms end-to-end from data and outperforming class
ical techniques by leveraging the unique spatial, temporal, and wavelength struc
tures of astronomical images. This paper introduces [AstroCompress](https://hugg
ingface.co/AstroCompress): a neural compression challenge for astrophysics data,
 featuring four new datasets (and one legacy dataset) with 16-bit unsigned integ
er imaging data in various modes: space-based, ground-based, multi-wavelength, a
nd time-series imaging. We provide code to easily access the data and benchmark
seven lossless compression methods (three neural and four non-neural, including
 all practical state-of-the-art algorithms).
Our results on lossless compression indicate that lossless neural compression te
chniques can enhance data collection at observatories, and provide guidance on t
he adoption of neural compression in scientific applications. Though the scope o
f this paper is restricted to lossless compression, we also comment on the poten
tial exploration of lossy compression methods in future studies.
****

Yushi Bai,Jiajie Zhang,Xin Lv,Linzhi Zheng,Siqi Zhu,Lei Hou,Yuxiao Dong,Jie Tang
,Juanzi Li

LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs

Current long context large language models (LLMs) can process inputs up to 100,0
00 tokens, yet struggle to generate outputs exceeding even a modest length of 2,
000 words. Through controlled experiments, we find that the model's effective ge
neration length is inherently bounded by the sample it has seen during supervise
d fine-tuning (SFT). In other words, their output limitation is due to the scarc
ity of long-output examples in existing SFT datasets. To address this, we introd
uce AgentWrite, an agent-based pipeline that decomposes ultra-long generation ta
sks into subtasks, enabling off-the-shelf LLMs to generate coherent outputs exce
eding 20,000 words. Leveraging AgentWrite, we construct LongWriter-6k, a dataset
 containing 6,000 SFT data with output lengths ranging from 2k to 32k words. By
incorporating this dataset into model training, we successfully scale the output
 length of existing models to over 10,000 words while maintaining output quality
. We also develop LongBench-Write, a comprehensive benchmark for evaluating ultr
a-long generation capabilities. Our 9B parameter model, further improved through
 DPO, achieves state-of-the-art performance on this benchmark, surpassing even m
uch larger proprietary models. In general, our work demonstrates that existing l
ong context LLM already possesses the potential for a larger output window--all
you need is data with extended output during model alignment to unlock this capa
bility.

****
Harry Zhang,Luca Carlone
CHAMP: Conformalized 3D Human Multi-Hypothesis Pose Estimators
We introduce CHAMP, a novel method for learning sequence-to-sequence, multi-hypothesis 3D human poses from 2D keypoints by leveraging a conditional distribution with a diffusion model. To predict a single output 3D pose sequence, we generate and aggregate multiple 3D pose hypotheses. For better aggregation results, we develop a method to score these hypotheses during training, effectively integrating conformal prediction into the learning process. This process results in a differentiable conformal predictor that is trained end-to-end with the 3D pose estimator. Post-training, the learned scoring model is used as the conformity score, and the 3D pose estimator is combined with a conformal predictor to select the most accurate hypotheses for downstream aggregation. Our results indicate that using a simple mean aggregation on the conformal prediction-filtered hypotheses set yields competitive results. When integrated with more sophisticated aggregation techniques, our method achieves state-of-the-art performance across various metrics and datasets while inheriting the probabilistic guarantees of conformal prediction.
****
Liang Sun,Yang Zhang,Weizhao He,Jiajun Wen,Linlin Shen,Weicheng Xie
DeeperForward: Enhanced Forward-Forward Training for Deeper and Better Performance
While backpropagation effectively trains models, it presents challenges related to bio-plausibility, resulting in high memory demands and limited parallelism. Recently, Hinton (2022) proposed the Forward-Forward (FF) algorithm for high-parallel local updates. FF leverages squared sums as the local update target, termed goodness, and employs L2 normalization to decouple goodness and extract new features. However, this design encounters issues with feature scaling and deactivated neurons, limiting its application mainly to shallow networks. This paper proposes a novel goodness design utilizing **layer normalization** and **mean goodness** to overcome these challenges, demonstrating performance improvements even in 17-layer CNNs. Experiments on CIFAR-10, MNIST, and Fashion-MNIST show significant advantages over existing FF-based algorithms, highlighting the potential of FF in deep models. Furthermore, the model parallel strategy is proposed to achieve highly efficient training based on the property of local updates.
****
Jindong Tian,Yuxuan Liang,Ronghui Xu,Peng Chen,Chenjuan Guo,Aoying Zhou,Lujia Pan,Zhongwen Rao,Bin Yang
Air Quality Prediction with Physics-Guided Dual Neural ODEs in Open Systems
Air pollution significantly threatens human health and ecosystems, necessitating effective air quality prediction to inform public policy. Traditional approaches are generally categorized into physics-based and data-driven models. Physics-based models usually struggle with high computational demands and closed-system assumptions, while data-driven models may overlook essential physical dynamics, confusing the capturing of spatiotemporal correlations. Although some physics-guided approaches combine the strengths of both models, they often face a mismatch between explicit physical equations and implicit learned representations. To address these challenges, we propose Air-DualODE, a novel physics-guided approach that integrates dual branches of Neural ODEs for air quality prediction. The first branch applies open-system physical equations to capture spatiotemporal dependencies for learning physics dynamics, while the second branch identifies the dependencies not addressed by the first in a fully data-driven way. These dual representations are temporally aligned and fused to enhance prediction accuracy. Our experimental results demonstrate that Air-DualODE achieves state-of-the-art performance in predicting pollutant concentrations across various spatial scales, thereby offering a promising solution for real-world air quality challenges.
****
Yuxiao Lu,Arunesh Sinha,Pradeep Varakantham
Semantic Loss Guided Data Efficient Supervised Fine Tuning for Safe Responses in LLMs

Large Language Models (LLMs) generating unsafe responses to toxic prompts is a s
ignificant issue in their applications. While various efforts aim to address thi
s safety concern, previous approaches often demand substantial human data collec
tion or rely on the less dependable option of using another LLM to generate corr
ective data. In this paper, we aim to take this problem and overcome limitations
 of requiring significant high-quality human data. Our method requires only a sm
all set of unsafe responses to toxic prompts, easily obtained from the unsafe LL
M itself. By employing a semantic cost combined with a negative Earth Mover Dist
ance (EMD) loss, we guide the LLM away from generating unsafe responses. Additio
nally, we propose a novel lower bound for EMD loss, enabling more efficient opti
mization. Our results demonstrate superior performance and data efficiency compa
red to baselines, and we further examine the nuanced effects of over-alignment a
nd potential degradation of language capabilities when using contrastive data.
****

Mazda Moayeri,Vidhisha Balachandran,Varun Chandrasekaran,Safoora Yousefi,Thomas
FEL,Soheil Feizi,Besmira Nushi,Neel Joshi,Vibhav Vineet
Unearthing Skill-level Insights for Understanding Trade-offs of Foundation Model
s
With models getting stronger, evaluations have grown more complex, testing multi
ple skills in one benchmark and even in the same instance at once. However, skil
l-wise performance is obscured when inspecting aggregate accuracy, under-utilizi
ng the rich signal modern benchmarks contain. We propose an automatic approach t
o recover the underlying skills relevant for any evaluation instance, by way of
inspecting model-generated {\em rationales}. After validating the relevance of r
ationale-parsed skills and inferring skills for 46k instances over 12 benchm
arks, we observe many skills to be common across benchmarks, resulting in the cu
ration of hundreds of \emph{skill-slices} (i.e. sets of instances testing a comm
on skill). Inspecting accuracy over these slices yields novel insights on model
trade-offs: e.g., compared to GPT-4o and Claude 3.5 Sonnet, on average, Gemini 1
.5 Pro is 18% more accurate in \emph{computing molar mass}, but $19\\%$ less
accurate in \emph{applying constitutional law}, despite the overall accuracies o
f the three models differing by a mere $0.4\\%$. Furthermore, we demonstrate the
 practical utility of our approach by showing that insights derived from skill s
lice analysis can generalize to held-out instances: when routing each instance t
o the model strongest on the relevant skills, we see a $3\\%$ accuracy improveme
nt over our 12 dataset corpus. Our skill-slices and framework open a new avenu
e in model evaluation, leveraging skill-specific analyses to unlock a more granu
lar and actionable understanding of model capabilities.
****

Weikang Meng,Yadan Luo,Xin Li,Dongmei Jiang,Zheng Zhang
PolaFormer: Polarity-aware Linear Attention for Vision Transformers
Linear attention has emerged as a promising alternative to softmax-based attenti
on, leveraging kernelized feature maps to reduce complexity from quadratic to li
near in sequence length. However, the non-negative constraint on feature maps an
d the relaxed exponential function used in approximation lead to significant inf
ormation loss compared to the original query-key dot products, resulting in less
 discriminative attention maps with higher entropy. To address the missing inter
actions driven by negative values in query-key pairs, we propose a polarity-awar
e linear attention mechanism that explicitly models both same-signed and opposit
e-signed query-key interactions, ensuring comprehensive coverage of relational i
nformation. Furthermore, to restore the spiky properties of attention maps, we p
rovide a theoretical analysis proving the existence of a class of element-wise f
unctions (with positive first and second derivatives) that can reduce entropy in
 the attention distribution. For simplicity, and recognizing the distinct contri
butions of each dimension, we employ a learnable power function for rescaling, a
llowing strong and weak attention signals to be effectively separated. Extensive
 experiments demonstrate that the proposed PolaFormer improves performance on va
rious vision tasks, enhancing both expressiveness and efficiency by up to 4.6%.
****

Yuze Zhao,Tianyun Ji,Wenjun Feng,Zhenya Huang,Qi Liu,Zhiding Liu,Yixiao Ma,Kai Z

hang,Enhong Chen
Unveiling the Magic of Code Reasoning through Hypothesis Decomposition and Amend ment
The reasoning abilities are one of the most enigmatic and captivating aspects of large language models (LLMs). Numerous studies are dedicated to exploring and e xpanding the boundaries of this reasoning capability. However, tasks that embody both reasoning and recall characteristics are often overlooked. In this paper, we introduce such a novel task, code reasoning, to provide a new perspective for the reasoning abilities of LLMs.
We summarize three meta-benchmarks based on established forms of logical reasoni ng, and instantiate these into eight specific benchmark tasks. Our testing on th ese benchmarks reveals that LLMs continue to struggle with identifying satisfact ory reasoning pathways. Additionally, we present a new pathway exploration pipel ine inspired by human intricate problem-solving methods. This Reflective Hypothe sis Decomposition and Amendment (RHDA) pipeline consists of the following iterat ive steps: (1) Proposing potential hypotheses based on observations and decompos ing them; (2) Utilizing tools to validate hypotheses and reflection outcomes; (3) Revising hypothesis in light of observations. Our approach effectively mitigat es logical chain collapses arising from forgetting or hallucination issues in mu lti-step reasoning, resulting in performance gains of up to $3\times$. Finally, we expanded this pipeline by applying it to simulate complex household tasks in real-world scenarios, specifically in VirtualHome, enhancing the handling of fai lure cases. We release our code and all of results at https://github.com/TnTWoW/ code_reasoning.
****
Haiyang SHEN,Yue Li,Desong Meng,Dongqi Cai,Sheng Qi,Li Zhang,Mengwei Xu,Yun Ma
ShortcutsBench: A Large-Scale Real-world Benchmark for API-based Agents
Recent advancements in integrating large language models (LLMs) with application programming interfaces (APIs) have gained significant interest in both academia and industry. Recent work demonstrates that these API-based agents exhibit rela tively strong autonomy and planning capabilities. However, their ability to hand le multi-dimensional difficulty levels, diverse task types, and real-world deman ds remains unknown.
In this paper, we introduce ShortcutsBench, a large-scale benchmark for the comp rehensive evaluation of API-based agents in solving real-world complex tasks. Sh ortcutsBench includes a wealth of real APIs from Apple Inc., refined user querie s, human-annotated
high-quality action sequences, detailed parameter filling values, and parameters requesting necessary input from the system or user. We put in significant effor t in collecting and processing the data. We revealed how existing benchmarks / d atasets struggle to accommodate the advanced reasoning capabilities of existing more intelligent LLMs. Moreover, our extensive evaluation of agents built with 5 leading open-source (size >= 57B) and 5 closed-source LLMs (e.g. Gemini-1.5-Pro and GPT-4o-mini) reveals significant limitations of existing API-based agents i n the whole process of handling complex queries related to API selection, parame ter filling, and requesting necessary input from the system and the user. These findings highlight the great challenges that API-based agents face in effectivel y fulfilling real and complex user queries. All datasets, code, experimental log s, and results are available at \url{https://anonymous.4open.science/r/Shortcuts Bench}.
****
Dexiong Chen,Till Hendrik Schulz,Karsten Borgwardt
Learning Long Range Dependencies on Graphs via Random Walks
Message-passing graph neural networks (GNNs) excel at capturing local relationsh ips but struggle with long-range dependencies in graphs. In contrast, graph tran sformers (GTs) enable global information exchange but often oversimplify the gra ph structure by representing graphs as sets of fixed-length vectors. This work i ntroduces a novel architecture that overcomes the shortcomings of both approache s by combining the long-range information of random walks with local message pas sing. By treating random walks as sequences, our architecture leverages recent a

dvances in sequence models to effectively capture long-range dependencies within these walks. Based on this concept, we propose a framework that offers (1) more expressive graph representations through random walk sequences, (2) the ability to utilize any sequence model for capturing long-range dependencies, and (3) the flexibility by integrating various GNN and GT architectures. Our experimental evaluations demonstrate that our approach achieves competitive performance on 19 graph and node benchmark datasets, notably outperforming existing methods by up to 13\% on the PascalVoc-SP and COCO-SP datasets.
Code: https://github.com/BorgwardtLab/NeuralWalker
****

Huayu Chen,Hang Su,Peize Sun,Jun Zhu
Toward Guidance-Free AR Visual Generation via Condition Contrastive Alignment
Classifier-Free Guidance (CFG) is a critical technique for enhancing the sample quality of visual generative models. However, in autoregressive (AR) multi-modal generation, CFG introduces design inconsistencies between language and visual content, contradicting the design philosophy of unifying different modalities for visual AR. Motivated by language model alignment methods, we propose Condition Contrastive Alignment (CCA) to facilitate guidance-free AR visual generation. Unlike guidance methods that alter the sampling process to achieve the ideal sampling distribution, CCA directly fine-tunes pretrained models to fit the same distribution target. Experimental results show that CCA can significantly enhance the guidance-free performance of all tested models with just one epoch of fine-tuning (1% of pretraining epochs) on the pretraining dataset. This largely removes the need for guided sampling in AR visual generation and cuts the sampling cost by half. Moreover, by adjusting training parameters, CCA can achieve trade-offs between sample diversity and fidelity similar to CFG. This experimentally confirms the strong theoretical connection between language-targeted alignment and visual-targeted guidance methods, unifying two previously independent research fields.
****

Wayne Wu,Honglin He,Jack He,Yiran Wang,Chenda Duan,Zhizheng Liu,Quanyi Li,Bolei Zhou
MetaUrban: An Embodied AI Simulation Platform for Urban Micromobility
Public urban spaces such as streetscapes and plazas serve residents and accommodate social life in all its vibrant variations. Recent advances in robotics and embodied AI make public urban spaces no longer exclusive to humans. Food delivery bots and electric wheelchairs have started sharing sidewalks with pedestrians, while robot dogs and humanoids have recently emerged in the street. **Micromobility** enabled by AI for short-distance travel in public urban spaces plays a crucial component in future transportation systems. It is essential to ensure the generalizability and safety of AI models used for maneuvering mobile machines. In this work, we present **MetaUrban**, a *compositional* simulation platform for the AI-driven urban micromobility research. MetaUrban can construct an *infinite* number of interactive urban scenes from compositional elements, covering a vast array of ground plans, object placements, pedestrians, vulnerable road users, and other mobile agents' appearances and dynamics. We design point navigation and social navigation tasks as the pilot study using MetaUrban for urban micromobility research and establish various baselines of Reinforcement Learning and Imitation Learning. We conduct extensive evaluation across mobile machines, demonstrating that heterogeneous mechanical structures significantly influence the learning and execution of AI policies. We perform a thorough ablation study, showing that the compositional nature of the simulated environments can substantially improve the generalizability and safety of the trained mobile agents. MetaUrban will be made publicly available to provide research opportunities and foster safe and trustworthy embodied AI and micromobility in cities. The code and data have been released.
****

YiFan Zhang,Huanyu Zhang,Haochen Tian,Chaoyou Fu,Shuangqing Zhang,Junfei Wu,Feng Li,Kun Wang,Qingsong Wen,Zhang Zhang,Liang Wang,Rong Jin
MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Sc

enarios that are Difficult for Humans?
Comprehensive evaluation of Multimodal Large Language Models (MLLMs) has recentl y garnered widespread attention in the research community. However, we observe t hat existing benchmarks present several common barriers that make it difficult t o measure the significant challenges that models face in the real world, includi ng: 1) small data scale leads to a large performance variance; 2) reliance on mo del-based annotations results in restricted data quality; 3) insufficient task d ifficulty, especially caused by the limited image resolution. To tackle these is sues, we introduce MME-RealWorld. Specifically, we collect more than 300 K ima ges from public datasets and the Internet, filtering $13,366$ high-quality image s for annotation. This involves the efforts of professional 25 annotators and 7 experts in MLLMs, contributing to $29,429$ question-answer pairs that cover 43 subtasks across 5 real-world scenarios, extremely challenging even for hu mans. As far as we know, **MME-RealWorld is the largest manually annotated bench mark to date, featuring the highest resolution and a targeted focus on real-worl d applications**. We further conduct a thorough evaluation involving 29 promin ent MLLMs, such as GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Our results sh ow that even the most advanced models struggle with our benchmarks, where none o f them reach 60\% accuracy. The challenges of perceiving high-resolution images and understanding complex real-world scenarios remain urgent issues to be addres sed. The data and evaluation code are released in our Project Page.
****
Linlin Yu,Bowen Yang,Tianhao Wang,Kangshuo Li,Feng Chen
Predictive Uncertainty Quantification for Bird's Eye View Segmentation: A Benchm ark and Novel Loss Function
The fusion of raw sensor data to create a Bird's Eye View (BEV) representation i s critical for autonomous vehicle planning and control. Despite the growing inte rest in using deep learning models for BEV semantic segmentation, anticipating s egmentation errors and enhancing the explainability of these models remain under explored. This paper introduces a comprehensive benchmark for predictive uncerta inty quantification in BEV segmentation, evaluating multiple uncertainty quantif ication methods across three popular datasets with three representative network architectures. Our study focuses on the effectiveness of quantified uncertainty in detecting misclassified and out-of-distribution (OOD) pixels while also impro ving model calibration. Through empirical analysis, we uncover challenges in exi sting uncertainty quantification methods and demonstrate the potential of eviden tial deep learning techniques, which capture both aleatoric and epistemic uncert ainty. To address these challenges, we propose a novel loss function, Uncertaint y-Focal-Cross-Entropy (UFCE), specifically designed for highly imbalanced data, along with a simple uncertainty-scaling regularization term that improves both u ncertainty quantification and model calibration for BEV segmentation.
****
Panagiotis Theodoropoulos,Nikolaos Komianos,Vincent Pacelli,Guan-Horng Liu,Evang elos Theodorou
Feedback Schrödinger Bridge Matching
Recent advancements in diffusion bridges for distribution transport problems hav e heavily relied on matching frameworks, yet existing methods often face a trade -off between scalability and access to optimal pairings during training.
Fully unsupervised methods make minimal assumptions but incur high computational costs, limiting their practicality. On the other hand, imposing full supervisio n of the matching process with optimal pairings improves scalability, however, i t can be infeasible in most applications.
To strike a balance between scalability and minimal supervision, we introduce Fe edback Schrödinger Bridge Matching (FSBM), a novel semi-supervised matching fram ework that incorporates a small portion (<8% of the entire dataset) of pre-ali gned pairs as state feedback to guide the transport map of non-coupled samples, thereby significantly improving efficiency. This is achieved by formulating a st atic Entropic Optimal Transport (EOT) problem with an additional term capturing the semi-supervised guidance. The generalized EOT objective is then recast into a dynamic formulation to leverage the scalability of matching frameworks. Extens

ive experiments demonstrate that FSBM accelerates training and enhances generali
zation by leveraging coupled pairs' guidance, opening new avenues for training m
atching frameworks with partially aligned datasets.
****

Angelika Romanou,Negar Foroutan,Anna Sotnikova,Sree Harsha Nelaturu,Shivalika Si
ngh,Rishabh Maheshwary,Micol Altomare,Zeming Chen,Mohamed A. Haggag,Snegha A,Alf
onso Amayuelas,Azril Hafizi Amirudin,Danylo Boiko,Michael Chang,Jenny Chim,Gal C
ohen,Aditya Kumar Dalmia,Abraham Diress,Sharad Duwal,Daniil Dzenhaliou,Daniel Fe
rnando Erazo Florez,Fabian Farestam,Joseph Marvin Imperial,Shayekh Bin Islam,Per
ttu Isotalo,Maral Jabbarishiviari,Börje F. Karlsson,Eldar Khalilov,Christopher K
lamm,Fajri Koto,Dominik Krzemi■ski,Gabriel Adriano de Melo,Syrielle Montariol,Yi
yang Nan,Joel Niklaus,Jekaterina Novikova,Johan Samir Obando Ceron,Debjit Paul,E
sther Ploeger,Jebish Purbey,Swati Rajwal,Selvan Sunitha Ravi,Sara Rydell,Roshan
Santhosh,Drishti Sharma,Marjana Prifti Skenduli,Arshia Soltani Moakhar,Bardia so
ltani moakhar,Ayush Kumar Tarun,Azmine Toushik Wasi,Thenuka Ovin Weerasinghe,Ser
han Yilmaz,Mike Zhang,Imanol Schlag,Marzieh Fadaee,Sara Hooker,Antoine Bosselut
INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge
The performance differential of large language models (LLM) between languages hi
nders their effective deployment in many regions, inhibiting the potential econo
mic and societal value of generative AI tools in many communities. However, the
development of functional LLMs in many languages (i.e., multilingual LLMs) is bo
ttlenecked by the lack of high-quality evaluation resources in languages other t
han English. Moreover, current practices in multilingual benchmark construction
often translate English resources, ignoring the regional and cultural knowledge
of the environments in which multilingual systems would be used. In this work, w
e construct an evaluation suite of 197,243 QA pairs from local exam sources to m
easure the capabilities of multilingual LLMs in a variety of regional contexts.
Our novel resource, INCLUDE, is a comprehensive knowledge- and reasoning-centric
 benchmark across 44 written languages that evaluates multilingual LLMs for perf
ormance in the actual language environments where they would be deployed.
****

Xiao Han,Saima Absar,Lu Zhang,Shuhan Yuan
Root Cause Analysis of Anomalies in Multivariate Time Series through Granger Cau
sal Discovery
Identifying the root causes of anomalies in multivariate time series is challeng
ing due to the complex dependencies among the series. In this paper, we propose
a comprehensive approach called AERCA that inherently integrates Granger causal
discovery with root cause analysis. By defining anomalies as interventions on th
e exogenous variables of time series, AERCA not only learns the Granger causalit
y among time series but also explicitly models the distributions of exogenous va
riables under normal conditions. AERCA then identifies the root causes of anomal
ies by highlighting exogenous variables that significantly deviate from their no
rmal states. Experiments on multiple synthetic and real-world datasets demonstra
te that AERCA can accurately capture the causal relationships among time series
and effectively identify the root causes of anomalies.
****

Maxwell A Xu,Jaya Narain,Gregory Darnell,Haraldur T Hallgrimsson,Hyewon Jeong,Da
rren Forde,Richard Andres Fineman,Karthik Jayaraman Raghuram,James Matthew Rehg,
Shirley You Ren
RelCon: Relative Contrastive Learning for a Motion Foundation Model for Wearable
 Data
We present RelCon, a novel self-supervised Relative Contrastive learning approac
h for training a motion foundation model from wearable accelerometry sensors. Fi
rst, a learnable distance measure is trained to capture motif similarity and dom
ain-specific semantic information such as rotation invariance. Then, the learned
 distance provides a measurement of semantic similarity between a pair of accele
rometry time-series, which we use to train our foundation model to model relativ
e relationships across time and across subjects. The foundation model is trained
 on 1 billion segments from 87,376 participants, and achieves strong performance
 across multiple downstream tasks, including human activity recognition and gait

metric regression. To our knowledge, we are the first to show the generalizabil
ity of a foundation model with motion data from wearables across distinct evalua
tion tasks.
****
Anish Mudide,Joshua Engels,Eric J Michaud,Max Tegmark,Christian Schroeder de Wit
t
Efficient Dictionary Learning with Switch Sparse Autoencoders
Sparse autoencoders (SAEs) are a recent technique for decomposing neural network
activations into human-interpretable features. However, in order for SAEs to id
entify all features represented in frontier models, it will be necessary to scal
e them up to very high width, posing a computational challenge. In this work, we
introduce Switch Sparse Autoencoders, a novel SAE architecture aimed at reducin
g the compute cost of training SAEs. Inspired by sparse mixture of experts model
s, Switch SAEs route activation vectors between smaller "expert" SAEs, enabling
SAEs to efficiently scale to many more features. We present experiments comparin
g Switch SAEs with other SAE architectures, and find that Switch SAEs deliver a
substantial Pareto improvement in the reconstruction vs. sparsity frontier for a
given fixed training compute budget. We also study the geometry of features acr
oss experts, analyze features duplicated across experts, and verify that Switch
SAE features are as interpretable as features found by other SAE architectures.
****
Patrik Reizinger,Siyuan Guo,Ferenc Huszár,Bernhard Schölkopf,Wieland Brendel
Identifiable Exchangeable Mechanisms for Causal Structure and Representation Lea
rning
Identifying latent representations or causal structures is important for good ge
neralization and downstream task performance. However, both fields developed rat
her independently.
We observe that several structure and representation identifiability methods, pa
rticularly those that require multiple environments, rely on
exchangeable non--i.i.d. (independent and identically distributed) data.
To formalize this connection,
we propose the Identifiable Exchangeable Mechanisms (IEM) framework to unify key
representation and causal structure learning methods. IEM provides a unified pr
obabilistic graphical model encompassing causal discovery, Independent Component
Analysis, and Causal Representation Learning.
With the help of the IEM model, we generalize the Causal de Finetti theorem of G
uo et al., 2022 by relaxing the necessary conditions for causal structure identi
fication in exchangeable data.
We term these conditions cause and mechanism variability, and show how they impl
y a duality condition in identifiable representation learning, leading to new id
entifiability results.
****
Yiming Wang,Pei Zhang,Baosong Yang,Derek F. Wong,Rui Wang
Latent Space Chain-of-Embedding Enables Output-free LLM Self-Evaluation
LLM self-evaluation relies on the LLM's own ability to estimate response correct
ness, which can greatly improve its deployment reliability.
In this research track, we propose the Chain-of-Embedding (CoE) in the latent sp
ace to enable LLMs to perform output-free self-evaluation. CoE consists of all p
rogressive hidden states produced during the inference time, which can be treate
d as the latent thinking path of LLMs. We find that when LLMs respond correctly
and incorrectly, their CoE features differ, these discrepancies assist us in est
imating LLM response correctness. Experiments in four diverse domains and seven
LLMs fully demonstrate the effectiveness of our method. Meanwhile, its label-fre
e design intent without any training and millisecond-level computational cost en
sure real-time feedback in large-scale scenarios.
More importantly, we provide interesting insights into LLM response correctness
from the perspective of hidden state changes inside LLMs.
****
Zhen Qin,Zhuqing Liu,Songtao Lu,Yingbin Liang,Jia Liu
DUET: Decentralized Bilevel Optimization without Lower-Level Strong Convexity

Decentralized bilevel optimization (DBO) provides a powerful framework for multi-agent systems to solve local bilevel tasks in a decentralized fashion without the need for a central server.
However, most existing DBO methods rely on lower-level strong convexity (LLSC) to guarantee unique solutions and a well-defined hypergradient for stationarity measure, hindering their applicability in many practical scenarios not satisfying LLSC.
To overcome this limitation, we introduce a new single-loop DBO algorithm called diminishing quadratically-regularized bilevel decentralized optimization (DUET), which eliminates the need for LLSC by introducing a diminishing quadratic regularization to the lower-level (LL) objective.
We show that DUET achieves an iteration complexity of $O(1/T^{1-5p-\frac{11}{4}\tau})$ for approximate KKT-stationary point convergence under relaxed assumptions, where p and τ are control parameters for LL learning rate and averaging, respectively.
In addition, our DUET algorithm incorporates gradient tracking to address data heterogeneity, a key challenge in DBO settings.
To the best of our knowledge, this is the first work to tackle DBO without LLSC under decentralized settings with data heterogeneity.
Numerical experiments validate the theoretical findings and demonstrate the practical effectiveness of our proposed algorithms.
****
Spencer Frei,Gal Vardi
Trained Transformer Classifiers Generalize and Exhibit Benign Overfitting In-Context
Transformers have the capacity to act as supervised learning algorithms: by properly encoding a set of labeled training (''in-context'') examples and an unlabeled test example into an input sequence of vectors of the same dimension, the forward pass of the transformer can produce predictions for that unlabeled test example. A line of recent work has shown that when linear transformers are pre-trained on random instances for linear regression tasks, these trained transformers make predictions using an algorithm similar to that of ordinary least squares.
In this work, we investigate the behavior of linear transformers trained on random linear classification tasks. Via an analysis of the implicit regularization of gradient descent, we characterize how many pre-training tasks and in-context examples are needed for the trained transformer to generalize well at test-time. We further show that in some settings, these trained transformers can exhibit ''benign overfitting in-context'': when in-context examples are corrupted by label flipping noise, the transformer memorizes all of its in-context examples (including those with noisy labels) yet still generalizes near-optimally for clean test examples.
****
Shobhita Sundaram,Julia Chae,Yonglong Tian,Sara Beery,Phillip Isola
Personalized Representation from Personalized Generation
Modern vision models excel at general purpose downstream tasks. It is unclear, however, how they may be used for personalized vision tasks, which are both fine-grained and data-scarce. Recent works have successfully applied synthetic data to general-purpose representation learning, while advances in T2I diffusion models have enabled the generation of personalized images from just a few real examples. Here, we explore a potential connection between these ideas, and formalize the challenge of using personalized synthetic data to learn personalized representations, which encode knowledge about an object of interest and may be flexibly applied to any downstream task relating to the target object. We introduce an evaluation suite for this challenge, including reformulations of two existing datasets and a novel dataset explicitly constructed for this purpose, and propose a contrastive learning approach that makes creative use of image generators. We show that our method improves personalized representation learning for diverse downstream tasks, from recognition to segmentation, and analyze characteristics of image generation approaches that are key to this gain.
****

Hao Cui,Zahra Shamsi,Gowoon Cheon,Xuejian Ma,Shutong Li,Maria Tikhanovskaya,Peter Christian Norgaard,Nayantara Mudur,Martyna Beata Plomecka,Paul Raccuglia,Yasaman Bahri,Victor V. Albert,Pranesh Srinivasan,Haining Pan,Philippe Faist,Brian A Rohr,Michael J. Statt,Dan Morris,Drew Purves,Elise Kleeman,Ruth Alcantara,Matthew Abraham,Muqthar Mohammad,Ean Phing VanLee,Chenfei Jiang,Elizabeth Dorfman,Eun-Ah Kim,Michael Brenner,Sameera S Ponda,Subhashini Venugopalan
CURIE: Evaluating LLMs on Multitask Scientific Long-Context Understanding and Reasoning
Scientific problem-solving involves synthesizing information while applying expert knowledge. We introduce CURIE, a scientific long-Context Understanding, Reasoning, and Information Extraction benchmark to measure the potential of Large Language Models (LLMs) in scientific problem-solving and assisting scientists in realistic workflows. This benchmark introduces ten challenging tasks with a total of 580 problems and solution pairs curated by experts in six disciplines - materials science, condensed matter physics, quantum computing, geo-spatial analysis, biodiversity, and proteins - covering both experimental and theoretical work-flows in science. We evaluate a range of closed and open LLMs on tasks in CURIE which requires domain expertise, comprehension of long in-context information,and multi-step reasoning. While Gemini Flash 2.0 and Claude-3 show consistent high comprehension across domains, the popular GPT-4o and command-R+ fail dramatically on protein sequencing tasks. With the best performance at 32% there is much room for improvement for all models. We hope that insights gained from CURIE can guide the future development of LLMs in sciences. Links to the data and evaluation code are in https://github.com/google/curie
****
Runpeng Yu,Xinchao Wang
Neural Phylogeny: Fine-Tuning Relationship Detection among Neural Networks
Given a collection of neural networks, can we determine which are parent models and which are child models fine-tuned from the parents?
In this work, we strive to answer this question
via introducing a new task termed as neural phylogeny detection, aimed at identifying the existence and direction of the fine-tuning relationship. Specifically, neural phylogeny detection attempts to identify all parent-child model pairs and determine, within each pair, which model is the parent and which is the child. We present two approaches for neural phylogeny detection: a learning-free method and a learning-based method. First, we propose a metric that leverages the distance from network parameters to a fake initialization to infer fine-tuning directions. By integrating this metric with traditional clustering algorithms, we propose a series of efficient, learning-free neural phylogeny detection methods. Second, we introduce a transformer-based neural phylogeny detector, which significantly enhances detection accuracy through a learning-based manner. Extensive experiments, ranging from shallow fully-connected networks to open-sourced Stable Diffusion and LLaMA models, progressively validate the effectiveness of both methods. The results demonstrate the reliability of both the learning-free and the learning-based approaches across various learning tasks and network architectures, as well as their ability to detect cross-generational phylogeny between ancestor models and their fine-tuned descendants.
****
Hongsin Lee,Seungju Cho,Changick Kim
Indirect Gradient Matching for Adversarial Robust Distillation
Adversarial training significantly improves adversarial robustness, but superior performance is primarily attained with large models.
This substantial performance gap for smaller models has spurred active research into adversarial distillation (AD) to mitigate the difference.
Existing AD methods leverage the teacher’s logits as a guide.
In contrast to these approaches, we aim to transfer another piece of knowledge from the teacher, the input gradient.
In this paper, we propose a distillation module termed Indirect Gradient Distillation Module (IGDM) that indirectly matches the student’s input gradient with that of the teacher.

Experimental results show that IGDM seamlessly integrates with existing AD methods, significantly enhancing their performance.
Particularly, utilizing IGDM on the CIFAR-100 dataset improves the AutoAttack accuracy from 28.06\% to 30.32\% with the ResNet-18 architecture and from 26.18\% to 29.32\% with the MobileNetV2 architecture when integrated into the SOTA method without additional data augmentation.
****
Zheng Chong,Xiao Dong,Haoxiang Li,shiyue Zhang,Wenqing Zhang,Hanqing Zhao,xujie zhang,Dongmei Jiang,Xiaodan Liang
CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models
Virtual try-on methods based on diffusion models achieve realistic effects but often require additional encoding modules, a large number of training parameters, and complex preprocessing, which increases the burden on training and inference. In this work, we re-evaluate the necessity of additional modules and analyze how to improve training efficiency and reduce redundant steps in the inference process. Based on these insights, we propose CatVTON, a simple and efficient virtual try-on diffusion model that transfers in-shop or worn garments of arbitrary categories to target individuals by concatenating them along spatial dimensions as inputs of the diffusion model. The efficiency of CatVTON is reflected in three aspects: (1) Lightweight network. CatVTON consists only of a VAE and a simplified denoising UNet, removing redundant image and text encoders as well as cross-attentions, and includes just 899.06M parameters. (2) Parameter-efficient training. Through experimental analysis, we identify self-attention modules as crucial for adapting pre-trained diffusion models to the virtual try-on task, enabling high-quality results with only 49.57M training parameters. (3) Simplified inference. CatVTON eliminates unnecessary preprocessing, such as pose estimation, human parsing, and captioning, requiring only a person image and garment reference to guide the virtual try-on process, reducing over 49% memory usage compared to other diffusion-based methods. Extensive experiments demonstrate that CatVTON achieves superior qualitative and quantitative results compared to baseline methods and demonstrates strong generalization performance in in-the-wild scenarios, despite being trained solely on public datasets with 73K samples.
****
Kang Liao,Zongsheng Yue,Zhouxia Wang,Chen Change Loy
Denoising as Adaptation: Noise-Space Domain Adaptation for Image Restoration
Although learning-based image restoration methods have made significant progress, they still struggle with limited generalization to real-world scenarios due to the substantial domain gap caused by training on synthetic data. Existing methods address this issue by improving data synthesis pipelines, estimating degradation kernels, employing deep internal learning, and performing domain adaptation and regularization. Previous domain adaptation methods have sought to bridge the domain gap by learning domain-invariant knowledge in either feature or pixel space. However, these techniques often struggle to extend to low-level vision tasks within a stable and compact framework. In this paper, we show that it is possible to perform domain adaptation via the noise space using diffusion models. In particular, by leveraging the unique property of how auxiliary conditional inputs influence the multi-step denoising process, we derive a meaningful *diffusion loss* that guides the restoration model in progressively aligning both restored synthetic and real-world outputs with a target clean distribution. We refer to this method as *denoising as adaptation*. To prevent shortcuts during joint training, we present crucial strategies such as channel-shuffling layer and residual-swapping contrastive learning in the diffusion model. They implicitly blur the boundaries between conditioned synthetic and real data and prevent the reliance of the model on easily distinguishable features. Experimental results on three classical image restoration tasks, namely denoising, deblurring, and deraining, demonstrate the effectiveness of the proposed method.
****
Jiahan Li,Tong Chen,Shitong Luo,Chaoran Cheng,Jiaqi Guan,Ruihan Guo,Sheng Wang,Ge Liu,Jian Peng,Jianzhu Ma
Hotspot-Driven Peptide Design via Multi-Fragment Autoregressive Extension

Peptides, short chains of amino acids, interact with target proteins, making the m a unique class of protein-based therapeutics for treating human diseases. Rece ntly, deep generative models have shown great promise in peptide generation. How ever, several challenges remain in designing effective peptide binders. First, n ot all residues contribute equally to peptide-target interactions. Second, the g enerated peptides must adopt valid geometries due to the constraints of peptide bonds. Third, realistic tasks for peptide drug development are still lacking.
To address these challenges, we introduce PepHAR, a hot-spot-driven autoregressi ve generative model for designing peptides targeting specific proteins. Building on the observation that certain hot spot residues have higher interaction poten tials, we first use an energy-based density model to fit and sample these key re sidues. Next, to ensure proper peptide geometry, we autoregressively extend pept ide fragments by estimating dihedral angles between residue frames. Finally, we apply an optimization process to iteratively refine fragment assembly, ensuring correct peptide structures.
By combining hot spot sampling with fragment-based extension, our approach enabl es \textit{de novo} peptide design tailored to a target protein and allows the i ncorporation of key hot spot residues into peptide scaffolds. Extensive experime nts, including peptide design and peptide scaffold generation, demonstrate the s trong potential of PepHAR in computational peptide binder design. The source cod e will be available at https://github.com/Ced3-han/PepHAR.
****
Lise Le Boudec,Emmanuel de Bezenac,Louis Serrano,Ramon Daniel Regueiro-Espino,Yu an Yin,Patrick Gallinari
Learning a Neural Solver for Parametric PDEs to Enhance Physics-Informed Methods
Physics-informed deep learning often faces optimization challenges due to the co mplexity of solving partial differential equations (PDEs), which involve explori ng large solution spaces, require numerous iterations, and can lead to unstable training. These challenges arise particularly from the ill-conditioning of the o ptimization problem, caused by the differential terms in the loss function. To a ddress these issues, we propose learning a solver, i.e., solving PDEs using a ph ysics-informed iterative algorithm trained on data. Our method learns to conditi on a gradient descent algorithm that automatically adapts to each PDE instance, significantly accelerating and stabilizing the optimization process and enabling faster convergence of physics-aware models. Furthermore, while traditional phys ics-informed methods solve for a single PDE instance, our approach addresses par ametric PDEs. Specifically, our method integrates the physical loss gradient wit h the PDE parameters to solve over a distribution of PDE parameters, including c oefficients, initial conditions, or boundary conditions. We demonstrate the effe ctiveness of our method through empirical experiments on multiple datasets, comp aring training and test-time optimization performance.
****
Zhenting Qi,Hongyin Luo,Xuliang Huang,Zhuokai Zhao,Yibo Jiang,Xiangjun Fan,Himab indu Lakkaraju,James R. Glass
Quantifying Generalization Complexity for Large Language Models
While large language models (LLMs) have shown exceptional capabilities in unders tanding complex queries
and performing sophisticated tasks, their generalization abilities are often dee ply entangled with
memorization, necessitating more precise evaluation.
To address this challenge, we introduce Scylla, a dynamic evaluation framework t hat quantitatively measures the generalization abilities of LLMs. Scylla disenta ngles generalization from memorization via assessing model performance on both i n-distribution (ID) and out-of-distribution (OOD) data through 20 tasks across 5 levels of complexity.
Through extensive experiments, we uncover a non-monotonic relationship between t ask complexity and the performance gap between ID and OOD
data, which we term the generalization valley.
Specifically, this phenomenon reveals a critical threshold---referred to
as critical complexity---where reliance on non-generalizable behavior peaks, ind

icating the
upper bound of LLMs' generalization capabilities.
As model size increases, the critical complexity shifts toward higher levels of task complexity,
suggesting that larger models can handle more complex reasoning tasks before over-relying on
memorization.
Leveraging Scylla and the concept of critical complexity, we benchmark 28 LLMs including
both open-sourced models such as LLaMA and Qwen families, and closed-sourced models like Claude and
GPT, providing a more robust evaluation and establishing a clearer
understanding of LLMs' generalization capabilities.
****
Noah Marshall,Ke Liang Xiao,Atish Agarwala,Elliot Paquette
To Clip or not to Clip: the Dynamics of SGD with Gradient Clipping in High-Dimensions
The success of modern machine learning is due in part to the adaptive optimization methods that have been developed to deal with the difficulties of training large models over complex datasets. One such method is gradient clipping: a practical procedure with limited theoretical underpinnings. In this work, we study clipping in a least squares problem under streaming SGD. We develop a theoretical analysis of the learning dynamics in the limit of large intrinsic dimension—a model and dataset dependent notion of dimensionality. In this limit we find a deterministic equation that describes the evolution of the loss and demonstrate that this equation predicts the path of clipped SGD on synthetic, CIFAR10, and Wikitext2 data. We show that with Gaussian noise clipping cannot improve SGD performance. Yet, in other noisy settings, clipping can provide benefits with tuning of the clipping threshold. We propose a simple heuristic for near optimal scheduling of the clipping threshold which requires the tuning of only one hyperparameter. We conclude with a discussion about the links between high-dimensional clipping and neural network training.
****
Andre Cornman,Jacob West-Roberts,Antonio Pedro Camargo,Simon Roux,Martin Beracochea,Milot Mirdita,Sergey Ovchinnikov,Yunha Hwang
The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling
Biological language model performance depends heavily on pretraining data quality, diversity, and size. While metagenomic datasets feature enormous biological diversity, their utilization as pretraining data has been limited due to challenges in data accessibility, quality filtering and deduplication. Here, we present the Open MetaGenomic (OMG) corpus, a genomic pretraining dataset totalling 3.1T base pairs and 3.3B protein coding sequences, obtained by combining two largest metagenomic dataset repositories (JGI's IMG and EMBL's MGnify). We first document the composition of the dataset and describe the quality filtering steps taken to remove poor quality data. We make the OMG corpus available as a mixed-modality genomic sequence dataset that represents multi-gene encoding genomic sequences with translated amino acids for protein coding sequences, and nucleic acids for intergenic sequences. We train the first mixed-modality genomic language model (gLM2) that leverages genomic context information to learn robust functional representations, as well as coevolutionary signals in protein-protein interfaces and genomic regulatory syntax. Furthermore, we show that deduplication in embedding space can be used to balance the corpus, demonstrating improved performance on downstream tasks. The OMG dataset is publicly hosted on the Hugging Face Hub at https://huggingface.co/datasets/tattabio/OMG and gLM2 is available at https://huggingface.co/tattabio/gLM2_650M.
****
Sam Gunn,Xuandong Zhao,Dawn Song
An Undetectable Watermark for Generative Image Models
We present the first undetectable watermarking scheme for generative image model

s.
Undetectability ensures that no efficient adversary can distinguish between wa
termarked and un-watermarked images, even after making many adaptive queries.
In particular, an undetectable watermark does not degrade image quality under an
y efficiently computable metric.
Our scheme works by selecting the initial latents of a diffusion model using a p
seudorandom error-correcting code (Christ and Gunn, 2024), a strategy which guar
antees undetectability and robustness. We experimentally demonstrate that our wa
termarks are quality-preserving and robust using Stable Diffusion 2.1.
Our experiments verify that, in contrast to _every prior scheme_ we tested, our
watermark does not degrade image quality.
Our experiments also demonstrate robustness: existing watermark removal attacks
fail to remove our watermark from images without significantly degrading the qua
lity of the images.
Finally, we find that we can robustly encode 512 bits in our watermark, and up t
o 2500 bits when the images are not subjected to watermark removal attacks.
Our code is available at https://github.com/XuandongZhao/PRC-Watermark.
****
Ningya Feng,Junwei Pan,Jialong Wu,Baixu Chen,Ximei Wang,QianLi,Xian Hu,Jie Jiang
,Mingsheng Long
Long-Sequence Recommendation Models Need Decoupled Embeddings
Lifelong user behavior sequences are crucial for capturing user interests and pr
edicting user responses in modern recommendation systems. A two-stage paradigm i
s typically adopted to handle these long sequences: a subset of relevant behavio
rs is first searched from the original long sequences via an attention mechanism
 in the first stage and then aggregated with the target item to construct a disc
riminative representation for prediction in the second stage. In this work, we i
dentify and characterize, for the first time, a neglected deficiency in existing
 long-sequence recommendation models: a single set of embeddings struggles with
learning both attention and representation, leading to interference between thes
e two processes. Initial attempts to address this issue with some common methods
 (e.g., linear projections---a technique borrowed from language processing) prov
ed ineffective, shedding light on the unique challenges of recommendation models
. To overcome this, we propose the Decoupled Attention and Representation Embedd
ings (DARE) model, where two distinct embedding tables are initialized and learn
ed separately to fully decouple attention and representation. Extensive experime
nts and analysis demonstrate that DARE provides more accurate searches of correl
ated behaviors and outperforms baselines with AUC gains up to 9‰ on public datas
ets and notable improvements on Tencent's advertising platform. Furthermore, dec
oupling embedding spaces allows us to reduce the attention embedding dimension a
nd accelerate the search procedure by 50\% without significant performance impac
t, enabling more efficient, high-performance online serving. Code in PyTorch for
 experiments, including model analysis, is available at https://github.com/thuml
/DARE.
****
Lichang Chen,Hexiang Hu,Mingda Zhang,Yiwen Chen,Zifeng Wang,YANDONG LI,Pranav Sh
yam,Tianyi Zhou,Heng Huang,Ming-Hsuan Yang,Boqing Gong
OmnixR: Evaluating Omni-modality Language Models on Reasoning across Modalities
We introduce \textbf{OmnixR}, an evaluation suite designed to benchmark state-of
-the-art Omni-modality Language Models (OLMs), such as GPT-4o and Gemini.
Evaluating OLMs, which integrate multiple modalities such as text, vision, and a
udio, presents unique challenges.
Particularly, the user message might often consist of multiple modalities, such
that OLMs have to establish holistic understanding and reasoning across modaliti
es to accomplish the task.
Existing benchmarks are limited to single-modality or dual-modality tasks (e.g.,
 image+text or video+text), overlooking comprehensive multi-modal assessments of
 model reasoning.
To address this, OmnixR offers two evaluation variants: (1) OmnixR-synth: a synt
hetic dataset generated automatically by translating text into multiple modaliti

es—audio, images, video, and hybrids Omnify!. (2) OmnixR-real: a real-world data set, manually curated and annotated by experts, for evaluating cross-modal reasoning in natural settings.
OmnixR presents a unique evaluation towards assessing OLMs over a diverse mix of modalities, such as a question that involves video, audio, and text, providing a rigorous cross-modal reasoning testbed than any existing benchmarks.
Our experiments find that all state-of-the-art OLMs struggles with OmnixR questions that require integrating information from multiple modalities to answer.
Further analysis highlight differences in reasoning behavior and underscoring the challenges of omni-modal AI alignment.
****
Fangkai Jiao,Geyang Guo,Xingxing Zhang,Nancy F. Chen,Shafiq Joty,Furu Wei
Preference Optimization for Reasoning with Pseudo Feedback
Preference optimization techniques, such as Direct Preference Optimization (DPO), are frequently employed to enhance the reasoning capabilities of large language models (LLMs) in domains like mathematical reasoning and coding, typically following supervised fine-tuning. These methods rely on high-quality labels for reasoning tasks to generate preference pairs; however, the availability of reasoning datasets with human-verified labels is limited.
In this study, we introduce a novel approach to generate pseudo feedback for reasoning tasks by framing the labeling of solutions to reason problems as an evaluation against associated \emph{test cases}.
We explore two forms of pseudo feedback based on test cases: one generated by frontier LLMs and the other by extending self-consistency to multi-test-case.
We conduct experiments on both mathematical reasoning and coding tasks using pseudo feedback for preference optimization, and observe improvements across both tasks. Specifically, using Mathstral-7B as our base model, we improve MATH results from 58.3 to 68.6, surpassing both NuminaMath-72B and GPT-4-Turbo-1106-preview. In GSM8K and College Math, our scores increase from 85.6 to 90.3 and from 34.3 to 42.3, respectively. Building on Deepseek-coder-7B-v1.5, we achieve a score of 24.3 on LiveCodeBench (from 21.1), surpassing Claude-3-Haiku.
****
Tung-Yu Wu,Melody Lo
U-shaped and Inverted-U Scaling behind Emergent Abilities of Large Language Models
Large language models (LLMs) have been shown to exhibit *emergent abilities* in some downstream tasks, where model performance stagnates at first and then improves sharply and unpredictably with scale beyond a threshold. In this work, we investigate the phenomenon by grouping questions based on difficulty level and provide a possible explanation for emergent abilities. Specifically, we observe U-shaped scaling for hard questions and inverted-U scaling followed by steady improvement for easy questions. The two scaling patterns initially offset each other, causing stagnant overall performance. The performance starts to soar when the scaling pattern of easy questions reverts from inverse to standard scaling, leading to emergent abilities. Based on this finding, we propose a simple yet effective pipeline, called *Slice-and-Sandwich*, to predict the emergence threshold and model performance beyond the threshold. Our code is publicly available at https://github.com/tony10101105/ExpEmergence.
****
Jiasheng Ye,Peiju Liu,Tianxiang Sun,Jun Zhan,Yunhua Zhou,Xipeng Qiu
Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance
Pretraining data of large language models composes multiple domains (e.g., web texts, academic papers, codes), whose mixture proportions crucially impact the competence of outcome models. While existing endeavors rely on heuristics or qualitative strategies to tune the proportions, we discover the quantitative predictability of model performance regarding the mixture proportions in function forms, which we refer to as the data mixing laws. Fitting such functions on sample mixtures unveils model performance on unseen mixtures before actual runs, thus guiding the selection of an ideal data mixture. Furthermore, we propose nested use o

f the scaling laws of training steps, model sizes, and our data mixing laws to p
redict the performance of large models trained on massive data under various mix
tures with only small-scale training. Experimental results verify that our meth
od effectively optimizes the training mixture of a 1B model trained for 100B tok
ens in RedPajama, reaching a performance comparable to the one trained for 48% m
ore steps on the default mixture. Extending the application of data mixing laws
to continual training accurately predicts the critical mixture proportion that a
voids catastrophic forgetting and outlooks the potential for dynamic data schedu
les.
****

Matteo Pagliardini,Pierre Ablin,David Grangier
The AdEMAMix Optimizer: Better, Faster, Older
Momentum based optimizers are central to a wide range of machine learning applic
ations. These typically rely on an Exponential Moving Average (EMA) of gradients
, which decays exponentially the present contribution of older gradients. This a
ccounts for gradients being local linear approximations which lose their relevan
ce as the iterate moves along the loss landscape. This work questions the use of
 a single EMA to accumulate past gradients and empirically demonstrates how this
 choice can be sub-optimal: a single EMA cannot simultaneously give a high weigh
t to the immediate past, and a non-negligible weight to older gradients. Buildin
g on this observation, we propose AdEMAMix, a simple modification of the Adam op
timizer with a mixture of two EMAs to better take advantage of past gradients. O
ur experiments on language modeling and image classification show---quite surpri
singly---that gradients can stay relevant for tens of thousands of steps. They h
elp to converge faster, and often to lower minima: e.g., a 1.3B parameter AdEM
AMix LLM trained on 101B tokens performs comparably to an AdamW model trained
on 197B tokens ($+95\%$). Moreover, our method significantly slows-down model
forgetting during training. Our work motivates further exploration of different
types of functions to leverage past gradients, beyond EMAs.
****

Teng Xiao,Yige Yuan,Zhengyu Chen,Mingxiao Li,Shangsong Liang,Zhaochun Ren,Vasant
 G Honavar
SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters
Existing preference optimization objectives for language model alignment require
 additional hyperparameters that must be extensively tuned to achieve optimal pe
rformance, increasing both the complexity and time required for fine-tuning larg
e language models. In this paper, we propose a simple yet effective hyperparamet
er-free preference optimization algorithm for alignment. We observe that promisi
ng performance can be achieved simply by optimizing inverse perplexity, which is
 calculated as the inverse of the exponentiated average log-likelihood of the ch
osen and rejected responses in the preference dataset. The resulting simple lear
ning objective, SimPER, is easy to implement and eliminates the need for expensi
ve hyperparameter tuning and a reference model, making it both computationally a
nd memory efficient. Extensive experiments on widely used real-world benchmarks,
 including MT-Bench, AlpacaEval 2, and 10 key benchmarks of the Open LLM Leaderb
oard with 5 base models, demonstrate that SimPER consistently and significantly
outperforms existing approaches—even without any hyperparameters or a reference
model. For example, despite its simplicity, SimPER outperforms state-of-the-art
methods by up to 5.7 points on AlpacaEval 2 and achieves the highest average ran
king across 10 benchmarks on the Open LLM Leaderboard. The source code for SimPE
R is publicly available at: https://github.com/tengxiao1/SimPER.
****

Qingtao Liu,Yu Cui,Zhengnan Sun,Gaofeng Li,Jiming Chen,Qi Ye
VTDexManip: A Dataset and Benchmark for Visual-tactile Pretraining and Dexterous
 Manipulation with Reinforcement Learning
Vision and touch are the most commonly used senses in human manipulation. While
leveraging human manipulation videos for robotic task pretraining has shown prom
ise in prior works, it is limited to image and language modalities and deploymen
t to simple parallel grippers. In this paper, aiming to address the limitations,
 we collect a vision-tactile dataset by humans manipulating 10 daily tasks and 1

82 objects. In contrast with the existing datasets, our dataset is the first visual-tactile dataset for complex robotic manipulation skill learning. Also, we introduce a novel benchmark, featuring six complex dexterous manipulation tasks and a reinforcement learning-based vision-tactile skill learning framework. 18 non-pretraining and pretraining methods within the framework are designed and compared to investigate the effectiveness of different modalities and pertaining strategies. Key findings based on our benchmark results and analyses experiments include: 1) Despite the tactile modality used in our experiments being binary and sparse, including it directly in the policy training boosts the success rate by about 20\% and joint pretraining it with vision gains a further 20\%. 2) Joint pretraining visual-tactile modalities exhibits strong adaptability in unknown tasks and achieves robust performance among all tasks. 3) Using binary tactile signals with vision is robust to viewpoint setting, tactile noise, and the binarization threshold, which facilitates to the visual-tactile policy to be deployed in reality. The dataset and benchmark are available at \url{https://github.com/LQTS/VTDexManip}.
****

Vitalis Vosylius,Edward Johns
Instant Policy: In-Context Imitation Learning via Graph Diffusion
Following the impressive capabilities of in-context learning with large transformers, In-Context Imitation Learning (ICIL) is a promising opportunity for robotics. We introduce Instant Policy, which learns new tasks instantly from just one or two demonstrations, achieving ICIL through two key components. First, we introduce inductive biases through a graph representation and model ICIL as a graph generation problem using a learned diffusion process, enabling structured reasoning over demonstrations, observations, and actions. Second, we show that such a model can be trained using pseudo-demonstrations – arbitrary trajectories generated in simulation – as a virtually infinite pool of training data. Our experiments, in both simulation and reality, show that Instant Policy enables rapid learning of various everyday robot tasks. We also show how it can serve as a foundation for cross-embodiment and zero-shot transfer to language-defined tasks.
****

Hongyu Guo,Yoshua Bengio,Shengchao Liu
AssembleFlow: Rigid Flow Matching with Inertial Frames for Molecular Assembly
Molecular assembly, where a cluster of rigid molecules aggregated into strongly correlated forms, is fundamental to determining the properties of materials. However, traditional numerical methods for simulating this process are computationally expensive, and existing generative models on material generation overlook the rigidity inherent in molecular structures, leading to unwanted distortions and invalid internal structures in molecules. To address this, we introduce AssembleFlow. AssembleFlow leverages inertial frames to establish reference coordinate systems at the molecular level for tracking the orientation and motion of molecules within the cluster. It further decomposes molecular $\text{SE}(3)$ transformations into translations in $\mathbb{R}^3$ and rotations in $\text{SO}(3)$, enabling explicit enforcement of both translational and rotational rigidity during each generation step within the flow matching framework. This decomposition also empowers distinct probability paths for each transformation group, effectively allowing for the separate learning of their velocity functions: the former, moving in Euclidean space, uses linear interpolation (LERP), while the latter, evolving in spherical space, employs spherical linear interpolation (SLERP) with a closed-form solution. Empirical validation on the benchmarking data COD-Cluster17 shows that AssembleFlow significantly outperforms six competitive deep learning baselines by at least 45\% in assembly matching scores while maintaining 100\% molecular integrity. Also, it matches the assembly performance of a widely used domain-specific simulation tool while reducing computational cost by 25-fold.
****

Ziwei Yang,Zheng Chen,Xin Liu,Rikuto Kotoge,Peng Chen,Yasuko Matsubara,Yasushi Sakurai,Jimeng Sun
GeSubNet: Gene Interaction Inference for Disease Subtype Network Generation
Retrieving gene functional networks from knowledge databases presents a challeng

e due to the mismatch between disease networks and subtype-specific variations. Current solutions, including statistical and deep learning methods, often fail to effectively integrate gene interaction knowledge from databases or explicitly learn subtype-specific interactions. To address this mismatch, we propose GeSubNet, which learns a unified representation capable of predicting gene interactions while distinguishing between different disease subtypes. Graphs generated by such representations can be considered subtype-specific networks. GeSubNet is a multi-step representation learning framework with three modules: First, a deep generative model learns distinct disease subtypes from patient gene expression profiles. Second, a graph neural network captures representations of prior gene networks from knowledge databases, ensuring accurate physical gene interactions. Finally, we integrate these two representations using an inference loss that leverages graph generation capabilities, conditioned on the patient separation loss, to refine subtype-specific information in the learned representation. GeSubNet consistently outperforms traditional methods, with average improvements of 30.6%, 21.0%, 20.1%, and 56.6% across four graph evaluation metrics, averaged over four cancer datasets. Particularly, we conduct a biological simulation experiment to assess how the behavior of selected genes from over 11,000 candidates affects subtypes or patient distributions. The results show that the generated network has the potential to identify subtype-specific genes with an 83% likelihood of impacting patient distribution shifts.
****
Sungwoo Jeong,Alex Townsend
Extending Mercer's expansion to indefinite and asymmetric kernels
Mercer's expansion and Mercer's theorem are cornerstone results in kernel theory. While the classical Mercer's theorem only considers continuous symmetric positive definite kernels, analogous expansions are effective in practice for indefinite and asymmetric kernels. In this paper we extend Mercer's expansion to continuous kernels, providing a rigorous theoretical underpinning for indefinite and asymmetric kernels. We begin by demonstrating that Mercer's expansion may not be pointwise convergent for continuous indefinite kernels, before proving that the expansion of continuous kernels with bounded variation uniformly in each variable separably converges pointwise almost everywhere, almost uniformly, and unconditionally almost everywhere. We also describe an algorithm for computing Mercer's expansion for general kernels and give new decay bounds on its terms.
****
Chenyu Zhou,Mengdan Zhang,Peixian Chen,Chaoyou Fu,Yunhang Shen,Xiawu Zheng,Xing Sun,Rongrong Ji
Learning Interleaved Image-Text Comprehension in Vision-Language Large Models
The swift progress of Multi-modal Large Models (MLLMs) has showcased their impressive ability to tackle tasks blending vision and language.
Yet, most current models and benchmarks cater to scenarios with a narrow scope of visual and textual contexts.
These models often fall short when faced with complex comprehension tasks, which involve navigating through a plethora of irrelevant and potentially misleading information in both text and image forms.
To bridge this gap, we introduce a new, more demanding task known as Interleaved Image-Text Comprehension (IITC).
This task challenges models to discern and disregard superfluous elements in both images and text to accurately answer questions and to follow intricate instructions to pinpoint the relevant image.
In support of this task, we further craft a new VEGA dataset, tailored for the IITC task on scientific content, and devised a subtask, Image-Text Association (ITA), to refine image-text correlation skills.
Our evaluation of four leading closed-source models, as well as various open-source models using VEGA, underscores the rigorous nature of IITC.
Even the most advanced models, such as Gemini-1.5-pro and GPT4V, only achieved modest success.
By employing a multi-task, multi-scale post-training strategy, we have set a robust baseline for MLLMs on the IITC task, attaining an 85.8% accuracy rate in

image association and a 0.508 Rouge score. These results validate the effectiv
eness of our dataset in improving MLLMs capabilities for nuanced image-text comp
rehension.
****
Ching-Yun Ko,Pin-Yu Chen,Payel Das,Youssef Mroueh,Soham Dan,Georgios Kollias,Sub
hajit Chaudhury,Tejaswini Pedapati,Luca Daniel
Large Language Models can Become Strong Self-Detoxifiers
Reducing the likelihood of generating harmful and toxic output is an essential t
ask when aligning large language models (LLMs). Existing methods mainly rely on
training an external reward model (i.e., another language model) or fine-tuning
the LLM using self-generated data to influence the outcome. In this paper, we sh
ow that LLMs have the capability of self-detoxification without external reward
model learning or retraining of the LM. We propose \textit{Self-disciplined Auto
regressive Sampling (SASA)}, a lightweight controlled decoding algorithm for tox
icity reduction of LLMs. SASA leverages the contextual representations from an L
LM to learn linear subspaces from labeled data characterizing toxic v.s. non-tox
ic output in analytical forms. When auto-completing a response token-by-token, S
ASA dynamically tracks the margin of the current output to steer the generation
away from the toxic subspace, by adjusting the autoregressive sampling strategy.
 Evaluated on LLMs of different scale and nature, namely Llama-3.1-Instruct (8B)
, Llama-2 (7B), and GPT2-L models with the RealToxicityPrompts, BOLD, and AttaQ
benchmarks, SASA markedly enhances the quality of the generated sentences relati
ve to the original models and attains comparable performance to state-of-the-art
 detoxification techniques, significantly reducing the toxicity level by only us
ing the LLM's internal representations.
****
Zhaochong An,Guolei Sun,Yun Liu,Runjia Li,Min Wu,Ming-Ming Cheng,Ender Konukoglu
,Serge Belongie
Multimodality Helps Few-shot 3D Point Cloud Semantic Segmentation
Few-shot 3D point cloud segmentation (FS-PCS) aims at generalizing models to seg
ment novel categories with minimal annotated support samples. While existing FS-
PCS methods have shown promise, they primarily focus on unimodal point cloud inp
uts, overlooking the potential benefits of leveraging multimodal information. In
 this paper, we address this gap by introducing a multimodal FS-PCS setup, utili
zing textual labels and the potentially available 2D image modality. Under this
easy-to-achieve setup, we present the MultiModal Few-Shot SegNet (MM-FSS), a mod
el effectively harnessing complementary information from multiple modalities. MM
-FSS employs a shared backbone with two heads to extract intermodal and unimodal
 visual features, and a pretrained text encoder to generate text embeddings. To
fully exploit the multimodal information, we propose a Multimodal Correlation Fu
sion (MCF) module to generate multimodal correlations, and a Multimodal Semantic
 Fusion (MSF) module to refine the correlations using text-aware semantic guidan
ce. Additionally, we propose a simple yet effective Test-time Adaptive Cross-mod
al Calibration (TACC) technique to mitigate training bias, further improving gen
eralization. Experimental results on S3DIS and ScanNet datasets demonstrate sign
ificant performance improvements achieved by our method. The efficacy of our app
roach indicates the benefits of leveraging commonly-ignored free modalities for
FS-PCS, providing valuable insights for future research. The code is available a
t github.com/ZhaochongAn/Multimodality-3D-Few-Shot.
****
Hojoon Lee,Dongyoon Hwang,Donghu Kim,Hyunseung Kim,Jun Jet Tai,Kaushik Subramani
an,Peter R. Wurman,Jaegul Choo,Peter Stone,Takuma Seno
SimBa: Simplicity Bias for Scaling Up Parameters in Deep Reinforcement Learning
Recent advances in CV and NLP have been largely driven by scaling up the number
of network parameters, despite traditional theories suggesting that larger netwo
rks are prone to overfitting.
These large networks avoid overfitting by integrating components that induce a s
implicity bias, guiding models toward simple and generalizable solutions.
However, in deep RL, designing and scaling up networks have been less explored.
Motivated by this opportunity, we present SimBa, an architecture designed to sca

le up parameters in deep RL by injecting a simplicity bias. SimBa consists of th
ree components: (i) an observation normalization layer that standardizes inputs
with running statistics, (ii) a residual feedforward block to provide a linear p
athway from the input to output, and (iii) a layer normalization to control feat
ure magnitudes.
By scaling up parameters with SimBa, the sample efficiency of various deep RL al
gorithms—including off-policy, on-policy, and unsupervised methods—is consistent
ly improved.
Moreover, solely by integrating SimBa architecture into SAC, it matches or surpa
sses state-of-the-art deep RL methods with high computational efficiency across
DMC, MyoSuite, and HumanoidBench.
These results demonstrate SimBa's broad applicability and effectiveness across d
iverse RL algorithms and environments.
****
Ye Li,Wenzhao Zheng,Xiaonan Huang,Kurt Keutzer
UniDrive: Towards Universal Driving Perception Across Camera Configurations
Vision-centric autonomous driving has demonstrated excellent performance with ec
onomical sensors. As the fundamental step, 3D perception aims to infer 3D inform
ation from 2D images based on 3D-2D projection. This makes driving perception mo
dels susceptible to sensor configuration (e.g., camera intrinsics and extrinsics
) variations. However, generalizing across camera configurations is important fo
r deploying autonomous driving models on different car models. In this paper, we
 present UniDrive, a novel framework for vision-centric autonomous driving to ac
hieve universal perception across camera configurations. We deploy a set of unif
ied virtual cameras and propose a ground-aware projection method to effectively
transform the original images into these unified virtual views. We further propo
se a virtual configuration optimization method by minimizing the expected projec
tion error between original cameras and virtual cameras. The proposed virtual ca
mera projection can be applied to existing 3D perception methods as a plug-and-p
lay module to mitigate the challenges posed by camera parameter variability, res
ulting in more adaptable and reliable driving perception models. To evaluate the
 effectiveness of our framework, we collect a dataset on CARLA by driving the sa
me routes while only modifying the camera configurations. Experimental results d
emonstrate that our method trained on one specific camera configuration can gene
ralize to varying configurations with minor performance degradation.
****
Xuehao Wang,Liyuan Wang,Binghuai Lin,Yu Zhang
HeadMap: Locating and Enhancing Knowledge Circuits in LLMs
Large language models (LLMs), through pretraining on extensive corpora, encompas
s rich semantic knowledge and exhibit the potential for efficient adaptation to
diverse downstream tasks. However, the intrinsic mechanisms underlying LLMs rema
in unexplored, limiting the efficacy of applying these models to downstream task
s. In this paper, we explore the intrinsic mechanisms of LLMs from the perspecti
ve of knowledge circuits. Specifically, considering layer dependencies, we propo
se a layer-conditioned locating algorithm to identify a series of attention head
s, which is a knowledge circuit of some tasks. Experiments demonstrate that simp
ly masking a small portion of attention heads in the knowledge circuit can signi
ficantly reduce the model's ability to make correct predictions. This suggests t
hat the knowledge flow within the knowledge circuit plays a critical role when t
he model makes a correct prediction. Inspired by this observation, we propose a
novel parameter-efficient fine-tuning method called HeadMap, which maps the acti
vations of these critical heads in the located knowledge circuit to the residual
 stream by two linear layers, thus enhancing knowledge flow from the knowledge c
ircuit in the residual stream. Extensive experiments conducted on diverse datase
ts demonstrate the efficiency and efficacy of the proposed method. Our code is a
vailable at https://github.com/XuehaoWangFi/HeadMap.
****
Kim Sung-Bin,Oh Hyun-Bin,JungMok Lee,Arda Senocak,Joon Son Chung,Tae-Hyun Oh
AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language
Models

Following the success of Large Language Models (LLMs), expanding their boundaries to new modalities represents a significant paradigm shift in multimodal understanding. Human perception is inherently multimodal, relying not only on text but also on auditory and visual cues for a complete understanding of the world. In recognition of this fact, audio-visual LLMs have recently emerged. Despite promising developments, the lack of dedicated benchmarks poses challenges for understanding and evaluating models. In this work, we show that audio-visual LLMs struggle to discern subtle relationships between audio and visual signals, leading to hallucinations and highlighting the need for reliable benchmarks. To address this, we introduce AVHBench, the first comprehensive benchmark specifically designed to evaluate the perception and comprehension capabilities of audio-visual LLMs. Our benchmark includes tests for assessing hallucinations,
as well as the cross-modal matching and reasoning abilities of these models. Our results reveal that most existing audio-visual LLMs struggle with hallucinations caused by cross-interactions between modalities, due to their limited capacity to perceive complex multimodal signals and their relationships. Additionally, we demonstrate that simple training with our AVHBench improves robustness of audio-visual LLMs against hallucinations. Dataset: https://github.com/kaist-ami/AVHBench
****

Lijie Fan,Tianhong Li,Siyang Qin,Yuanzhen Li,Chen Sun,Michael Rubinstein,Deqing Sun,Kaiming He,Yonglong Tian

Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens

Scaling up autoregressive models in vision has not proven as beneficial as in large language models.
In this work, we investigate this scaling problem in the context of text-to-image generation, focusing on two critical factors: whether models use discrete or continuous tokens, and whether tokens are generated in a random or fixed raster order using BERT- or GPT-like transformer architectures. Our empirical results show that, while all models scale effectively in terms of validation loss, their evaluation performance -- measured by FID, GenEval score, and visual quality -- follows different trends. Models based on continuous tokens achieves significantly better visual quality than those using discrete tokens. Furthermore, the generation order and attention mechanisms significantly affect the GenEval score: random-order models achieve notably better GenEval scores compared to raster-order models.
Inspired by these findings, we train Fluid, a random-order autoregressive model on continuous tokens. Fluid 10.5B model achieves a new state-of-the-art zeor-shot FID of 6.16 on MS-COCO 30K, and 0.69 overall score on the GenEval benchmark. We hope our findings and results will encourage future efforts to further bridge the scaling gap between vision and language models.
****

Ricardo Dominguez-Olmedo,Florian E. Dorner,Moritz Hardt

Training on the Test Task Confounds Evaluation and Emergence

We study a fundamental problem in the evaluation of large language models that we call training on the test task. Unlike wrongful practices like training on the test data, leakage, or data contamination, training on the test task is not a malpractice. Rather, the term describes a growing set of techniques to include task-relevant data in the pretraining stage of a language model. We demonstrate that training on the test task confounds both relative model evaluations and claims about emergent capabilities. We argue that the seeming superiority of one model family over another may be explained by a different degree of training on the test task. To this end, we propose an effective method to adjust for the effect of training on the test task on benchmark evaluations. Put simply, to fine-tune each model under comparison on the same task-relevant data before evaluation. Lastly, we show that instances of emergent behavior disappear gradually as models train on the test task. Our work promotes a new perspective on the evaluation of large language models with broad implications for benchmarking and the study of emergent capabilities.

****
Hiroshi Takahashi,Tomoharu Iwata,Atsutoshi Kumagai,Yuuki Yamanaka,Tomoya Yamashi
ta
Positive-Unlabeled Diffusion Models for Preventing Sensitive Data Generation
Diffusion models are powerful generative models but often generate sensitive dat
a that are unwanted by users,
mainly because the unlabeled training data frequently contain such sensitive dat
a.
Since labeling all sensitive data in the large-scale unlabeled training data is
impractical,
we address this problem by using a small amount of labeled sensitive data.
In this paper,
we propose positive-unlabeled diffusion models,
which prevent the generation of sensitive data using unlabeled and sensitive dat
a.
Our approach can approximate the evidence lower bound (ELBO) for normal (negativ
e) data using only unlabeled and sensitive (positive) data.
Therefore, even without labeled normal data,
we can maximize the ELBO for normal data and minimize it for labeled sensitive d
ata,
ensuring the generation of only normal data.
Through experiments across various datasets and settings,
we demonstrated that our approach can prevent the generation of sensitive images
 without compromising image quality.
****
Omer Madmon,Idan Pipano,Itamar Reinman,Moshe Tennenholtz
On the Convergence of No-Regret Dynamics in Information Retrieval Games with Pro
portional Ranking Functions
Publishers who publish their content on the web act strategically, in a behavior
 that can be modeled within the online learning framework.
Regret, a central concept in machine learning, serves as a canonical measure for
 assessing the performance of learning agents within this framework.
We prove that any proportional content ranking function with a concave activatio
n function induces games in which no-regret learning dynamics converge.
Moreover, for proportional ranking functions, we prove the equivalence of the co
ncavity of the activation function, the social concavity of the induced games an
d the concavity of the induced games.
We also study the empirical trade-offs between publishers' and users' welfare, u
nder different choices of the activation function, using a state-of-the-art no-r
egret dynamics algorithm. Furthermore, we demonstrate how the choice of the rank
ing function and changes in the ecosystem structure affect these welfare measure
s, as well as the dynamics' convergence rate.
****
Qiushi Huang,Tom Ko,Lilian Tang,Yu Zhang
ComLoRA: A Competitive Learning Approach for Enhancing LoRA
We propose a Competitive Low-Rank Adaptation (ComLoRA) framework to address the
limitations of the LoRA method, which either lacks capacity with a single rank-$
r$ LoRA or risks inefficiency and overfitting with a larger rank-Kr LoRA, wher
e K is an integer larger than 1. The proposed ComLoRA method initializes K d
istinct LoRA components, each with rank r, and allows them to compete during t
raining. This competition drives each LoRA component to outperform the others, i
mproving overall model performance. The best-performing LoRA is selected based o
n validation metrics, ensuring that the final model outperforms a single rank-$r
$ LoRA and matches the effectiveness of a larger rank-Kr LoRA, all while avoid
ing extra computational overhead during inference. To the best of our knowledge,
 this is the first work to introduce and explore competitive learning in the con
text of LoRA optimization. The ComLoRA's code is available at https://github.com
/hqsiswiliam/comlora.
****
Zhengrui Guo,Fangxu Zhou,Wei Wu,Qichen Sun,Lishuang Feng,Jinzhuo Wang,Hao Chen

BLEND: Behavior-guided Neural Population Dynamics Modeling via Privileged Knowle
dge Distillation
Modeling the nonlinear dynamics of neuronal populations represents a key pursuit
 in computational neuroscience. Recent research has increasingly focused on join
tly modeling neural activity and behavior to unravel their interconnections. Des
pite significant efforts, these approaches often necessitate either intricate mo
del designs or oversimplified assumptions. Given the frequent absence of perfect
ly paired neural-behavioral datasets in real-world scenarios when deploying thes
e models, a critical yet understudied research question emerges: how to develop
a model that performs well using only neural activity as input at inference, whi
le benefiting from the insights gained from behavioral signals during training?

To this end, we propose **BLEND**, the **B**ehavior-guided neura**L** population
 dynamics mod**E**lling framework via privileged k**N**owledge **D**istillation.
 By considering behavior as privileged information, we train a teacher model tha
t takes both behavior observations (privileged features) and neural activities (
regular features) as inputs. A student model is then distilled using only neural
 activity. Unlike existing methods, our framework is model-agnostic and avoids m
aking strong assumptions about the relationship between behavior and neural acti
vity. This allows BLEND to enhance existing neural dynamics modeling architectur
es without developing specialized models from scratch. Extensive experiments acr
oss neural population activity modeling and transcriptomic neuron identity predi
ction tasks demonstrate strong capabilities of BLEND, reporting over 50% improve
ment in behavioral decoding and over 15% improvement in transcriptomic neuron id
entity prediction after behavior-guided distillation. Furthermore, we empiricall
y explore various behavior-guided distillation strategies within the BLEND frame
work and present a comprehensive analysis of effectiveness and implications for
model performance. Code will be made available at https://github.com/dddavid4rea
l/BLEND.
****
Suraj Anand,Michael A. Lepori,Jack Merullo,Ellie Pavlick
Dual Process Learning: Controlling Use of In-Context vs. In-Weights Strategies w
ith Weight Forgetting
Language models have the ability to perform in-context learning (ICL), allowing
 them to flexibly adapt their behavior based on context. This contrasts with in-w
eights learning (IWL), where memorized information is encoded in model paramete
rs after iterated observations of data. An ideal model should be able to flexibl
y deploy both of these abilities. Despite their apparent ability to learn in-con
text, language models are known to struggle when faced with unseen or rarely see
n tokens (Land & Bartolo, 2024). Hence, we study $\textbf{structural in-context
 learning}$, which we define as the ability of a model to execute in-context lea
rning on arbitrary novel tokens -- so called because the model must generalize o
n the basis of e.g. sentence structure or task structure, rather than content en
coded in token embeddings. We study structural in-context algorithms on both syn
thetic and naturalistic tasks using toy models, masked language models, and auto
regressive language models. We find that structural ICL appears before quickly d
isappearing early in LM pretraining. While it has been shown that ICL can dimini
sh during training (Singh et al., 2023), we find that prior work does not accoun
t for structural ICL. Building on Chen et al. (2024) 's active forgetting method
, we introduce pretraining and finetuning methods that can modulate the preferen
ce for structural ICL and IWL. Importantly, this allows us to induce a $\textit{
dual process strategy}$ where in-context and in-weights solutions coexist within
 a single model.
****
Jerry Yao-Chieh Hu,Wei-Po Wang,Ammar Gilani,Chenyang Li,Zhao Song,Han Liu
Fundamental Limits of Prompt Tuning Transformers: Universality, Capacity and Eff
iciency
We investigate the statistical and computational limits of prompt tuning for tra
nsformer-based foundation models. Our key contributions are that prompt tuning
on *single-head* transformers with only a *single* self-attention layer: (i) is

universal, and (ii) supports efficient (even almost-linear time) algorithms under the Strong Exponential Time Hypothesis (SETH). Statistically, we prove that prompt tuning on such the simplest possible transformers are universal approximators for sequence-to-sequence Lipschitz functions. In addition, we provide an exponential-in-dL and -in-$(1/\epsilon)$ lower bound on the required soft-prompt tokens for prompt tuning to memorize any dataset with 1-layer, 1-head transformers. Computationally, we identify a phase transition in the efficiency of prompt tuning, determined by the norm of the *soft-prompt-induced* keys and queries, and provide an upper bound criterion. Beyond this criterion, no sub-quadratic (efficient) algorithm for prompt tuning exists under SETH. Within this criterion, we showcase our theory by proving the existence of almost-linear time prompt tuning inference algorithms. These fundamental limits provide important necessary conditions for designing expressive and efficient prompt tuning methods for practitioners.
****
Jian Gao,Weidong Cao,Junyi Yang,Xuan Zhang
AnalogGenie: A Generative Engine for Automatic Discovery of Analog Circuit Topologies
The massive and large-scale design of foundational semiconductor integrated circuits (ICs) is crucial to sustaining the advancement of many emerging and future technologies, such as generative AI, 5G/6G, and quantum computing.
Excitingly, recent studies have shown the great capabilities of foundational models in expediting the design of digital ICs.
Yet, applying generative AI techniques to accelerate the design of analog ICs remains a significant challenge due to critical domain-specific issues, such as the lack of a comprehensive dataset and effective representation methods for analog circuits.
This paper proposes, $\textbf{AnalogGenie}$, a $\underline{\textbf{Gen}}$erat$\underline{\textbf{i}}$ve $\underline{\textbf{e}}$ngine for automatic design/discovery of $\underline{\textbf{Analog}}$ circuit topologies--the most challenging and creative task in the conventional manual design flow of analog ICs.
AnalogGenie addresses two key gaps in the field: building a foundational comprehensive dataset of analog circuit topology and developing a scalable sequence-based graph representation universal to analog circuits.
Experimental results show the remarkable generation performance of AnalogGenie in broadening the variety of analog ICs, increasing the number of devices within a single design, and discovering unseen circuit topologies far beyond any prior arts.
Our work paves the way to transform the longstanding time-consuming manual design flow of analog ICs to an automatic and massive manner powered by generative AI.
Our source code is available at https://github.com/xz-group/AnalogGenie.
****
Aditya Ramesh,Shivam Bhardwaj,Aditya Saibewar,Manohar Kaul
EFFICIENT JAILBREAK ATTACK SEQUENCES ON LARGE LANGUAGE MODELS VIA MULTI-ARMED BANDIT-BASED CONTEXT SWITCHING
Content warning: This paper contains examples of harmful language and content.
Recent advances in large language models (LLMs) have made them increasingly vulnerable to jailbreaking attempts, where malicious users manipulate models into generating harmful content. While existing approaches rely on either single-step attacks that trigger immediate safety responses or multi-step methods that inefficiently iterate prompts using other LLMs, we introduce ``Sequence of Context" (SoC) attacks that systematically alter conversational context through strategically crafted context-switching queries (CSQs). We formulate this as a multi-armed bandit (MAB) optimization problem, automatically learning optimal sequences of CSQs that gradually weaken the model's safety boundaries. Our theoretical analysis provides tight bounds on both the expected sequence length until successful jailbreak and the convergence of cumulative rewards. Empirically, our method achieves a 95\% attack success rate, surpassing PAIR by 63.15\%, AutoDAN by 60\%, and ReNeLLM by 50\%. We evaluate our attack across multiple open-source LLMs includ

ing Llama and Mistral variants. Our findings highlight critical vulnerabilities in current LLM safeguards and emphasize the need for defenses that consider sequential attack patterns rather than relying solely on static prompt filtering or iterative refinement.
****
Tianming Zhao,Chunqiu xia,Xiaomin Chang,Chunhao Li,Wei Li,Albert Zomaya
Competitive Fair Scheduling with Predictions
Beyond the worst-case analysis of algorithms, the learning-augmented framework considers that an algorithm can leverage possibly imperfect predictions about the unknown variables to have guarantees tied to the prediction quality. We consider online non-clairvoyant scheduling to minimize the max-stretch under this framework, where the scheduler can access job size predictions. We present a family of algorithms: Relaxed-Greedy (RG) with an $O(\eta^3 \cdot \sqrt{P})$ competitive ratio, where η denotes the prediction error for job sizes and P the maximum job size ratio; Adaptive Relaxed-Greedy with an $O(\lambda^{0.5} \cdot \eta^{2.5} \cdot \sqrt{P})$ competitive ratio, where λ denotes the error for the minimum job size; Predictive Relaxed-Greedy with an $O(\lambda^{0.5} \cdot \varphi^{0.5} \cdot \eta \cdot \max \\\{ \eta, \varphi \\\} \cdot \sqrt{P})$ competitive ratio, where φ denotes the error for the maximum job size. We also present *${RG}^x$*, an algorithm that represents a trade-off between consistency and smoothness, with an $O(\eta^{2+2x} \cdot P^{1-x})$ competitive ratio. We introduce a general method using resource augmentation to bound robustness, resulting in *RR*-augmented *RG*, with a $(1 + \epsilon)$-speed $O(\min \\\{ \eta^3 \sqrt{P}, \frac{n}{\epsilon} \\\})$ competitive ratio. Finally, we conduct simulations on synthetic and real-world datasets to evaluate the practical performance of these algorithms.
****
QIUHAO Zeng,Jerry Huang,Peng Lu,Gezheng Xu,Boxing Chen,Charles Ling,Boyu Wang
ZETA: Leveraging Z-order Curves for Efficient Top-k Attention
Over recent years, the Transformer has become a fundamental building block for sequence modeling architectures. Yet at its core is the use of self-attention, whose memory and computational cost grow quadratically with the sequence length N, rendering it prohibitively expensive for long sequences. A promising approach is top-k attention, which selects only the k most relevant tokens and achieves performance comparable to vanilla self-attention while significantly reducing space and computational demands. However, causal masks require the current query token to only attend to past tokens, preventing existing top-k attention methods from efficiently searching for the most relevant tokens in parallel, thereby limiting training efficiency. In this work, we propose ZETA, leveraging Z-Order Curves for Efficient Top-k Attention, to enable parallel querying of past tokens for entire sequences. We first theoretically show that the choice of key and query dimensions involves a trade-off between the curse of dimensionality and the preservation of relative distances after projection. In light of this insight, we propose reducing the dimensionality of keys and queries in contrast to values and further leveraging Z-order curves to map low-dimensional keys and queries into one-dimensional space, which permits parallel sorting, thereby largely improving the efficiency for top-k token selection. Experimental results demonstrate that ZETA~matches the performance of standard attention on synthetic tasks Associative Recall and outperforms attention and its variants on Long-Range Arena and WikiText-103 language modeling.
****
Chengyi Cai,Zesheng Ye,Lei Feng,Jianzhong Qi,Feng Liu
Attribute-based Visual Reprogramming for Vision-Language Models
Visual reprogramming (VR) reuses pre-trained vision models for downstream image classification tasks by adding trainable noise patterns to inputs. When applied to vision-language models (e.g., CLIP), existing VR approaches follow the same pipeline used in vision models (e.g., ResNet, ViT), where ground-truth class labels are inserted into fixed text templates to guide the optimization of VR patterns. This label-based approach, however, overlooks the rich information and diverse attribute-guided textual representations that CLIP can exploit, which may l

ead to the misclassification of samples. In this paper, we propose ***Attr**ibut e-based **V**isual **R**eprogramming* (AttrVR) for CLIP, utilizing ***des**cript ive **attr**ibutes* (DesAttrs) and ***dist**inctive **attr**ibutes* (DistAttrs), which respectively represent common and unique feature descriptions for differe nt classes. Besides, as images of the same class may reflect different attribute s after VR, AttrVR iteratively refines patterns using the k-nearest DesAttrs a nd DistAttrs for each image sample, enabling more dynamic and sample-specific op timization. Theoretically, AttrVR is shown to reduce intra-class variance and in crease inter-class separation. Empirically, it achieves superior performance in 12 downstream tasks for both ViT-based and ResNet-based CLIP. The success of Att rVR facilitates more effective integration of VR from unimodal vision models int o vision-language models. Our code is available at https://github.com/tmlr-group /AttrVR.
****

Dorian Guyot,Alexandra Anna Lassota

Minimalistic Predictions for Online Class Constraint Scheduling

We consider online scheduling with class constraints. That is, we are given m machines, each with k class slots. Upon receiving a job j with class c_j, an algorithm needs to allocate j on some machine i. The goal is to minimize the makespan while not assigning more than k different classes onto each machi ne.

While the offline case is well understood and even (E)PTAS results are known [Ja nsen, Lassota, Maack SPAA'20, Chen Jansen Luo Zhang COCOA'16], the online case a dmits strong impossibility results in classical competitive analysis [Epstein, L assota, Levin, Maack, Rohwedder STACS'22].

We overcome these daunting results by investigating the problem in a learning-au gmented setting where an algorithm can access possibly erroneous predictions. We present new algorithms with competitive ratios independent of m and tight low er bounds for several classical and problem-specific prediction models. We there by give a structured overview of what additional information helps in the design of better scheduling algorithms.
****

Bedionita Soro,Bruno Andreis,Hayeon Lee,Wonyong Jeong,Song Chong,Frank Hutter,Su ng Ju Hwang

Diffusion-based Neural Network Weights Generation

Transfer learning is a cornerstone of modern deep learning, yet it remains const rained by challenges in model selection and the overhead of extensive model stor age. In this work, we present Diffusion-based Neural Network Weights Generation, D2NWG, a novel framework that leverages diffusion processes to synthesize task-specific network weights. By modeling the distribution of weights from a diverse ensemble of pretrained models and conditioning the generation process on datase t characteristics, task descriptions, and architectural specifications, D2NWG ci rcumvents the need for storing and searching through massive model repositories. We evaluate D2NWG across multiple experimental settings. On in-distribution tas ks, our framework achieves performance that is on par with or superior to conven tional pretrained models, while also serving as an effective initialization stra tegy for novel domains, resulting in faster convergence and a 6\% improvement in few-shot learning scenarios. Extensive ablation studies further indicate that o ur approach scales robustly with increased diversity and volume of pretrained mo dels. Moreover, D2NWG demonstrates significant promise for large language model applications. In evaluations on the OpenLM leaderboard, our method improved LLaM A-3-2-1B-Instruct performance by 3\% on challenging mathematical reasoning tasks , with a consistent gain of 0.36\% across a range of benchmarks. These findings establish D2NWG as a versatile and powerful framework for neural network weight generation, offering a scalable solution to the limitations of traditional trans fer learning.
****

Kepan Nan,Rui Xie,Penghao Zhou,Tiehan Fan,Zhenheng Yang,Zhijie Chen,Xiang Li,Jia n Yang,Ying Tai

OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation
Text-to-video (T2V) generation has recently garnered significant attention thank s to the large multi-modality model Sora. However, T2V generation still faces tw o important challenges: 1) Lacking a precise open sourced high-quality dataset. The previously popular video datasets, e.g.WebVid-10M and Panda-70M, overly emph asized large scale, resulting in the inclusion of many low-quality videos and short, imprecise captions. Therefore, it is challenging but crucial to collect a precise high-quality dataset while maintaining a scale of millions for T2V gene ration. 2) Ignoring to fully utilize textual information. Recent T2V methods hav e focused on vision transformers, using a simple cross attention module for vide o generation, which falls short of making full use of semantic information from text tokens. To address these issues, we introduce OpenVid-1M, a precise high-qu ality dataset with expressive captions. This open-scenario dataset contains over 1 million text-video pairs, facilitating research on T2V generation. Furthermor e, we curate 433K 1080p videos from OpenVid-1M to create OpenVidHD-0.4M, advanci ng high-definition video generation. Additionally, we propose a novel Multi-moda l Video Diffusion Transformer (MVDiT) capable of mining both structure informati on from visual tokens and semantic information from text tokens. Extensive exper iments and ablation studies verify the superiority of OpenVid-1M over previous d atasets and the effectiveness of our MVDiT.
****
Chenbin Zhang,Zhiqiang Hu,Jiang Chuchu,Wen Chen,JIE XU,Shaoting Zhang
Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation
Drug-target binding affinity prediction is a fundamental task for drug discovery . It has been extensively explored in literature and promising results are repor ted. However, in this paper, we demonstrate that the results may be misleading a nd cannot be well generalized to real practice. The core observation is that the canonical randomized split of a test set in conventional evaluation leaves the test set dominated by samples with high similarity to the training set. The perf ormance of models is severely degraded on samples with lower similarity to the t raining set but the drawback is highly overlooked in current evaluation. As a re sult, the performance can hardly be trusted when the model meets low-similarity samples in real practice. To address this problem, we propose a framework of sim ilarity aware evaluation in which a novel split methodology is proposed to adapt to any desired distribution. This is achieved by a formulation of optimization problems which are approximately and efficiently solved by gradient descent. We perform extensive experiments across five representative methods in four dataset s for two typical target evaluations and compare them with various counterpart m ethods. Results demonstrate that the proposed split methodology can significantl y better fit desired distributions and guide the development of models.
****
Ziyu Zhao,Tao Shen,Didi Zhu,Zexi Li,Jing Su,Xuwu Wang,Fei Wu
Merging LoRAs like Playing LEGO: Pushing the Modularity of LoRA to Extremes Thro ugh Rank-Wise Clustering
Low-Rank Adaptation (LoRA) has emerged as a popular technique for fine-tuning la rge language models (LLMs) to various domains due to its modular design and wide spread availability on platforms like Huggingface. This modularity has sparked i nterest in combining multiple LoRAs to significantly enhance LLM capabilities. H owever, existing methods for LoRA composition primarily focus on task-specific a daptations that require additional training, and current model merging technique s often fail to fully leverage LoRA's modular nature, leading to parameter inter ference and performance degradation.
In this paper, we explore the possibility of disassembling and reassembling mult iple LoRAs at a finer granularity, much like assembling LEGO blocks. We introduc e the concept of Minimal Semantic Units (MSUs), where the parameters correspondi ng to each rank in LoRA function as independent units. These MSUs exhibit proper ties such as permutation invariance and concatenation-summation equivalence, all owing for flexible combinations to form new LoRAs. Building on these insights, w e propose the LoRA-LEGO framework. This framework conducts rank-wise parameter c

lustering by grouping MSUs from different LoRAs into k clusters. The centroid of each cluster serves as a representative MSU, enabling the assembly of a merged LoRA with an adjusted rank of k. Additionally, we apply a dual reweighting strategy to optimize the scale of the merged LoRA. Experiments across various benchmarks demonstrate that our method outperforms existing approaches in LoRA merging.
****
Floris Holstege,Bram Wouters,Noud Van Giersbergen,Cees Diks
Optimizing importance weighting in the presence of sub-population shifts
A distribution shift between the training and test data can severely harm performance of machine learning models. Importance weighting addresses this issue by assigning different weights to data points during training. We argue that existing heuristics for determining the weights are suboptimal, as they neglect the increase of the variance of the estimated model due to the limited sample size of the training data. We interpret the optimal weights in terms of a bias-variance trade-off, and propose a bi-level optimization procedure in which the weights and model parameters are optimized simultaneously. We apply this framework to existing importance weighting techniques for last-layer retraining of deep neural networks in the presence of sub-population shifts and show empirically that optimizing weights significantly improves generalization performance.
****
Cong Chen,Mingyu Liu,Chenchen Jing,Yizhou Zhou,Fengyun Rao,Hao Chen,Bo Zhang,Chunhua Shen
PerturboLLaVA: Reducing Multimodal Hallucinations with Perturbative Visual Training
This paper aims to address the challenge of hallucinations in Multimodal Large Language Models (MLLMs) particularly for dense image captioning tasks. To tackle the challenge, we identify the current lack of a metric that finely measures the caption quality in concept level. We hereby introduce HalFscore, a novel metric built upon the language graph and is designed to evaluate both the accuracy and completeness of dense captions at a
granular level. Additionally, we identify the root cause of hallucination as the model's over-reliance on its language prior. To address this, we propose PerturboLLaVA, which reduces the model's reliance on the language prior by incorporating adversarially perturbed text during training. This method enhances the model's focus on visual inputs, effectively reducing hallucinations and producing accurate, image-grounded descriptions without incurring additional computational overhead. PerturboLLaVA significantly improves the fidelity of generated captions, outperforming existing approaches in handling multimodal hallucinations and achieving improved performance across general multimodal benchmarks.
****
Patrick Schnell,Luca Guastoni,Nils Thuerey
Temporal Difference Learning: Why It Can Be Fast and How It Will Be Faster
Temporal difference (TD) learning represents a fascinating paradox: It is the prime example of a divergent algorithm that has not vanished after its instability was proven. On the contrary, TD continues to thrive in reinforcement learning (RL), suggesting that it provides significant compensatory benefits. Empirical evidence supports this, as many RL tasks require substantial computational resources, and TD delivers a crucial speed advantage that makes these tasks solvable. However, it is limited to cases where the divergence issues are absent or negligible for unknown reasons. So far, the theoretical foundations behind the speed-up are also unclear. In our work, we address these shortcomings of TD by employing techniques for analyzing iterative schemes developed over the past century. Our analysis reveals that TD possesses a mechanism that enables efficient mapping into the smallest eigenspace—an operation previously thought to necessitate costly matrix inversion. Notably, this effect is independent of the conditioning of the problem, making it particularly well-suited for RL tasks characterized by rapidly increasing condition numbers, e.g. through delayed rewards. Our novel theoretical understanding allows us to develop a scalable algorithm that integrates TD’s speed with the reliable convergence of gradient descent (GD). We additionall

y validate these improvements through a rigorous mathematical proof in two dimen
sions, as well as experiments on problems where TD and GD falter, providing valu
able insights into the future of optimization techniques in artificial intellige
nce
****

Shansan Gong,Shivam Agarwal,Yizhe Zhang,Jiacheng Ye,Lin Zheng,Mukai Li,Chenxin A
n,Peilin Zhao,Wei Bi,Jiawei Han,Hao Peng,Lingpeng Kong
Scaling Diffusion Language Models via Adaptation from Autoregressive Models
Diffusion Language Models (DLMs) have emerged as a promising new paradigm for te
xt generative modeling, potentially addressing limitations of autoregressive (AR
) models. However, current DLMs have been studied at a smaller scale compared to
 their AR counterparts and lack fair comparison on language modeling benchmarks.
 Additionally, training diffusion models from scratch at scale remains challengi
ng. Given the prevalence of open-source AR language models, we propose adapting
these models to build text diffusion models. We demonstrate connections between
AR and diffusion modeling objectives and introduce a simple continual pre-traini
ng approach for training diffusion models. Through systematic evaluation on lang
uage modeling, reasoning, and commonsense benchmarks, we show that we can conver
t AR models ranging from 127M to 7B parameters (GPT2 and LLaMA) into diffusion m
odels DiffuGPT and DiffuLLaMA, using less than 200B tokens for training. Our exp
erimental results reveal that these models outperform earlier DLMs and are compe
titive with their AR counterparts. We release a suite of DLMs (127M-355M-7B) cap
able of generating fluent text, performing in-context learning, filling in the m
iddle without prompt re-ordering, and following instructions.
****

Xiangyu Zhao,Zhiwang Zhou,zhangwenlong,Yihao Liu,Xiangyu Chen,Junchao Gong,Hao C
hen,Ben Fei,Shiqi Chen,Wanli Ouyang,Xiao-Ming Wu,LEI BAI
WeatherGFM: Learning a Weather Generalist Foundation Model via In-context Learni
ng
The Earth's weather system involves intricate weather data modalities and divers
e weather understanding tasks, which hold significant value to human life.
Existing data-driven models focus on single weather understanding tasks (e.g., w
eather forecasting).
While these models have achieved promising results, they fail to tackle various
complex tasks within a single and unified model.
Moreover, the paradigm that relies on limited real observations for a single sce
nario hinders the model's performance upper bound.
Inspired by the in-context learning paradigm from visual foundation models and l
arge language models, in this paper, we introduce the first generalist weather g
eneralist foundation model (WeatherGFM) to address weather understanding tasks i
n a unified manner.
Specifically, we first unify the representation and definition for diverse weath
er understanding tasks.
Subsequently, we design weather prompt formats to handle different weather data
modalities, including single, multiple, and temporal modalities.
Finally, we adopt a visual prompting question-answering paradigm for the trainin
g of unified weather understanding tasks.
Extensive experiments indicate that our WeatherGFM can effectively handle up to
12 weather understanding tasks, including weather forecasting, super-resolution,
 weather image translation, and post-processing. Our method also showcases gener
alization ability on unseen tasks. The source code is available at https://githu
b.com/xiangyu-mm/WeatherGFM.
****

Chentao Cao,Zhun Zhong,Zhanke Zhou,Tongliang Liu,Yang Liu,Kun Zhang,Bo Han
Noisy Test-Time Adaptation in Vision-Language Models
Test-time adaptation (TTA) aims to address distribution shifts between source an
d target data by relying solely on target data during testing. In open-world sce
narios, models often encounter noisy samples, i.e., samples outside the in-distr
ibution (ID) label space. Leveraging the zero-shot capability of pre-trained vis
ion-language models (VLMs), this paper introduces Zero-Shot Noisy TTA (ZS-NTTA),

focusing on adapting the model to target data with noisy samples during test-time in a zero-shot manner. In the preliminary study, we reveal that existing TTA methods suffer from a severe performance decline under ZS-NTTA, often lagging behind even the frozen model. We conduct comprehensive experiments to analyze this phenomenon, revealing that the negative impact of unfiltered noisy data outweighs the benefits of clean data during model updating. In addition, as these methods adopt the adapting classifier to implement ID classification and noise detection sub-tasks, the ability of the model in both sub-tasks is largely hampered. Based on this analysis, we propose a novel framework that decouples the classifier and detector, focusing on developing an individual detector while keeping the classifier (including the backbone) frozen. Technically, we introduce the Adaptive Noise Detector (AdaND), which utilizes the frozen model's outputs as pseudo-labels to train a noise detector for detecting noisy samples effectively. To address clean data streams, we further inject Gaussian noise during adaptation, preventing the detector from misclassifying clean samples as noisy. Beyond the ZS-NTTA, AdaND can also improve the zero-shot out-of-distribution (ZS-OOD) detection ability of VLMs. Extensive experiments show that our method outperforms in both ZS-NTTA and ZS-OOD detection. On ImageNet, AdaND achieves a notable improvement of 8.32% in harmonic mean accuracy (Acc_H) for ZS-NTTA and 9.40% in FPR95 for ZS-OOD detection, compared to state-of-the-art methods. Importantly, AdaND is computationally efficient and comparable to the model-frozen method. The code is publicly available at: https://github.com/tmlr-group/ZS-NTTA.
****
Tarun Suresh,Revanth Gangi Reddy,Yifei Xu,Zach Nussbaum,Andriy Mulyar,Brandon Duderstadt,Heng Ji
CoRNStack: High-Quality Contrastive Data for Better Code Retrieval and Reranking
Effective code retrieval plays a crucial role in advancing code generation, bug fixing, and software maintenance, particularly as software systems increase in complexity. While current code embedding models have demonstrated promise in retrieving code snippets for small-scale, well-defined tasks, they often underperform in more demanding real-world applications such as bug localization within GitHub repositories. We hypothesize that a key issue is their reliance on noisy and inconsistent datasets for training, which impedes their ability to generalize to more complex retrieval scenarios. To address these limitations, we introduce CoRNStack, a large-scale, high-quality contrastive training dataset for code that spans multiple programming languages. This dataset is curated using consistency filtering to eliminate noisy positives and is further enriched with mined hard negatives, thereby facilitating more effective learning. We demonstrate that contrastive training of embedding models using CoRNStack leads to state-of-the-art performance across a variety of code retrieval tasks. Furthermore, the dataset can be leveraged for training code reranking models, a largely underexplored area compared to text reranking. Our finetuned code reranking model significantly improves the ranking quality over the retrieved results. Finally, by employing our code retriever and reranker together, we demonstrate significant improvements in function localization for GitHub issues, an important
component of real-world software development.
****
Lu Dai,Yijie Xu,Jinhui Ye,Hao Liu,Hui Xiong
SePer: Measure Retrieval Utility Through The Lens Of Semantic Perplexity Reduction
Large Language Models (LLMs) have demonstrated improved generation performance by incorporating externally retrieved knowledge, a process known as retrieval-augmented generation (RAG). Despite the potential of this approach, existing studies evaluate RAG effectiveness by 1) assessing retrieval and generation components jointly, which obscures retrieval's distinct contribution, or 2) examining retrievers using traditional metrics such as NDCG, which creates a gap in understanding retrieval's true utility in the overall generation process. To address the above limitations, in this work, we introduce an automatic evaluation method that measures retrieval quality through the lens of information gain within the RAG framework. Specifically, we propose Semantic Perplexity (SePer), a metric that c

aptures the LLM's internal belief about the correctness of the retrieved informa
tion. We quantify the utility of retrieval by the extent to which it reduces sem
antic perplexity post-retrieval. Extensive experiments demonstrate that SePer no
t only aligns closely with human preferences but also offers a more precise and
efficient evaluation of retrieval utility across diverse RAG scenarios.
****

Yu Liu,Baoxiong Jia,Ruijie Lu,Junfeng Ni,Song-Chun Zhu,Siyuan Huang
Building Interactable Replicas of Complex Articulated Objects via Gaussian Splat
ting

Building interactable replicas of articulated objects is a key challenge in comp
uter vision. Existing methods often fail to effectively integrate information ac
ross different object states, limiting the accuracy of part-mesh reconstruction
and part dynamics modeling, particularly for complex multi-part articulated obje
cts. We introduce ArtGS, a novel approach that leverages 3D Gaussians as a flexi
ble and efficient representation to address these issues. Our method incorporate
s canonical Gaussians with coarse-to-fine initialization and updates for alignin
g articulated part information across different object states, and employs a ski
nning-inspired part dynamics modeling module to improve both part-mesh reconstru
ction and articulation learning. Extensive experiments on both synthetic and rea
l-world datasets, including a new benchmark for complex multi-part objects, demo
nstrate that ArtGS achieves state-of-the-art performance in joint parameter esti
mation and part mesh reconstruction. Our approach significantly improves reconst
ruction quality and efficiency, especially for multi-part articulated objects. A
dditionally, we provide comprehensive analyses of our design choices, validating
 the effectiveness of each component to highlight potential areas for future imp
rovement.
****

Rushang Karia,Daniel Richard Bramblett,Daksh Dobhal,Siddharth Srivastava
AutoEval: Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Task
s

This paper presents AutoEval, a novel benchmark for scaling Large Language Model
 (LLM) assessment in formal tasks with clear notions of correctness, such as tru
th maintenance in translation and logical reasoning. AutoEval is the first bench
marking paradigm that offers several key advantages necessary for scaling object
ive evaluation of LLMs without human labeling: (a) ability to evaluate LLMs of i
ncreasing sophistication by auto-generating tasks at different levels of difficu
lty; (b) auto-generation of ground truth that eliminates dependence on expensive
 and time-consuming human annotation; (c) the use of automatically generated, ra
ndomized datasets that mitigate the ability of successive LLMs to overfit to sta
tic datasets used in many contemporary benchmarks. Empirical analysis shows that
 an LLM's performance on AutoEval is highly indicative of its performance on a d
iverse array of other benchmarks focusing on translation and reasoning tasks, ma
king it a valuable autonomous evaluation paradigm in settings where hand-curated
 datasets can be hard to obtain and/or update.
****

Jiajie Li,Brian R Quaranto,Chenhui Xu,Ishan Mishra,Ruiyang Qin,Dancheng Liu,Pete
r C W Kim,Jinjun Xiong
Recognize Any Surgical Object: Unleashing the Power of Weakly-Supervised Data

We present RASO, a foundation model designed to Recognize Any Surgical Object, o
ffering robust open-set recognition capabilities across a broad range of surgica
l procedures and object classes, in both surgical images and videos. RASO levera
ges a novel weakly-supervised learning framework that generates tag-image-text p
airs automatically from large-scale unannotated surgical lecture videos, signifi
cantly reducing the need for manual annotations. Our scalable data generation pi
peline gathers 2,200 surgical procedures and produces 3.6 million tag annotation
s across 2,066 unique surgical tags. Our experiments show that RASO achieves imp
rovements of 2.9 mAP, 4.5 mAP, 10.6 mAP, and 7.2 mAP on four standard surgical b
enchmarks respectively in zero-shot settings, and surpasses state-of-the-art mod
els in supervised surgical action recognition tasks. We will open-source our cod
e, model, and dataset to facilitate further research.

****
Oussama Boussif,Lena Nehale Ezzine,Joseph D Viviano,Micha■ Koziarski,Moksh Jain,
Nikolay Malkin,Emmanuel Bengio,Rim Assouel,Yoshua Bengio
Action abstractions for amortized sampling
As trajectories sampled by policies used by reinforcement learning (RL) and gene
rative flow networks (GFlowNets) grow longer, credit assignment and exploration
become more challenging, and the long planning horizon hinders mode discovery an
d generalization.
The challenge is particularly pronounced in entropy-seeking RL methods, such as
generative flow networks, where the agent must learn to sample from a structured
distribution and discover multiple high-reward states, each of which take many
steps to reach.
To tackle this challenge, we propose an approach to incorporate the discovery of
action abstractions, or high-level actions, into the policy optimization proces
s.
Our approach involves iteratively extracting action subsequences commonly used a
cross many high-reward trajectories and `chunking' them into a single action tha
t is added to the action space.
In empirical evaluation on synthetic and real-world environments, our approach d
emonstrates improved sample efficiency performance in discovering diverse high-r
eward objects, especially on harder exploration problems.
We also observe that the abstracted high-order actions are potentially interpret
able, capturing the latent structure of the reward landscape of the action space
.
This work provides a cognitively motivated approach to action abstraction in RL
and is the first demonstration of hierarchical planning in amortized sequential
sampling.
****
Siyu Chen,Beining Wu,Miao Lu,Zhuoran Yang,Tianhao Wang
Can Neural Networks Achieve Optimal Computational-statistical Tradeoff? An Analy
sis on Single-Index Model
In this work, we tackle the following question: Can neural networks trained with
gradient-based methods achieve the optimal statistical-computational tradeoff i
n learning Gaussian single-index models?
Prior research has shown that any polynomial-time algorithm under the statistica
l query (SQ) framework requires $\Omega(d^{s^\star/2}\lor d)$ samples, where $s^
\star$ is the generative exponent representing the intrinsic difficulty of learn
ing the underlying model.
However, it remains unknown whether neural networks can achieve this sample comp
lexity.
Inspired by prior techniques such as label transformation and landscape smoothin
g for learning single-index models, we propose a unified gradient-based algorith
m for training a two-layer neural network in polynomial time.
Our method is adaptable to a variety of loss and activation functions, covering
a broad class of existing approaches.
We show that our algorithm learns a feature representation that strongly aligns
with the unknown signal $\theta^\star$, with sample complexity $\tilde O (d^{s^\
star/2} \lor d)$, matching the SQ lower bound up to a polylogarithmic factor for
all generative exponents $s^\star\geq 1$.
Furthermore, we extend our approach to the setting where $\theta^\star$ is k-s
parse for $k = o(\sqrt{d})$ by introducing a novel weight perturbation technique
that leverages the sparsity structure.
We derive a corresponding SQ lower bound
of order $\tilde\Omega(k^{s^\star})$, matched by our method up to a polylogarith
mic factor.
Our framework, especially the weight perturbation technique, is of independent i
nterest, and suggests potential gradient-based solutions to other problems such
as sparse tensor PCA.
****
Alonso Granados,Reza Ebrahimi,Jason Pacheco

Risk-Sensitive Variational Actor-Critic: A Model-Based Approach
Risk-sensitive reinforcement learning (RL) with an entropic risk measure typical
ly requires knowledge of the transition kernel or performs unstable updates w.r.
t. exponential Bellman equations. As a consequence, algorithms that optimize thi
s objective have been restricted to tabular or low-dimensional continuous enviro
nments. In this work we leverage the connection between the entropic risk measur
e and the RL-as-inference framework to develop a risk-sensitive variational acto
r-critic algorithm (rsVAC). Our work extends the variational framework to incorp
orate stochastic rewards and proposes a variational model-based actor-critic app
roach that modulates policy risk via a risk parameter. We consider, both, the r
isk-seeking and risk-averse regimes and present rsVAC learning variants for each
 setting. Our experiments demonstrate that this approach produces risk-sensitiv
e policies and yields improvements in both tabular and risk-aware variants of co
mplex continuous control tasks in MuJoCo.
****
Rui Dai,Sile Hu,Xu Shen,Yonggang Zhang,Xinmei Tian,Jieping Ye
Leveraging Submodule Linearity Enhances Task Arithmetic Performance in LLMs
Task arithmetic is a straightforward yet highly effective strategy for model mer
ging, enabling the resultant model to exhibit multi-task capabilities. Recent re
search indicates that models demonstrating linearity enhance the performance of
task arithmetic. In contrast to existing methods that rely on the global lineari
zation of the model, we argue that this linearity already exists within the mode
l's submodules. In particular, we present a statistical analysis and show that s
ubmodules (e.g., layers, self-attentions, and MLPs) exhibit significantly higher
 linearity than the overall model. Based on these findings, we propose an innova
tive model merging strategy that independently merges these submodules. Especial
ly, we derive a closed-form solution for optimal merging weights grounded in the
 linear properties of these submodules. Experimental results demonstrate that ou
r method consistently outperforms the standard task arithmetic approach and othe
r established baselines across different model scales and various tasks. This re
sult highlights the benefits of leveraging the linearity of submodules and provi
des a new perspective for exploring solutions for effective and practical multi-
task model merging.
****
Siddharth Swaroop,Mohammad Emtiyaz Khan,Finale Doshi-Velez
Connecting Federated ADMM to Bayes
We provide new connections between two distinct federated learning approaches ba
sed on (i) ADMM and (ii) Variational Bayes (VB), and propose new variants by com
bining their complementary strengths. Specifically, we show that the dual variab
les in ADMM naturally emerge through the "site" parameters used in VB with isotr
opic Gaussian covariances. Using this, we derive two versions of ADMM from VB th
at use flexible covariances and functional regularisation, respectively. Through
 numerical experiments, we validate the improvements obtained in performance. Th
e work shows connection between two fields that are believed to be fundamentally
 different and combines them to improve federated learning.
****
Martin Kuo,Jingyang Zhang,Jianyi Zhang,Minxue Tang,Louis DiValentin,Aolin Ding,J
ingwei Sun,William Chen,Amin Hass,Tianlong Chen,Yiran Chen,Hai Li
Proactive Privacy Amnesia for Large Language Models: Safeguarding PII with Negli
gible Impact on Model Utility
With the rise of large language models (LLMs), increasing research has recognize
d
their risk of leaking personally identifiable information (PII) under malicious
attacks. Although efforts have been made to protect PII in LLMs, existing method
s
struggle to balance privacy protection with maintaining model utility. In this p
aper,
inspired by studies of amnesia in cognitive science, we propose a novel approach
,
Proactive Privacy Amnesia (PPA), to safeguard PII in LLMs while preserving their

utility. This mechanism works by actively identifying and forgetting key memorie
s
most closely associated with PII in sequences, followed by a memory implanting
using suitable substitute memories to maintain the LLM’s functionality. We condu
ct
evaluations across multiple models to protect common PII, such as phone numbers
and physical addresses, against prevalent PII-targeted attacks, demonstrating th
e
superiority of our method compared with other existing defensive techniques. The
results show that our PPA method completely eliminates the risk of phone number
exposure by 100% and significantly reduces the risk of physical address exposure
by 9.8% – 87.6%, all while maintaining comparable model utility performance.
****
Kehan Wen,Yutong Hu,Yao Mu,Lei Ke
M^3PC: Test-time Model Predictive Control using Pretrained Masked Trajectory Mod
el
Recent work in Offline Reinforcement Learning (RL) has shown that a unified tra
nsformer trained under a masked auto-encoding objective can effectively capture
the relationships between different modalities (e.g., states, actions, rewards)
within given trajectory datasets. However, this information has not been fully e
xploited during the inference phase, where the agent needs to generate an optima
l policy instead of just reconstructing masked components from unmasked. Given t
hat a pretrained trajectory model can act as both a Policy Model and a World Mod
el with appropriate mask patterns, we propose using Model Predictive Control (MP
C) at test time to leverage the model's own predictive capacity to guide its act
ion selection. Empirical results on D4RL and RoboMimic show that our inference-p
hase MPC significantly improves the decision-making performance of a pretrained
trajectory model without any additional parameter training. Furthermore, our fra
mework can be adapted to Offline to Online (O2O) RL and Goal Reaching RL, result
ing in more substantial performance gains when an additional online interaction
budget is given, and better generalization capabilities when different task targ
ets are specified. Code is available: \href{https://github.com/wkh923/m3pc}{\tex
ttt{https://github.com/wkh923/m3pc}}.
****
Binghui Li,Yuanzhi Li
Adversarial Training Can Provably Improve Robustness: Theoretical Analysis of Fe
ature Learning Process Under Structured Data
Adversarial training is a widely-applied approach to training deep neural networ
ks to be robust against adversarial perturbation. However, although adversarial
training has achieved empirical success in practice, it still remains unclear wh
y adversarial examples exist and how adversarial training methods improve model
robustness. In this paper, we provide a theoretical understanding of adversarial
 examples and adversarial training algorithms from the perspective of feature le
arning theory. Specifically, we focus on a multiple classification setting, wher
e the structured data can be composed of two types of features: the robust featu
res, which are resistant to perturbation but sparse, and the non-robust features
, which are susceptible to perturbation but dense. We train a two-layer smoothed
 ReLU convolutional neural network to learn our structured data. First, we prove
 that by using standard training (gradient descent over the empirical risk), the
 network learner primarily learns the non-robust feature rather than the robust
feature, which thereby leads to the adversarial examples that are generated by p
erturbations aligned with negative non-robust feature directions. Then, we consi
der the gradient-based adversarial training algorithm, which runs gradient ascen
t to find adversarial examples and runs gradient descent over the empirical risk
 at adversarial examples to update models. We show that the adversarial training
 method can provably strengthen the robust feature learning and suppress the non
-robust feature learning to improve the network robustness. Finally, we also emp
irically validate our theoretical findings with experiments on real-image datase
ts, including MNIST, CIFAR10 and SVHN.
****

Hyesu Lim,Jinho Choi,Jaegul Choo,Steffen Schneider
Sparse autoencoders reveal selective remapping of visual concepts during adaptat ion
Adapting foundation models for specific purposes has become a standard approach to build machine learning systems for downstream applications. Yet, it is an ope n question which mechanisms take place during adaptation. Here we develop a new Sparse Autoencoder (SAE) for the CLIP vision transformer, named PatchSAE, to ext ract interpretable concepts at granular levels (e.g., shape, color, or semantics of an object) and their patch-wise spatial attributions. We explore how these c oncepts influence the model output in downstream image classification tasks and investigate how recent state-of-the-art prompt-based adaptation techniques chang e the association of model inputs to these concepts. While activations of concep ts slightly change between adapted and non-adapted models, we find that the majo rity of gains on common adaptation tasks can be explained with the existing conc epts already present in the non-adapted foundation model. This work provides a c oncrete framework to train and use SAEs for Vision Transformers and provides ins ights into explaining adaptation mechanisms.
****

Ximing Lu,Melanie Sclar,Skyler Hallinan,Niloofar Mireshghallah,Jiacheng Liu,Seun gju Han,Allyson Ettinger,Liwei Jiang,Khyathi Chandu,Nouha Dziri,Yejin Choi
AI as Humanity’s Salieri: Quantifying Linguistic Creativity of Language Models v ia Systematic Attribution of Machine Text against Web Text
Creativity has long been considered one of the most difficult aspect of human in telligence for AI to mimic. However, the rise of Large Language Models (LLMs), l ike ChatGPT, has raised questions about whether AI can match or even surpass hum an creativity. We present CREATIVITY INDEX as the first step to quantify the lin guistic creativity of a text by reconstructing it from existing text snippets on the web. CREATIVITY INDEX is motivated by the hypothesis that the seemingly rem arkable creativity of LLMs may be attributable in large part to the creativity o f human-written texts on the web. To compute CREATIVITY INDEX efficiently, we in troduce DJ SEARCH, a novel dynamic programming algorithm that can search verbati m and near-verbatim matches of text snippets from a given document against the w eb. Experiments reveal that the CREATIVITY INDEX of professional human authors i s on average 66.2% higher than that of LLMs, and that alignment reduces the CREA TIVITY INDEX of LLMs by an average of 30.1%. In addition, we find that distingui shed authors like Hemingway exhibit measurably higher CREATIVITY INDEX compared to other human writers. Finally, we demonstrate that CREATIVITY INDEX can be use d as a surprisingly effective criterion for zero-shot machine text detection, su rpassing the strongest existing zero-shot system, DetectGPT, by a significant ma rgin of 30.2%, and even outperforming the strongest supervised system, GhostBust er, in five out of six domains.
****

Christopher Rawles,Sarah Clinckemaillie,Yifan Chang,Jonathan Waltz,Gabrielle Lau ,Marybeth Fair,Alice Li,William E Bishop,Wei Li,Folawiyo Campbell-Ajala,Daniel K enji Toyama,Robert James Berry,Divya Tyamagundlu,Timothy P Lillicrap,Oriana Riva
AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents
Autonomous agents that execute human tasks by controlling computers can enhance human productivity and application accessibility. However, progress in this fiel d will be driven by realistic and reproducible benchmarks. We present AndroidWor ld, a fully functional Android environment that provides reward signals for 116 programmatic tasks across 20 real-world Android apps. Unlike existing interactiv e environments, which provide a static test set, AndroidWorld dynamically constr ucts tasks that are parameterized and expressed in natural language in unlimited ways, thus enabling testing on a much larger and more realistic suite of tasks. To ensure reproducibility, each task includes dedicated initialization, success -checking, and tear-down logic, which modifies and inspects the device’s system state.

We experiment with baseline agents to test AndroidWorld and provide initial resu lts on the benchmark. Our best agent can complete 30.6% of AndroidWorld's tasks,

leaving ample room for future work. Furthermore, we adapt a popular desktop web agent to work on Android, which we find to be less effective on mobile, suggesting future research is needed to achieve universal, cross-platform agents. Finally, we also conduct a robustness analysis, showing that task variations can significantly affect agent performance, demonstrating that without such testing, agent performance metrics may not fully reflect practical challenges. AndroidWorld and the experiments in this paper are available at https://github.com/google-research/android_world.
****
Arjun Krishna,Edward S. Hu,Dinesh Jayaraman
The Value of Sensory Information to a Robot
A decision-making agent, such as a robot, must observe and react to any new task-relevant information that becomes available from its environment. We seek to study a fundamental scientific question: what value does sensory information hold to an agent at various moments in time during the execution of a task? Towards this, we empirically study agents of varying architectures, generated with varying policy synthesis approaches (imitation, RL, model-based control), on diverse robotics tasks. For each robotic agent, we characterize its regret in terms of performance degradation when state observations are withheld from it at various task states for varying lengths of time. We find that sensory information is surprisingly rarely task-critical in many commonly studied task setups. Task characteristics such as stochastic dynamics largely dictate the value of sensory information for a well-trained robot; policy architectures such as planning vs. reactive control generate more nuanced second-order effects. Further, sensing efficiency is curiously correlated with task proficiency: in particular, fully trained high-performing agents are more robust to sensor loss than novice agents early in their training. Overall, our findings characterize the tradeoffs between sensory information and task performance in practical sequential decision making tasks, and pave the way towards the design of more resource-efficient decision-making agents.
****
Daogao Liu,Kunal Talwar
Adaptive Batch Size for Privately Finding Second-Order Stationary Points
There is a gap between finding a first-order stationary point (FOSP) and a second-order stationary point (SOSP) under differential privacy constraints, and it remains unclear whether privately finding an SOSP is more challenging than finding an FOSP. Specifically, Ganesh et al. (2023) claimed that an α-SOSP can be found with $\alpha=\Tilde{O}(\frac{1}{n^{1/3}}+(\frac{\sqrt{d}}{n\epsilon})^{3/7})$, where n is the dataset size, d is the dimension, and ϵ is the differential privacy parameter.
However, a recent analysis revealed an issue in their saddle point escape procedure, leading to weaker guarantees.
Building on the SpiderBoost algorithm framework, we propose a new approach that uses adaptive batch sizes and incorporates the binary tree mechanism.
Our method not only corrects this issue but also improves the results for privately finding an SOSP, achieving $\alpha=\Tilde{O}(\frac{1}{n^{1/3}}+(\frac{\sqrt{d}}{n\epsilon})^{1/2})$.
This improved bound matches the state-of-the-art for finding a FOSP, suggesting that privately finding an SOSP may be achievable at no additional cost.
****
Lesi Chen,Chengchang Liu,Jingzhao Zhang
Second-Order Min-Max Optimization with Lazy Hessians
This paper studies second-order methods for convex-concave minimax optimization.

Monteiro & Svaiter (2012) proposed a method to solve the problem with an optimal iteration complexity of
$\mathcal{O}(\epsilon^{-3/2})$ to find an ϵ-saddle point. However, it is unclear whether the
computational complexity, $\mathcal{O}((N+ d^2) d \epsilon^{-2/3})$, can be improved. In the above, we follow Doikov et al. (2023) and assume the complexity of

obtaining a first-order oracle as N and the complexity of obtaining a second-order oracle as dN.
In this paper, we show that the computation cost can be reduced by reusing Hessian across iterations. Our methods take the overall computational complexity of $\tilde{\mathcal{O}}((N+d^2)(d+ d^{2/3}\epsilon^{-2/3}))$, which improves those of previous methods by a factor of $d^{1/3}$.
Furthermore, we generalize our method to strongly-convex-strongly-concave minimax problems and establish the complexity of $\tilde{\mathcal{O}}((N+d^2) (d + d^{2/3} \kappa^{2/3}))$ when the condition number of the problem is κ, enjoying a similar speedup upon the state-of-the-art method.
Numerical experiments on both real and synthetic datasets also verify the efficiency of our method.
****
Neehar Kondapaneni,Oisin Mac Aodha,Pietro Perona
Representational Similarity via Interpretable Visual Concepts
How do two deep neural networks differ in how they arrive at a decision? Measuring the similarity of deep networks has been a long-standing open question. Most existing methods provide a single number to measure the similarity of two networks at a given layer, but give no insight into what makes them similar or dissimilar. We introduce an interpretable representational similarity method (RSVC) to compare two networks. We use RSVC to discover shared and unique visual concepts between two models. We show that some aspects of model differences can be attributed to unique concepts discovered by one model that are not well represented in the other. Finally, we conduct extensive evaluation across different vision model architectures and training protocols to demonstrate its effectiveness.
****
Wei Xiao,Tsun-Hsuan Wang,Chuang Gan,Ramin Hasani,Mathias Lechner,Daniela Rus
SafeDiffuser: Safe Planning with Diffusion Probabilistic Models
Diffusion models have shown promise in data-driven planning. While these planners are commonly employed in applications where decisions are critical, they still lack established safety guarantees. In this paper, we address this limitation by introducing SafeDiffuser, a method to equip diffusion models with safety guarantees via control barrier functions. The key idea of our approach is to embed finite-time diffusion invariance, i.e., a form of specification consisting of safety constraints, into the denoising diffusion procedure. This way we enable data generation under safety constraints. We show that SafeDiffusers maintain the generative performance of diffusion models while also providing robustness in safe data generation. We evaluate our method on a series of tasks, including maze path generation, legged robot locomotion, and 3D space manipulation, and demonstrate the advantages of robustness over vanilla diffusion models.
****
Zizhuo Zhang,Lijun Wu,Kaiyuan Gao,Jiangchao Yao,Tao Qin,Bo Han
Fast and Accurate Blind Flexible Docking
Molecular docking that predicts the bound structures of small molecules (ligands) to their protein targets, plays a vital role in drug discovery. However, existing docking methods often face limitations: they either overlook crucial structural changes by assuming protein rigidity or suffer from low computational efficiency due to their reliance on generative models for structure sampling. To address these challenges, we propose FABFlex, a fast and accurate regression-based multi-task learning model designed for realistic blind flexible docking scenarios, where proteins exhibit flexibility and binding pocket sites are unknown (blind). Specifically, FABFlex's architecture comprises three specialized modules working in concert: (1) A pocket prediction module that identifies potential binding sites, addressing the challenges inherent in blind docking scenarios. (2) A ligand docking module that predicts the bound (holo) structures of ligands from their unbound (apo) states. (3) A pocket docking module that forecasts the holo structures of protein pockets from their apo conformations. Notably, FABFlex incorporates an iterative update mechanism that serves as a conduit between the ligand and pocket docking modules, enabling continuous structural refinements. This approach effectively integrates the three subtasks of blind flexible docking—pocket

identification, ligand conformation prediction, and protein flexibility modelin
g—into a unified, coherent framework. Extensive experiments on public benchmark
datasets demonstrate that FABFlex not only achieves superior effectiveness in pr
edicting accurate binding modes but also exhibits a significant speed advantage
(208$\times$) compared to existing state-of-the-art methods. Our code is release
d at~\url{https://github.com/tmlr-group/FABFlex}.
****
Nikolaos Dimitriadis,Pascal Frossard,François Fleuret
Pareto Low-Rank Adapters: Efficient Multi-Task Learning with Preferences
Multi-task trade-offs in machine learning can be addressed via Pareto Front Lear
ning (PFL) methods that parameterize the Pareto Front (PF) with a single model.
PFL permits to select the desired operational point during inference, contrary t
o traditional Multi-Task Learning (MTL) that optimizes for a single trade-off de
cided prior to training. However, recent PFL methodologies suffer from limited s
calability, slow convergence, and excessive memory requirements, while exhibitin
g inconsistent mappings from preference to objective space. We introduce PaLoRA,
a novel parameter-efficient method that addresses these limitations in two ways
. First, we augment any neural network architecture with task-specific low-rank
adapters and continuously parameterize the Pareto Front in their convex hull. Ou
r approach steers the original model and the adapters towards learning general a
nd task-specific features, respectively. Second, we propose a deterministic samp
ling schedule of preference vectors that reinforces this division of labor, enab
ling faster convergence and strengthening the validity of the mapping from prefe
rence to objective space throughout training. Our experiments show that PaLoRA o
utperforms state-of-the-art MTL and PFL baselines across various datasets, scale
s to large networks, reducing the memory overhead $23.8-31.7$ times compared wit
h competing PFL baselines in scene understanding benchmarks.
****
Zexu Sun,Yiju Guo,Yankai Lin,Xu Chen,Qi Qi,Xing Tang,xiuqiang He,Ji-Rong Wen
Uncertainty and Influence aware Reward Model Refinement for Reinforcement Learni
ng from Human Feedback
Reinforcement Learning from Human Feedback (RLHF) has emerged as a standard and
effective approach for training large language models (LLMs) with human preferen
ces. In this framework, a learned reward model approximates human preferences an
d guides policy optimization, making it crucial to develop an accurate reward mo
del. However, without the ``true'' reward function, challenges arise when the re
ward model is an imperfect proxy for human preference. Since the policy optimiza
tion continuously shifts the human preference training dataset's distribution. T
he fixed reward model suffers from this problem of off-distribution, especially
the on policy methods. While collecting new preference data can mitigate this is
sue, it is costly and challenging to optimize. Thus, reusing the policy interact
ion samples becomes a possible way to further refine the reward model. To tackle
these challenges, we introduce a novel method \textbf{U}ncertainty-\textbf{G}ra
dient based \textbf{D}ata \textbf{A}ugmentation (\textbf{UGDA} for short) to enh
ance reward modeling by leveraging policy samples to maintain on-distribution pe
rformance. Specifically, UGDA selects interaction samples based on the uncertain
ty of the reward ensembles and the gradient based influence of policy optimizati
on. After the reward relabeling of selected samples, we use supervised learning
to refine the reward ensembles, then get the retrained policy. Extensive experim
ents demonstrate that by leveraging UGDA to select a few samples without the cos
tly human preference data collection, we can improve the ability of the policy a
nd surpass the state-of-the-art methods.
****
Huiyang Yi,Yanyan He,Duxin Chen,Mingyu Kang,He Wang,Wenwu Yu
THE ROBUSTNESS OF DIFFERENTIABLE CAUSAL DISCOVERY IN MISSPECIFIED SCENARIOS
Causal discovery aims to learn causal relationships between variables from targe
ted data, making it a fundamental task in machine learning. However, causal disc
overy algorithms often rely on unverifiable causal assumptions, which are usuall
y difficult to satisfy in real-world data, thereby limiting the broad applicatio
n of causal discovery in practical scenarios. Inspired by these considerations,

this work extensively benchmarks the empirical performance of various mainstream causal discovery algorithms, which assume i.i.d. data, under eight model assumption violations. Our experimental results show that differentiable causal discovery methods exhibit robustness under the metrics of Structural Hamming Distance and Structural Intervention Distance of the inferred graphs in commonly used challenging scenarios, except for scale variation. We also provide the theoretical explanations for the performance of differentiable causal discovery methods. Finally, our work aims to comprehensively benchmark the performance of recent differentiable causal discovery methods under model assumption violations, and provide the standard for reasonable evaluation of causal discovery, as well as to further promote its application in real-world scenarios.
****

Anna Susmelj,Mael Macuglia,Natasa Tagasovska,Reto Sutter,Sebastiano Caprara,Jean-Philippe Thiran,Ender Konukoglu
Uncertainty modeling for fine-tuned implicit functions
Implicit functions such as Neural Radiance Fields (NeRFs), occupancy networks, and signed distance functions (SDFs) have become pivotal in computer vision for reconstructing detailed object shapes from sparse views. Achieving optimal performance with these models can be challenging due to the extreme sparsity of inputs and distribution shifts induced by data corruptions. To this end, large, noise-free synthetic datasets can serve as shape priors to help models fill in gaps, but the resulting reconstructions must be approached with caution. Uncertainty estimation is crucial for assessing the quality of these reconstructions, particularly in identifying areas where the model is uncertain about the parts it has inferred from the prior. In this paper, we introduce Dropsembles, a novel method for uncertainty estimation in tuned implicit functions. We demonstrate the efficacy of our approach through a series of experiments, starting with toy examples and progressing to a real-world scenario. Specifically, we train a Convolutional Occupancy Network on synthetic anatomical data and test it on low-resolution MRI segmentations of the lumbar spine. Our results show that Dropsembles achieve the accuracy and calibration levels of deep ensembles but with significantly less computational cost.
****

Samir Yitzhak Gadre,Georgios Smyrnis,Vaishaal Shankar,Suchin Gururangan,Mitchell Wortsman,Rulin Shao,Jean Mercat,Alex Fang,Jeffrey Li,Sedrick Keh,Rui Xin,Marianna Nezhurina,Igor Vasiljevic,Luca Soldaini,Jenia Jitsev,Alex Dimakis,Gabriel Ilharco,Pang Wei Koh,Shuran Song,Thomas Kollar,Yair Carmon,Achal Dave,Reinhard Heckel,Niklas Muennighoff,Ludwig Schmidt
Language models scale reliably with over-training and on downstream tasks
Scaling laws are useful guides for derisking expensive training runs, as they predict performance of large models using cheaper, small-scale experiments. However, there remain gaps between current scaling studies and how language models are ultimately trained and evaluated. For instance, scaling is usually studied in the compute-optimal training regime (i.e., "Chinchilla optimal" regime). In contrast, models are often over-trained to reduce inference costs. Moreover, scaling laws mostly predict loss on next-token prediction, but models are usually compared on downstream task performance. To address both shortcomings, we create a testbed of 104 models with 0.011B to 6.9B parameters trained with various numbers of tokens on three data distributions. First, we fit scaling laws that extrapolate in both the amount of over-training and the number of model parameters. This enables us to predict the validation loss of a 1.4B parameter, 900B token run (i.e., 32$\times$ over-trained) and a 6.9B parameter, 138B token run (i.e., a compute-optimal run)––each from experiments that take 300$\times$ less compute. Second, we relate the perplexity of a language model to its downstream task performance by proposing a power law. We use this law to predict top-1 error averaged over downstream tasks for the two aforementioned models, using experiments that take 20$\times$ less compute.
****

Sunghyeon Woo,Sol Namkung,Sunwoo Lee,Inho Jeong,Beomseok Kim,Dongsuk Jeon
PaCA: Partial Connection Adaptation for Efficient Fine-Tuning

Prior parameter-efficient fine-tuning (PEFT) algorithms reduce memory usage and computational costs of fine-tuning large neural network models by training only a few additional adapter parameters, rather than the entire model. However, the reduction in computational costs due to PEFT does not necessarily translate to a reduction in training time; although the computational costs of the adapter layers are much smaller than the pretrained layers, it is well known that those two types of layers are processed sequentially on GPUs, resulting in significant latency overhead. LoRA and its variants avoid this latency overhead by merging the low-rank adapter matrices with the pretrained weights during inference. However, those layers cannot be merged during training since the pretrained weights must remain frozen while the low-rank adapter matrices are updated continuously over the course of training. Furthermore, LoRA and its variants do not reduce activation memory, as the first low-rank adapter matrix still requires the input activations to the pretrained weights to compute weight gradients. To mitigate this issue, we propose **Pa**rtial **C**onnection **A**daptation (**PaCA**), which fine-tunes randomly selected partial connections within the pretrained weights instead of introducing adapter layers in the model. PaCA not only enhances training speed by eliminating the time overhead due to the sequential processing of the adapter and pretrained layers but also reduces activation memory since only partial activations, rather than full activations, need to be stored for gradient computation. Compared to LoRA, PaCA reduces training time by 22% and total memory usage by 16%, while maintaining comparable accuracy across various fine-tuning scenarios, such as fine-tuning on the MMLU dataset and instruction tuning on the Oasst1 dataset. PaCA can also be combined with quantization, enabling the fine-tuning of large models such as LLaMA3.1-70B. In addition, PaCA enables training with 23% longer sequence and improves throughput by 16\% on both NVIDIA A100 GPU and INTEL Gaudi2 HPU compared to LoRA. The code is available at https://github.com/WooSunghyeon/paca.
****

Denis Blessing,Xiaogang Jia,Gerhard Neumann

End-to-end Learning of Gaussian Mixture Priors for Diffusion Sampler

Diffusion models optimized via variational inference (VI) have emerged as a promising tool for generating samples from unnormalized target densities. These models create samples by simulating a stochastic differential equation, starting from a simple, tractable prior, typically a Gaussian distribution. However, when the support of this prior differs greatly from that of the target distribution, diffusion models often struggle to explore effectively or suffer from large discretization errors. Moreover, learning the prior distribution can lead to mode-collapse, exacerbated by the mode-seeking nature of reverse Kullback-Leibler divergence commonly used in VI.

To address these challenges, we propose end-to-end learnable Gaussian mixture priors (GMPs). GMPs offer improved control over exploration, adaptability to target support, and increased expressiveness to counteract mode collapse. We further leverage the structure of mixture models by proposing a strategy to iteratively refine the model through the addition of mixture components during training. Our experimental results demonstrate significant performance improvements across a diverse range of real-world and synthetic benchmark problems when using GMPs without requiring additional target evaluations.
****

Oleh Kolner,Thomas Ortner,Stanis■aw Wo■niak,Angeliki Pantazi

Mind the GAP: Glimpse-based Active Perception improves generalization and sample efficiency of visual reasoning

Human capabilities in understanding visual relations are far superior to those of AI systems, especially for previously unseen objects. For example, while AI systems struggle to determine whether two such objects are visually the same or different, humans can do so with ease. Active vision theories postulate that the learning of visual relations is grounded in actions that we take to fixate objects and their parts by moving our eyes. In particular, the low-dimensional spatial information about the corresponding eye movements is hypothesized to facilitate the representation of relations between different image parts. Inspired by thes

e theories, we develop a system equipped with a novel Glimpse-based Active Perce
ption (GAP) that sequentially glimpses at the most salient regions of the input
image and processes them at high resolution. Importantly, our system leverages t
he locations stemming from the glimpsing actions, along with the visual content
around them, to represent relations between different parts of the image. The re
sults suggest that the GAP is essential for extracting visual relations that go
beyond the immediate visual content. Our approach reaches state-of-the-art perfo
rmance on several visual reasoning tasks being more sample-efficient, and genera
lizing better to out-of-distribution visual inputs than prior models.
****

Didi Zhu,Yibing Song,Tao Shen,Ziyu Zhao,Jinluan Yang,Min Zhang,Chao Wu
REMEDY: Recipe Merging Dynamics in Large Vision-Language Models
Model merging has emerged as a powerful technique for combining task-specific vi
sion models into a unified and multi-functional model. Previous methods represen
ted by task arithmetic, have demonstrated effectiveness and scalability in this
domain. When large vision-language models (LVLMs) arise with model size scaling
up, this design becomes challenging to fuse different instruction-tuned LVLMs fo
r generalization enhancement. The large scale and multi-modal nature of LVLMs pr
esent unique obstacles, including constructing reusable and modular components t
o accommodate the multi-component architecture of LVLMs and the requirement for
dynamic fusion based on multi-modal input tokens. To address these challenges, w
e propose the \textbf{RE}cipe \textbf{ME}rging \textbf{DY}namics (REMEDY) method
, a scalable and flexible paradigm for model merging in LVLMs. We first define r
eusable modules termed \textit{recipes} including the projector and shallow LLM
layers, enhancing visual-language understanding. Then, we introduce a modality-a
ware allocator dynamically generates weights in a one-shot manner based on input
 relevance to existing recipes, enabling efficient cross-modal knowledge integra
tion. REMEDY thus offers an adaptive solution for LVLMs to tackle both seen (i.e
., multi-task learning) and unseen (i.e., zero-shot generalization) tasks. Exper
imental results demonstrate that our method consistently improves performance on
 both seen and unseen tasks, underscoring the effectiveness of REMEDY in diverse
 multi-modal scenarios.
****

Pengxin Guo,Shuang Zeng,Yanran Wang,Huijie Fan,Feifei Wang,Liangqiong Qu
Selective Aggregation for Low-Rank Adaptation in Federated Learning
We investigate LoRA in federated learning through the lens of the asymmetry anal
ysis of the learned A and B matrices. In doing so, we uncover that A matri
ces are responsible for learning general knowledge, while B matrices focus on
capturing client-specific knowledge. Based on this finding, we introduce Federat
ed Share-A Low-Rank Adaptation (FedSA-LoRA), which employs two low-rank trainabl
e matrices A and B to model the weight update, but only A matrices are sha
red with the server for aggregation. Moreover, we delve into the relationship be
tween the learned A and B matrices in other LoRA variants, such as rsLoRA an
d VeRA, revealing a consistent pattern. Consequently, we extend our FedSA-LoRA m
ethod to these LoRA variants, resulting in FedSA-rsLoRA and FedSA-VeRA. In this
way, we establish a general paradigm for integrating LoRA with FL, offering guid
ance for future work on subsequent LoRA variants combined with FL. Extensive exp
erimental results on natural language understanding and generation tasks demonst
rate the effectiveness of the proposed method. Our code is available at https://
github.com/Pengxin-Guo/FedSA-LoRA.
****

Assaf Ben-Kish,Itamar Zimerman,Shady Abu-Hussein,Nadav Cohen,Amir Globerson,Lior
 Wolf,Raja Giryes
DeciMamba: Exploring the Length Extrapolation Potential of Mamba
Long-range sequence processing poses a significant challenge for Transformers du
e to their quadratic complexity in input length. A promising alternative is Mamb
a, which demonstrates high performance and achieves Transformer-level capabiliti
es while requiring substantially fewer computational resources. In this paper we
 explore the length-generalization capabilities of Mamba, which we find to be re
latively limited. Through a series of visualizations and analyses we identify th

at the limitations arise from a restricted effective receptive field, dictated b
y the sequence length used during training. To address this constraint, we intro
duce DeciMamba, a context-extension method specifically designed for Mamba. This
 mechanism, built on top of a hidden filtering mechanism embedded within the S6
layer, enables the trained model to extrapolate well even without additional tra
ining. Empirical experiments over real-world long-range NLP tasks show that Deci
Mamba can extrapolate to context lengths that are significantly longer than the
ones seen during training, while enjoying faster inference. We will release our
code and models.
****
Xiang Li,Cristina Mata,Jongwoo Park,Kumara Kahatapitiya,Yoo Sung Jang,Jinghuan S
hang,Kanchana Ranasinghe,Ryan D Burgert,Mu Cai,Yong Jae Lee,Michael S Ryoo
LLaRA: Supercharging Robot Learning Data for Vision-Language Policy
Vision Language Models (VLMs) have recently been leveraged to generate robotic a
ctions, forming Vision-Language-Action (VLA) models. However, directly adapting
a pretrained VLM for robotic control remains challenging, particularly when cons
trained by a limited number of robot demonstrations. In this work, we introduce
LLaRA: Large Language and Robotics Assistant, a framework that formulates robot
action policy as visuo-textual conversations and enables an efficient transfer o
f a pretrained VLM into a powerful VLA, motivated by the success of visual instr
uction tuning in Computer Vision. First, we present an automated pipeline to gen
erate conversation-style instruction tuning data for robots from existing behavi
or cloning datasets, aligning robotic actions with image pixel coordinates. Furt
her, we enhance this dataset in a self-supervised manner by defining six auxilia
ry tasks, without requiring any additional action annotations. We show that a VL
M finetuned with a limited amount of such datasets can produce meaningful action
 decisions for robotic control. Through experiments across multiple simulated an
d real-world tasks, we demonstrate that LLaRA achieves state-of-the-art performa
nce while preserving the generalization capabilities of large language models. T
he code, datasets, and pretrained models are available at https://github.com/Los
tXine/LLaRA.
****
Paul Gavrikov,Jovita Lukasik,Steffen Jung,Robert Geirhos,Muhammad Jehanzeb Mirza
,Margret Keuper,Janis Keuper
Can We Talk Models Into Seeing the World Differently?
Unlike traditional vision-only models, vision language models (VLMs) offer an in
tuitive way to access visual content through language prompting by combining a l
arge language model (LLM) with a vision encoder. However, both the LLM and the v
ision encoder come with their own set of biases, cue preferences, and shortcuts,
 which have been rigorously studied in uni-modal models. A timely question is ho
w such (potentially misaligned) biases and cue preferences behave under multi-mo
dal fusion in VLMs.
As a first step towards a better understanding, we investigate a particularly we
ll-studied vision-only bias - the texture vs. shape bias and the dominance of lo
cal over global information.
As expected, we find that VLMs inherit this bias to some extent from their visio
n encoders. Surprisingly, the multi-modality alone proves to have important effe
cts on the model behavior, i.e., the joint training and the language querying ch
ange the way visual cues are processed.
While this direct impact of language-informed training on a model's visual perce
ption is intriguing, it raises further questions on our ability to actively stee
r a model's output so that its prediction is based on particular visual cues of
the user's choice.
Interestingly, VLMs have an inherent tendency to recognize objects based on shap
e information, which is different from what a plain vision encoder would do. Fur
ther active steering towards shape-based classifications through language prompt
s is however limited. In contrast, active VLM steering towards texture-based dec
isions through simple natural language prompts is often more successful.
****
Jiannan Huang,Jun Hao Liew,Hanshu Yan,Yuyang Yin,Yao Zhao,Humphrey Shi,Yunchao W

ei
ClassDiffusion: More Aligned Personalization Tuning with Explicit Class Guidance
Recent text-to-image customization works have proven successful in generating im
ages of given concepts by fine-tuning diffusion models on a few examples. Howeve
r, tuning-based methods inherently tend to overfit the concepts, resulting in fa
ilure to create the concept under multiple conditions (*e.g.*, headphone is miss
ing when generating "a <sks>`dog wearing a headphone"). Interestingly, we notice
 that the base model before fine-tuning exhibits the capability to compose the b
ase concept with other elements (*e.g.*, "a dog wearing a headphone"), implying
that the compositional ability only disappears after personalization tuning. We
observe a semantic shift in the customized concept after fine-tuning, indicating
 that the personalized concept is not aligned with the original concept, and fur
ther show through theoretical analyses that this semantic shift leads to increas
ed difficulty in sampling the joint conditional probability distribution, result
ing in the loss of the compositional ability. Inspired by this finding, we prese
nt **ClassDiffusion**, a technique that leverages a **semantic preservation loss
** to explicitly regulate the concept space when learning a new concept. Althoug
h simple, this approach effectively prevents semantic drift during the fine-tuni
ng process of the target concepts. Extensive qualitative and quantitative experi
ments demonstrate that the use of semantic preservation loss effectively improve
s the compositional abilities of fine-tuning models. Lastly, we also extend our
ClassDiffusion to personalized video generation, demonstrating its flexibility.
****
Santiago Cortes-Gomez,Carlos Miguel Patiño,Yewon Byun,Steven Wu,Eric Horvitz,Bry
an Wilder
Utility-Directed Conformal Prediction: A Decision-Aware Framework for Actionable
 Uncertainty Quantification
There is increasing interest in ``decision-focused" machine learning methods whi
ch train models to account for how their predictions are used in downstream opti
mization problems. Doing so can often improve performance on subsequent decision
 problems. However, current methods for uncertainty quantification do not incorp
orate any information at all about downstream decisions. We develop a framework
based on conformal prediction to produce prediction sets that account for a down
stream decision loss function, making them more appropriate to inform high-stake
s decision-making. Our approach harnesses the strengths of conformal methods—mod
ularity, model-agnosticism, and statistical coverage guarantees—while incorporat
ing downstream decisions and user-specified utility functions. We prove that our
 methods retain standard coverage guarantees. Empirical evaluation across a ran
ge of datasets and utility metrics demonstrates that our methods achieve signifi
cantly lower decision loss compared to standard conformal methods. Additionally,
 we present a real-world use case in healthcare diagnosis, where our method effe
ctively incorporates the hierarchical structure of dermatological diseases. It s
uccessfully generates sets with coherent diagnostic meaning, aiding the triage p
rocess during dermatology diagnosis and illustrating how our method can ground h
igh-stakes decision-making on external domain knowledge.
****
Johannes Hertrich,Tim Jahn,Michael Quellmalz
Fast Summation of Radial Kernels via QMC Slicing
The fast computation of large kernel sums is a challenging task, which arises as
 a subproblem in any kernel method. We approach the problem by slicing, which r
elies on random projections to one-dimensional subspaces and fast Fourier summat
ion. We prove bounds for the slicing error and propose a quasi-Monte Carlo (QMC)
 approach for selecting the projections based on spherical quadrature rules. Num
erical examples demonstrate that our QMC-slicing approach significantly outperfo
rms existing methods like (QMC-)random Fourier features, orthogonal Fourier feat
ures or non-QMC slicing on standard test datasets.
****
Hoang Phuc Hau Luu,Hanlin Yu,Bernardo Williams,Marcelo Hartmann,Arto Klami
Stochastic variance-reduced Gaussian variational inference on the Bures-Wasserst
ein manifold

Optimization in the Bures-Wasserstein space has been gaining popularity in the m
achine learning community since it draws connections between variational inferen
ce and Wasserstein gradient flows. The variational inference objective function
of Kullback–Leibler divergence can be written as the sum of the negative entropy
 and the potential energy, making forward-backward Euler the method of choice. N
otably, the backward step admits a closed-form solution in this case, facilitati
ng the practicality of the scheme. However, the forward step is not exact since
the Bures-Wasserstein gradient of the potential energy involves "intractable" ex
pectations. Recent approaches propose using the Monte Carlo method -- in practic
e a single-sample estimator -- to approximate these terms, resulting in high var
iance and poor performance. We propose a novel variance-reduced estimator based
on the principle of control variates. We theoretically show that this estimator
has a smaller variance than the Monte-Carlo estimator in scenarios of interest.
We also prove that variance reduction helps improve the optimization bounds of t
he current analysis. We demonstrate that the proposed estimator gains order-of-m
agnitude improvements over the previous Bures-Wasserstein methods.
****

Shiguang Wu,Yaqing Wang,Quanming Yao
Why In-Context Learning Models are Good Few-Shot Learners?
We explore in-context learning (ICL) models from a learning-to-learn perspective
. Unlike studies that identify specific learning algorithms in ICL models, we co
mpare ICL models with typical meta-learners to understand their superior perform
ance. We theoretically prove the expressiveness of ICL models as learning algori
thms and examine their learnability and generalizability.
Our findings show that ICL with transformers
can effectively construct data-dependent learning algorithms instead of directly
 follow existing ones
(including gradient-based, metric-based, and amortization-based meta-learners).
The construction of such learning algorithm is determined by the pre-training pr
ocess, as a function fitting the training distribution, which raises generalizab
ility as an important issue.
With above understanding, we propose strategies to transfer techniques for class
ical deep networks to meta-level to further improve ICL. As examples, we impleme
nt meta-level meta-learning for domain adaptability with limited data and meta-l
evel curriculum learning for accelerated convergence during pre-training, demons
trating their empirical effectiveness.
****

Qihang Zhang,Yinghao Xu,Chaoyang Wang,Hsin-Ying Lee,Gordon Wetzstein,Bolei Zhou,
Ceyuan Yang
3DitScene: Editing Any Scene via Language-guided Disentangled Gaussian Splatting
Scene image editing is crucial for entertainment, photography, and advertising d
esign. Existing methods solely focus on either 2D individual object or 3D global
 scene editing. This results in a lack of a unified approach to effectively cont
rol and manipulate scenes at the 3D level with different levels of granularity.
In this work, we propose 3DitScene, a novel and unified scene editing framework
leveraging language-guided disentangled Gaussian Splatting that enables seamless
 editing from 2D to 3D, allowing precise control over scene composition and indi
vidual objects. We first incorporate 3D Gaussians that are refined through gener
ative priors and optimization techniques. Language features from CLIP then intro
duce semantics into 3D geometry for object disentanglement. With the disentangle
d Gaussians, 3DitScene allows for manipulation at both the global and individual
 levels, revolutionizing creative expression and empowering control over scenes
and objects. Experimental results demonstrate the effectiveness and versatility
of 3DitScene in scene image editing.
****

Fu-Yun Wang,Yunhao Shui,Jingtan Piao,Keqiang Sun,Hongsheng Li
Diffusion-NPO: Negative Preference Optimization for Better Preference Aligned Ge
neration of Diffusion Models
Diffusion models have made substantial advances in image generation, yet models
trained on large, unfiltered datasets often yield outputs misaligned with human

preferences. Numerous methods have already been proposed to fine-tune pre-traine
d diffusion models, achieving notable improvements in aligning generated outputs
 with human preferences. However, we point out that existing preference alignmen
t methods neglect the critical role of handling unconditional/negative-condition
al outputs, leading to a diminished capacity to avoid generating undesirable ou
tcomes. This oversight limits the efficacy of classifier-free guidance (CFG), wh
ich relies on the contrast between conditional generation and unconditional/nega
tive-conditional generation to optimize output quality. In response, we propose
a straightforward but consistently effective approach that involves training a m
odel specifically attuned to negative preferences. This method does not require
new training strategies or datasets but rather involves minor modifications to e
xisting techniques. Our approach integrates seamlessly with models such as SD15,
 SDXL, video diffusion models and models that have undergone preference optimiza
tion, consistently enhancing their ability to produce more human preferences ali
gned outputs.
****

Han-Hung Lee,Yiming Zhang,Angel X Chang
Duoduo CLIP: Efficient 3D Understanding with Multi-View Images
We introduce Duoduo CLIP, a model for 3D representation learning that learns sha
pe encodings from multi-view images instead of point-clouds. The choice of multi
-view images allows us to leverage 2D priors from off-the-shelf CLIP models to f
acilitate fine-tuning with 3D data. Our approach not only shows better generaliz
ation compared to existing point cloud methods, but also reduces GPU requirement
s and training time. In addition, the model is modified with cross-view attentio
n to leverage information across multiple frames of the object which further boo
sts performance. Notably, our model is permutation invariant to the order of mul
ti-view images while being pose-free. Compared to the current SOTA point cloud m
ethod that requires 480 A100 hours to train 1 billion model parameters we only r
equire 57 A5000 hours and 87 million parameters. Multi-view images also provide
more flexibility including being able to encode objects with a variable number o
f images, and performance scales when more views are used. In contrast, point cl
oud based methods require an entire scan or model of the object. We showcase thi
s flexibility with benchmarks from images of real-world objects. Our model also
achieves better performance in more fine-grained text to shape retrieval, demons
trating better text-and-shape alignment than point cloud based models.
****

Henry Zheng,Hao Shi,Qihang Peng,Yong Xien Chng,Rui Huang,Yepeng Weng,zhongchao s
hi,Gao Huang
DenseGrounding: Improving Dense Language-Vision Semantics for Ego-centric 3D Vis
ual Grounding
Enabling intelligent agents to comprehend and interact with 3D environments thro
ugh natural language is crucial for advancing robotics and human-computer intera
ction. A fundamental task in this field is ego-centric 3D visual grounding, wher
e agents locate target objects in real-world 3D spaces based on verbal descripti
ons. However, this task faces two significant challenges: (1) loss of fine-grain
ed visual semantics due to sparse fusion of point clouds with ego-centric multi-
view images, (2) limited textual semantic context due to arbitrary language desc
riptions. We propose DenseGrounding, a novel approach designed to address these
issues by enhancing both visual and textual semantics. For visual features, we i
ntroduce the Hierarchical Scene Semantic Enhancer, which retains dense semantics
 by capturing fine-grained global scene features and facilitating cross-modal al
ignment. For text descriptions, we propose a Language Semantic Enhancer that lev
erage large language models to provide rich context and diverse language descrip
tions with additional context during model training. Extensive experiments show
that DenseGrounding significantly outperforms existing methods in overall accura
cy, achieving improvements of **5.81%** and **7.56%** when trained on the compre
hensive full training dataset and smaller mini subset, respectively, further adv
ancing the SOTA in ego-centric 3D visual grounding. Our method also achieves **1
st place** and receives **Innovation Award** in the 2024 Autonomous Grand Challe
nge Multi-view 3D Visual Grounding Track, validating its effectiveness and robus

tness.
****
Jiate Li,Meng Pang,Yun Dong,Jinyuan Jia,Binghui Wang
Provably Robust Explainable Graph Neural Networks against Graph Perturbation Attacks
Explaining Graph Neural Network (XGNN) has gained growing attention to facilitate the trust of using GNNs, which is the mainstream method to learn graph data. Despite their growing attention, Existing XGNNs focus on improving the explanation performance, and its robustness under attacks is largely unexplored. We noticed that an adversary can slightly perturb the graph structure such that the explanation result of XGNNs is largely changed. Such vulnerability of XGNNs could cause serious issues particularly in safety/security-critical applications. In this paper, we take the first step to study the robustness of XGNN against graph perturbation attacks, and propose XGNNCert, the first provably robust XGNN. Particularly, our XGNNCert can provably ensure the explanation result for a graph under the worst-case graph perturbation attack is close to that without the attack, while not affecting the GNN prediction, when the number of perturbed edges is bounded. Evaluation results on multiple graph datasets and GNN explainers show the effectiveness of XGNNCert.
****
Yi Li,Wai Ming Tai
Near-optimal Active Regression of Single-Index Models
The active regression problem of the single-index model is to solve $\min_x \lVert f(Ax)-b\rVert_p$, where A is fully accessible and b can only be accessed via entry queries, with the goal of minimizing the number of queries to the entries of b.
When f is Lipschitz, previous results only obtain constant-factor approximations. This work presents the first algorithm that provides a $(1+\varepsilon)$-approximation solution by querying $\tilde{O}(d^{\frac{p}{2}\vee 1}/\varepsilon^{p\vee 2})$ entries of b. This query complexity is also shown to be optimal up to logarithmic factors for $p\in [1,2]$ and the ε-dependence of $1/\varepsilon^p$ is shown to be optimal for $p>2$.
****
Alexander Atanasov,Alexandru Meterez,James B Simon,Cengiz Pehlevan
The Optimization Landscape of SGD Across the Feature Learning Strength
We consider neural networks (NNs) where the final layer is down-scaled by a fixed hyperparameter γ.
Recent work has identified γ as controlling the strength of feature learning.
As γ increases, network evolution changes from "lazy" kernel dynamics to "rich" feature-learning dynamics, with a host of associated benefits including improved performance on common tasks.
In this work, we conduct a thorough empirical investigation of the effect of scaling γ across a variety of models and datasets in the online training setting.
We first examine the interaction of γ with the learning rate η, identifying several scaling regimes in the γ-η plane which we explain theoretically using a simple model.
We find that the optimal learning rate η^* scales non-trivially with γ. In particular, $\eta^* \propto \gamma^2$ when $\gamma \ll 1$ and $\eta^* \propto \gamma^{2/L}$ when $\gamma \gg 1$ for a feed-forward network of depth L.
Using this optimal learning rate scaling, we proceed with an empirical study of the under-explored ``ultra-rich'' $\gamma \gg 1$ regime.
We find that networks in this regime display characteristic loss curves, starting with a long plateau followed by a drop-off, sometimes followed by one or more additional staircase steps.
We find networks of different large γ values optimize along similar trajectories up to a reparameterization of time.
We further find that optimal online performance is often found at large γ and could be missed if this hyperparameter is not tuned.

Our findings indicate that analytical study of the large-γ limit may yiel
d useful insights into the dynamics of representation learning in performant mod
els.
****

Chaodong Xiao,Minghan Li,Zhengqiang ZHANG,Deyu Meng,Lei Zhang
Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fus
ion
Selective state space models (SSMs), such as Mamba, highly excel at capturing lo
ng-range dependencies in 1D sequential data, while their applications to 2D visi
on tasks still face challenges. Current visual SSMs often convert images into 1D
 sequences and employ various scanning patterns to incorporate local spatial dep
endencies. However, these methods are limited in effectively capturing the compl
ex image spatial structures and the increased computational cost caused by the l
engthened scanning paths. To address these limitations, we propose Spatial-Mamba
, a novel approach that establishes neighborhood connectivity directly in the st
ate space. Instead of relying solely on sequential state transitions, we introdu
ce a structure-aware state fusion equation, which leverages dilated convolutions
 to capture image spatial structural dependencies, significantly enhancing the f
low of visual contextual information. Spatial-Mamba proceeds in three stages: in
itial state computation in a unidirectional scan, spatial context acquisition th
rough structure-aware state fusion, and final state computation using the observ
ation equation. Our theoretical analysis shows that Spatial-Mamba unifies the or
iginal Mamba and linear attention under the same matrix multiplication framework
, providing a deeper understanding of our method. Experimental results demonstra
te that Spatial-Mamba, even with a single scan, attains or surpasses the state-o
f-the-art SSM-based models in image classification, detection and segmentation.
Source codes and trained models can be found at \url{ https://github.com/EdwardC
hasel/Spatial-Mamba }.
****

Jiajian Xie,Shengyu Zhang,Mengze Li,chengfei lv,Zhou Zhao,Fei Wu
EcoFace: Audio-Visual Emotional Co-Disentanglement Speech-Driven 3D Talking Face
 Generation
Speech-driven 3D facial animation has attracted significant attention due to its
 wide range of applications in animation production and virtual reality. Recent
research has explored speech-emotion disentanglement to enhance facial expressio
ns rather than manually assigning emotions. However, this approach face issues s
uch as feature confusion, emotions weakening and mean-face. To address these iss
ues, we present EcoFace, a framework that (1) proposes a novel collaboration obj
ective to provide a explicit signal for emotion representation learning from the
 speaker's expressive movements and produced sounds, constructing an audio-visua
l joint and coordinated emotion space that is independent of speech content. (2)
 constructs a universal facial motion distribution space determined by speech fe
atures and implement speaker-specific generation. Extensive experiments show tha
t our method achieves more generalized and emotionally realistic talking face ge
neration compared to previous methods.
****

Yushun Zhang,Congliang Chen,Ziniu Li,Tian Ding,Chenwei Wu,Diederik P Kingma,Yiny
u Ye,Zhi-Quan Luo,Ruoyu Sun
Adam-mini: Use Fewer Learning Rates To Gain More
We propose Adam-mini, an optimizer that achieves on-par or better performance th
an AdamW with 50% less memory footprint. Adam-mini reduces memory by cutting d
own the learning rate resources in Adam (i.e., $1/\sqrt{v}$). By delving into th
e Hessian structure of neural nets, we find Adam’s v might not function at its
 full potential as effectively as we expected. We find that ≥ 99.9% of thes
e learning rates in v could be harmlessly removed if we (1) carefully partitio
n the parameters into blocks following our proposed principle on Hessian structu
re; (2) assign a single but good learning rate to each parameter block. We then
provide one simple way to find good learning rates and propose Adam-mini. Empiri
cally, we verify that Adam-mini performs on par or better than AdamW on various
language models sized from 39M to 13B for pre-training, supervised fine-tuning,

and RLHF. The reduced memory footprint of Adam-mini also alleviates communicatio
n overheads among GPUs, thereby increasing throughput. For instance, Adam-mini a
chieves 49.6% higher throughput than AdamW when pre-training Llama 2-7B on $2\
times$ A800-80GB GPUs, which saves 33% wall-clock time for pre-training.
****

Seong Ho Pahng,Sahand Hormoz
Improving Graph Neural Networks by Learning Continuous Edge Directions
Graph Neural Networks (GNNs) traditionally employ a message-passing mechanism th
at resembles diffusion over undirected graphs, which often leads to homogenizati
on of node features and reduced discriminative power in tasks such as node class
ification. Our key insight for addressing this limitation is to assign fuzzy edg
e directions---that can vary continuously from node i pointing to node j to
vice versa---to the edges of a graph so that features can preferentially flow in
 one direction between nodes to enable long-range information transmission acros
s the graph. We also introduce a novel complex-valued Laplacian for directed gra
phs with fuzzy edges where the real and imaginary parts represent information fl
ow in opposite directions. Using this Laplacian, we propose a general framework,
 called Continuous Edge Direction (CoED) GNN, for learning on graphs with fuzzy
edges and prove its expressivity limits using a generalization of the Weisfeiler
-Leman (WL) graph isomorphism test for directed graphs with fuzzy edges. Our arc
hitecture aggregates neighbor features scaled by the learned edge directions and
 processes the aggregated messages from in-neighbors and out-neighbors separatel
y alongside the self-features of the nodes. Since continuous edge directions are
 differentiable, they can be learned jointly with the GNN weights via gradient-b
ased optimization. CoED GNN is particularly well-suited for graph ensemble data
where the graph structure remains fixed but multiple realizations of node featur
es are available, such as in gene regulatory networks, web connectivity graphs,
and power grids. We demonstrate through extensive experiments on both synthetic
and real graph ensemble datasets that learning continuous edge directions signif
icantly improves performance both for undirected and directed graphs compared wi
th existing methods.
****

Xingwei Qu,yuelin bai,Yinghao Ma,Ziya Zhou,Ka Man Lo,Jiaheng Liu,Ruibin Yuan,Lej
un Min,Xueling Liu,Tianyu Zhang,Xeron Du,Shuyue Guo,Yiming Liang,Yizhi LI,Shangd
a Wu,Junting Zhou,Tianyu Zheng,Ziyang Ma,Fengze Han,Wei Xue,Gus Xia,Emmanouil Be
netos,Xiang Yue,Chenghua Lin,Xu Tan,Wenhao Huang,Jie Fu,Ge Zhang
MuPT: A Generative Symbolic Music Pretrained Transformer
In this paper, we explore the application of Large Language Models (LLMs) to the
 pre-training of music. While the prevalent use of MIDI in music modeling is wel
l-established, our findings suggest that LLMs are inherently more compatible wit
h ABC Notation, which aligns more closely with their design and strengths, there
by enhancing the model's performance in musical composition.
To address the challenges associated with misaligned measures from different tra
cks during generation, we propose the development of a $\underline{S}$ynchronize
d $\underline{M}$ulti-$\underline{T}$rack ABC Notation ($\textbf{SMT-ABC Notatio
n}$), which aims to preserve coherence across multiple musical tracks.
Our contributions include a series of models capable of handling up to 8192 toke
ns, covering 90\% of the symbolic music data in our training set. Furthermore, w
e explore the implications of the $\underline{S}$ymbolic $\underline{M}$usic $\u
nderline{S}$caling Law ($\textbf{SMS Law}$) on model performance. The results in
dicate a promising research direction in music generation, offering extensive re
sources for further research through our open-source contributions.
****

Michael Scholkemper,Xinyi Wu,Ali Jadbabaie,Michael T Schaub
Residual Connections and Normalization Can Provably Prevent Oversmoothing in GNN
s
Residual connections and normalization layers have become standard design choice
s for graph neural networks (GNNs), and were proposed as solutions to the mitiga
te the oversmoothing problem in GNNs. However, how exactly these methods help al
leviate the oversmoothing problem from a theoretical perspective is not well und

erstood. In this work, we provide a formal and precise characterization of (line arized) GNNs with residual connections and normalization layers. We establish th at (a) for residual connections, the incorporation of the initial features at ea ch layer can prevent the signal from becoming too smooth, and determines the sub space of possible node representations; (b) batch normalization prevents a compl ete collapse of the output embedding space to a one-dimensional subspace through the individual rescaling of each column of the feature matrix. This results in the convergence of node representations to the top-k eigenspace of the message-p assing operator; (c) moreover, we show that the centering step of a normalizatio n layer — which can be understood as a projection — alters the graph signal in m essage-passing in such a way that relevant information can become harder to extr act. Building on the last theoretical insight, we introduce GraphNormv2, a novel and principled normalization layer. GraphNormv2 features a learnable centering step designed to preserve the integrity of the original graph signal. Experiment al results corroborate the effectiveness of our method, demonstrating improved p erformance across various GNN architectures and tasks.
****

Qixin Zhang,Zongqi Wan,Yu Yang,Li Shen,Dacheng Tao
Near-Optimal Online Learning for Multi-Agent Submodular Coordination: Tight Appr oximation and Communication Efficiency
Coordinating multiple agents to collaboratively maximize submodular functions in unpredictable environments is a critical task with numerous applications in mac hine learning, robot planning and control. The existing approaches, such as the OSG algorithm, are often hindered by their poor approximation guarantees and th e rigid requirement for a fully connected communication graph. To address these challenges, we firstly present a $\textbf{MA-OSMA}$ algorithm, which employs the multi-linear extension to transfer the discrete submodular maximization problem into a continuous optimization, thereby allowing us to reduce the strict depend ence on a complete graph through consensus techniques. Moreover, $\textbf{MA-OSM A}$ leverages a novel surrogate gradient to avoid sub-optimal stationary points. To eliminate the computationally intensive projection operations in $\textbf{MA -OSMA}$, we also introduce a projection-free $\textbf{MA-OSEA}$ algorithm, which effectively utilizes the KL divergence by mixing a uniform distribution. Theore tically, we confirm that both algorithms achieve a regret bound of $\widetilde{O }(\sqrt{\frac{C_{T}T}{1-\beta}})$ against a $(\frac{1-e^{-c}}{c})$-approximatio n to the best comparator in hindsight, where C_{T} is the deviation of maximiz er sequence, β is the spectral gap of the network and c is the joint cur vature of submodular objectives. This result significantly improves the $(\frac{ 1}{1+c})$-approximation provided by the state-of-the-art OSG algorithm. Finally, we demonstrate the effectiveness of our proposed algorithms through simulation-based multi-target tracking.
****

Harin Lee,Min-hwan Oh
Minimax Optimal Reinforcement Learning with Quasi-Optimism
In our quest for a reinforcement learning (RL) algorithm that is both practical and provably optimal, we introduce EQO (Exploration via Quasi-Optimism). Unlike existing minimax optimal approaches, EQO avoids reliance on empirical variances and employs a simple bonus term proportional to the inverse of the state-action visit count. Central to EQO is the concept of *quasi-optimism*, where estimated values need not be fully optimistic, allowing for a simpler yet effective explor ation strategy. The algorithm achieves the sharpest known regret bound for tabul ar RL under the mildest assumptions, proving that fast convergence can be attain ed with a practical and computationally efficient approach. Empirical evaluation s demonstrate that EQO consistently outperforms existing algorithms in both regr et performance and computational efficiency, providing the best of both theoreti cal soundness and practical effectiveness.
****

Junqi Jiang,Tom Bewley,Saumitra Mishra,Freddy Lecue,Manuela Veloso
Interpreting Language Reward Models via Contrastive Explanations
Reward models (RMs) are a crucial component in the alignment of large language m

odels’ (LLMs) outputs with human values. RMs approximate human preferences over possible LLM responses to the same prompt by predicting and comparing reward scores. However, as they are typically modified versions of LLMs with scalar output heads, RMs are large black boxes whose predictions are not explainable. More transparent RMs would enable improved trust in the alignment of LLMs. In this work, we propose to use contrastive explanations to explain any binary response comparison made by an RM. Specifically, we generate a diverse set of new comparisons similar to the original one to characterise the RM’s local behaviour. The perturbed responses forming the new comparisons are generated to explicitly modify manually specified high-level evaluation attributes, on which analyses of RM behaviour are grounded. In quantitative experiments, we validate the effectiveness of our method for finding high-quality contrastive explanations. We then showcase the qualitative usefulness of our method for investigating global sensitivity of RMs to each evaluation attribute, and demonstrate how representative examples can be automatically extracted to explain and compare behaviours of different RMs. We see our method as a flexible framework for RM explanation, providing a basis for more interpretable and trustworthy LLM alignment.
****

Xiaoyi Zhu,Zengfeng Huang
Lipschitz Bandits in Optimal Space
This paper considers the Lipschitz bandit problem, where the set of arms is continuous and the expected reward is a Lipschitz function over the arm space. This problem has been extensively studied. Prior algorithms need to store the reward information of all visited arms, leading to significant memory consumption. We address this issue by introducing an algorithm named Log-space Lipschitz bandits (Log-Li), which achieves an optimal (up to logarithmic factors) regret of $\widetilde{O}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ while only uses $O\left(\log T\right)$ bits of memory. Additionally, we provide a complexity analysis for this problem, demonstrating that $\Omega\left(\log T\right)$ bits of space are necessary for any algorithm to achieve the optimal regret. We also conduct numerical simulations, and the results show that our new algorithm achieves regret comparable to the state-of-the-art while reducing memory usage by orders of magnitude.
****

Yuhang Wang,Hanwei Guo,Sizhe Wang,Long Qian,Xuguang Lan
Bootstrapped Model Predictive Control
Model Predictive Control (MPC) has been demonstrated to be effective in continuous control tasks. When a world model and a value function are available, planning a sequence of actions ahead of time leads to a better policy. Existing methods typically obtain the value function and the corresponding policy in a model-free manner. However, we find that such an approach struggles with complex tasks, resulting in poor policy learning and inaccurate value estimation. To address this problem, we leverage the strengths of MPC itself. In this work, we introduce Bootstrapped Model Predictive Control (BMPC), a novel algorithm that performs policy learning in a bootstrapped manner. BMPC learns a network policy by imitating an MPC expert, and in turn, uses this policy to guide the MPC process. Combined with model-based TD-learning, our policy learning yields better value estimation and further boosts the efficiency of MPC. We also introduce a lazy reanalyze mechanism, which enables computationally efficient imitation learning. Our method achieves superior performance over prior works on diverse continuous control tasks. In particular, on challenging high-dimensional locomotion tasks, BMPC significantly improves data efficiency while also enhancing asymptotic performance and training stability, with comparable training time and smaller network sizes. Code is available at https://github.com/wertyuilife2/bmpc.
****

Joshua Huang Wu,Hari Koneru,James Russell Ravenel,Anshuman Sabath,James Michael Roach,Shaun Sze-Xian Lim,MICHAEL R TADROSS,Alex H Williams,Timothy W DUNN
Disentangling 3D Animal Pose Dynamics with Scrubbed Conditional Latent Variables
Methods for tracking lab animal movements in unconstrained environments have become increasingly common and powerful tools for neuroscience. The prevailing hypothesis is that animal behavior in these environments comprises sequences of disc

rete stereotyped body movements ("motifs" or "actions"). However, the same actio
n can occur at different speeds or heading directions, and the same action may m
anifest slightly differently across subjects due to, for example, variation in b
ody size. These and other forms of nuisance variability complicate attempts to q
uantify animal behavior in terms of discrete action sequences and draw meaningfu
l comparisons across individual subjects. To address this, we present a framewor
k for motion analysis that uses conditional variational autoencoders in conjunct
ion with adversarial learning paradigms to disentangle behavioral factors. We de
monstrate the utility of this approach in downstream tasks such as clustering, d
ecodability, and motion synthesis. Further, we apply our technique to improve di
sease detection in a Parkinsonian mouse model.
****
Yair Schiff,Subham Sekhar Sahoo,Hao Phung,Guanghan Wang,Sam Boshar,Hugo Dalla-to
rre,Bernardo P de Almeida,Alexander M Rush,Thomas PIERROT,Volodymyr Kuleshov
Simple Guidance Mechanisms for Discrete Diffusion Models
Diffusion models for continuous data gained widespread adoption owing to their h
igh quality generation and control mechanisms. However, controllable diffusion o
n discrete data faces challenges given that continuous guidance methods do not d
irectly apply to discrete diffusion. Here, we provide a straightforward derivati
on of classifier-free and classifier-based guidance for discrete diffusion, as w
ell as a new class of diffusion models that leverage uniform noise and that are
more guidable because they can continuously edit their outputs. We improve the q
uality of these models with a novel continuous-time variational lower bound that
 yields state-of-the-art performance, especially in settings involving guidance
or fast generation. Empirically, we demonstrate that our guidance mechanisms com
bined with uniform noise diffusion improve controllable generation relative to a
utoregressive and diffusion baselines on several discrete data domains, includin
g genomic sequences, small molecule design, and discretized image generation.
****
Sheng-Chieh Lin,Chankyu Lee,Mohammad Shoeybi,Jimmy Lin,Bryan Catanzaro,Wei Ping
MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS
State-of-the-art retrieval models typically address a straightforward search sce
nario, in which retrieval tasks are fixed (e.g., finding a passage to answer a s
pecific question) and only a single modality is supported for both queries and r
etrieved results. This paper introduces techniques for advancing information ret
rieval with multimodal large language models (MLLMs), enabling a broader search
scenario, termed universal multimodal retrieval, where multiple modalities and d
iverse retrieval tasks are accommodated. To this end, we first study fine-tuning
 an MLLM as a bi-encoder retriever on 10 datasets with 16 retrieval tasks. Our e
mpirical results show that the fine-tuned MLLM retriever is capable of understan
ding challenging queries, composed of both text and image, but it underperforms
compared to a smaller CLIP retriever in cross-modal retrieval tasks due to the m
odality bias exhibited by MLLMs. To address the issue, we propose modality-aware
 hard negative mining to mitigate the modality bias exhibited by MLLM retrievers
. Second, we propose continuously fine-tuning the universal multimodal retriever
 to enhance its text retrieval capability while preserving multimodal retrieval
capability. As a result, our model, MM-Embed, achieves state-of-the-art performa
nce on the multimodal retrieval benchmark M-BEIR, which spans multiple domains a
nd tasks, while also surpassing the state-of-the-art text retrieval model, NV-Em
bed-v1, on the MTEB retrieval benchmark. Finally, we explore prompting the off-t
he-shelf MLLMs as zero-shot rerankers to refine the ranking of the candidates fr
om the multimodal retriever. We find that, through prompt-and-reranking, MLLMs c
an further improve multimodal retrieval when the user queries (e.g., text-image
composed queries) are more complex and challenging to understand. These findings
 also pave the way for advancing universal multimodal retrieval in the future. W
e release the model weights at: https://huggingface.co/nvidia/MM-Embed.
****
Seungeun Rho,Laura Smith,Tianyu Li,Sergey Levine,Xue Bin Peng,Sehoon Ha
Language Guided Skill Discovery
Skill discovery methods enable agents to learn diverse emergent behaviors withou

t explicit rewards. To make learned skills useful for downstream tasks, obtainin g a semantically diverse repertoire of skills is crucial. While some approaches use discriminators to acquire distinguishable skills and others focus on increas ing state coverage, the direct pursuit of ‘semantic diversity’ in skills remains underexplored. We hypothesize that leveraging the semantic knowledge of large l anguage models (LLM) can lead us to improve semantic diversity of resulting beha viors. In this sense, we introduce Language Guided Skill Discovery (LGSD), a ski ll discovery framework that aims to directly maximize the semantic diversity bet ween skills. LGSD takes user prompts as input and outputs a set of semantically distinctive skills. The prompts serve as a means to constrain the search space i nto a semantically desired subspace, and the generated LLM outputs guide the age nt to visit semantically diverse states within the subspace. We demonstrate that LGSD enables legged robots to visit different user-intended areas on a plane by simply changing the prompt. Furthermore, we show that language guidance aids in discovering more diverse skills compared to five existing skill discovery metho ds in robot-arm manipulation environments. Lastly, LGSD provides a simple way of utilizing learned skills via natural language.
****
Ed Davis,Ian Gallagher,Daniel John Lawson,Patrick Rubin-Delanchy
Valid Conformal Prediction for Dynamic GNNs
Dynamic graphs provide a flexible data abstraction for modelling many sorts of r eal-world systems, such as transport, trade, and social networks. Graph neural n etworks (GNNs) are powerful tools allowing for different kinds of prediction and inference on these systems, but getting a handle on uncertainty, especially in dynamic settings, is a challenging problem.

In this work we propose to use a dynamic graph representation known in the tenso r literature as the unfolding, to achieve valid prediction sets via conformal pr ediction. This representation, a simple graph, can be input to any standard GNN and does not require any modification to existing GNN architectures or conformal prediction routines.

One of our key contributions is a careful mathematical consideration of the diff erent inference scenarios which can arise in a dynamic graph modelling context. For a range of practically relevant cases, we obtain valid prediction sets with almost no assumptions, even dispensing with exchangeability. In a more challengi ng scenario, which we call the semi-inductive regime, we achieve valid predictio n under stronger assumptions, akin to stationarity.

We provide real data examples demonstrating validity, showing improved accuracy over baselines, and sign-posting different failure modes which can occur when th ose assumptions are violated.
****
Yiwei Li,Sekeun Kim,Zihao Wu,Hanqi Jiang,Yi Pan,Pengfei Jin,Sifan Song,Yucheng S hi,Xiaowei Yu,Tianze Yang,Tianming Liu,Quanzheng Li,Xiang Li
ECHOPulse: ECG Controlled Echocardio-gram Video Generation
Echocardiography (ECHO) is essential for cardiac assessments, but its video qual ity and interpretation heavily relies on manual expertise, leading to inconsiste nt results from clinical and portable devices. ECHO video generation offers a so lution by improving automated monitoring through synthetic data and generating h igh-quality videos from routine health data. However, existing models often face high computational costs, slow inference, and rely on complex conditional promp ts that require experts' annotations. To address these challenges, we propose EC HOPulse, an ECG-conditioned ECHO video generation model. ECHOPulse introduces tw o key advancements: (1) it accelerates ECHO video generation by leveraging VQ-VA E tokenization and masked visual token modeling for fast decoding, and (2) it co nditions on readily accessible ECG signals, which are highly coherent with ECHO videos, bypassing complex conditional prompts. To the best of our knowledge, thi s is the first work to use time-series prompts like ECG signals for ECHO video g eneration. ECHOPulse not only enables controllable synthetic ECHO data generatio

n but also provides updated cardiac function information for disease monitoring and prediction beyond ECG alone. Evaluations on three public and private dataset s demonstrate state-of-the-art performance in ECHO video generation across both qualitative and quantitative measures. Additionally, ECHOPulse can be easily gen eralized to other modality generation tasks, such as cardiac MRI, fMRI, and 3D C T generation. We will make the synthetic ECHO dataset, along with the code and m odel, publicly available upon acceptance.
****
Leo Zhang,Kianoosh Ashouritaklimi,Yee Whye Teh,Rob Cornish
SymDiff: Equivariant Diffusion via Stochastic Symmetrisation
We propose SymDiff, a method for constructing equivariant diffusion models using the framework of stochastic symmetrisation. SymDiff resembles a learned data au gmentation that is deployed at sampling time, and is lightweight, computationall y efficient, and easy to implement on top of arbitrary off-the-shelf models. In contrast to previous work, SymDiff typically does not require any neural network components that are intrinsically equivariant, avoiding the need for complex pa rameterisations or the use of higher-order geometric features. Instead, our meth od can leverage highly scalable modern architectures as drop-in replacements for these more constrained alternatives. We show that this additional flexibility y ields significant empirical benefit for E(3)-equivariant molecular generation. T o the best of our knowledge, this is the first application of symmetrisation to generative modelling, suggesting its potential in this domain more generally.
****
Zhanfeng Mo,Long-Kai Huang,Sinno Jialin Pan
Parameter and Memory Efficient Pretraining via Low-rank Riemannian Optimization
Pretraining large language models often requires significant computational resou rces and memory due to their vast parameter amount. An effective approach to enh ance parameter efficiency in both training and inference is to parameterize each full-size weight as the product of two trainable low-rank factors. While low-ra nk fine-tuning has achieved great success, low-rank pretraining remains challeng ing as it requires learning extensive knowledge from scratch under the restricti ve low-rank parameterization. During standard low-rank pretraining, separately o ptimizing the low-rank factors introduces redundant information from the full gr adient, which hinders the learning process. To achieve efficient yet effective l ow-rank pretraining, we propose a **Lo**w-rank **R**iemannian **O**ptimizer (**L ORO**). At each LORO update step, the low-rank factor pairs are jointly updated to ensure their full-size product moves along the steepest descent direction on the low-rank manifold, without the need to compute any memory-intensive full-siz e matrices or gradients. Hence, our LORO finds low-rank models that achieve high performance comparable to full-size pretrained models, while significantly redu cing memory usage and accelerating both training and inference. A LLaMA 1B model pretrained with LORO achieves a perplexity score of 2\% better than the full-si ze baseline, with a 54\% reduction in model memory, a $\times1.8$ speedup in tra ining, and a $\times2.2$ speedup in inference. The code is available on https:// github.com/mzf666/LORO-main.
****
Augustinos D Saravanos,Hunter Kuperman,Alex Oshin,Arshiya Taj Abdul,Vincent Pace lli,Evangelos Theodorou
Deep Distributed Optimization for Large-Scale Quadratic Programming
Quadratic programming (QP) forms a crucial foundation in optimization, appearing in a broad spectrum of domains and serving as the basis for more advanced algor ithms. Consequently, as the scale and complexity of modern applications continue to grow, the development of efficient and reliable QP algorithms becomes increa singly vital. In this context, this paper introduces a novel deep learning-aided distributed optimization architecture designed for tackling large-scale QP prob lems. First, we combine the state-of-the-art Operator Splitting QP (OSQP) method with a consensus approach to derive **DistributedQP**, a new method tailored fo r network-structured problems, with convergence guarantees to optimality. Subseq uently, we unfold this optimizer into a deep learning framework, leading to **De epDistributedQP**, which leverages learned policies to accelerate reaching to de

sired accuracy within a restricted amount of iterations. Our approach is also th
eoretically grounded through Probably Approximately Correct (PAC)-Bayes theory,
providing generalization bounds on the expected optimality gap for unseen proble
ms. The proposed framework, as well as its centralized version **DeepQP**, signi
ficantly outperform their standard optimization counterparts on a variety of tas
ks such as randomly generated problems, optimal control, linear regression, tran
sportation networks and others. Notably, DeepDistributedQP demonstrates strong g
eneralization by training on small problems and scaling to solve much larger one
s (up to 50K variables and 150K constraints) using the same policy. Moreover, it
 achieves orders-of-magnitude improvements in wall-clock time compared to OSQP.
The certifiable performance guarantees of our approach are also demonstrated, en
suring higher-quality solutions over traditional optimizers.
****

Wenxuan Huang,Zijie Zhai,Yunhang Shen,Shaosheng Cao,Fei Zhao,Xiangfeng Xu,Zheyu
Ye,Shaohui Lin

Dynamic-LLaVA: Efficient Multimodal Large Language Models via Dynamic Vision-lan
guage Context Sparsification

Multimodal Large Language Models (MLLMs) have achieved remarkable success in vis
ion understanding, reasoning, and interaction. However, the inference computatio
n and memory increase progressively with the generation of output tokens during
decoding, directly affecting the efficacy of MLLMs. Existing methods attempt to
reduce the vision context redundancy to achieve efficient MLLMs. Unfortunately,
the efficiency benefits of the vision context reduction in the prefill stage gra
dually diminish during the decoding stage. To address this problem, we proposed
a dynamic vision-language context sparsification framework Dynamic-LLaVA, which
dynamically reduces the redundancy of vision context in the prefill stage and de
creases the memory and computation overhead of the generated language context du
ring decoding. Dynamic-LLaVA designs a tailored sparsification inference scheme
for different inference modes, i.e., prefill, decoding with and without KV cache
, to achieve efficient inference of MLLMs. In practice, Dynamic-LLaVA can reduce
 computation consumption by $\sim$75\% in the prefill stage. Meanwhile, througho
ut the entire generation process of MLLMs, Dynamic-LLaVA reduces the $\sim$50\%
computation consumption under decoding without KV cache, while saving $\sim$50\%
 GPU memory overhead when decoding with KV cache, due to the vision-language con
text sparsification. Extensive experiments also demonstrate that Dynamic-LLaVA a
chieves efficient inference for MLLMs with negligible understanding and generati
on ability degradation or even performance gains compared to the full-context in
ference baselines. Code is available at https://github.com/Osilly/dynamic_llava.
****

Runzhe Wu,Ayush Sekhari,Akshay Krishnamurthy,Wen Sun

Computationally Efficient RL under Linear Bellman Completeness for Deterministic
 Dynamics

We study computationally and statistically efficient Reinforcement Learning algo
rithms for the *linear Bellman Complete* setting. This setting uses linear funct
ion approximation to capture value functions and unifies existing models like li
near Markov Decision Processes (MDP) and Linear Quadratic Regulators (LQR). Whi
le it is known from the prior works that this setting is statistically tractable
, it remained open whether a computationally efficient algorithm exists. Our wor
k provides a computationally efficient algorithm for the linear Bellman complete
 setting that works for MDPs with large action spaces, random initial states, an
d random rewards but relies on the underlying dynamics to be deterministic. Our
approach is based on randomization: we inject random noise into least squares re
gression problems to perform optimistic value iteration. Our key technical contr
ibution is to carefully design the noise to only act in the null space of the tr
aining data to ensure optimism while circumventing a subtle error amplification
issue.
****

Cevahir Koprulu,Franck Djeumou,ufuk topcu

Neural Stochastic Differential Equations for Uncertainty-Aware Offline RL

Offline model-based reinforcement learning (RL) offers a principled approach to

using a learned dynamics model as a simulator to optimize a control policy.
Despite the near-optimal performance of existing approaches on benchmarks with h
igh-quality datasets, most struggle on datasets with low state-action space cove
rage or suboptimal demonstrations.
We develop a novel offline model-based RL approach that particularly shines in l
ow-quality data regimes while maintaining competitive performance on high-qualit
y datasets.
Neural Stochastic Differential Equations for Uncertainty-aware, Offline RL (NUNO
) learns a dynamics model as neural stochastic differential equations (SDE),
where its drift term can leverage prior physics knowledge as inductive bias.
In parallel, its diffusion term provides distance-aware estimates of model uncer
tainty by matching the dynamics' underlying stochasticity near the training data
 regime while providing high but bounded estimates beyond it.
To address the so-called model exploitation problem in offline model-based RL, N
UNO builds on existing studies by penalizing and adaptively truncating neural SD
E's rollouts according to uncertainty estimates.
Our empirical results in D4RL and NeoRL MuJoCo benchmarks evidence that NUNO out
performs state-of-the-art methods in low-quality datasets by up to 93% while mat
ching or surpassing their performance by up to 55% in some high-quality counterp
arts.
****
Martin Rohbeck,Edward De Brouwer,Charlotte Bunne,Jan-Christian Huetter,Anne Bito
n,Kelvin Y. Chen,Aviv Regev,Romain Lopez
Modeling Complex System Dynamics with Flow Matching Across Time and Conditions
Modeling the dynamics of complex real-world systems from temporal snapshot data
is crucial for understanding phenomena such as gene regulation, climate change,
and financial market fluctuations. Researchers have recently proposed a few meth
ods based either on the Schroedinger Bridge or Flow Matching to tackle this prob
lem, but these approaches remain limited in their ability to effectively combine
 data from multiple time points and different experimental settings. This integr
ation is essential in real-world scenarios where observations from certain combi
nations of time points and experimental conditions are missing, either because o
f experimental costs or sensory failure. To address this challenge, we propose a
 novel method named Multi-Marginal Flow Matching (MMFM). MMFM first constructs a
 flow using smooth spline-based interpolation across time points and conditions
and regresses it with a neural network using the classifier-free guided Flow Mat
ching framework. This framework allows for the sharing of contextual information
 about the dynamics across multiple trajectories. We demonstrate the effectivene
ss of our method on both synthetic and real-world datasets, including a recent s
ingle-cell genomics data set with around a hundred chemical perturbations across
 time points. Our results show that MMFM significantly outperforms existing meth
ods at imputing data at missing time points.
****
Eshaan Nichani,Jason D. Lee,Alberto Bietti
Understanding Factual Recall in Transformers via Associative Memories
Large language models have demonstrated an impressive ability to perform factual
 recall. Prior work has found that transformers trained on factual recall tasks
can store information at a rate proportional to their parameter count. In our wo
rk, we show that shallow transformers can use a combination of associative memor
ies to obtain such near optimal storage capacity. We begin by proving that the s
torage capacities of both linear and MLP associative memories scale linearly wit
h parameter count. We next introduce a synthetic factual recall task, and prove
that a transformer with a single layer of self-attention followed by an MLP can
obtain 100\% accuracy on the task whenever either the total number of self-atten
tion parameters or MLP parameters scales (up to log factors) linearly with the n
umber of facts. In particular, the transformer can trade off between using the v
alue matrices or the MLP as an associative memory to store the dataset of facts.
 We complement these expressivity results with an analysis of the gradient flow
trajectory of a simplified linear attention model trained on our factual recall
task, where we show that the model exhibits sequential learning behavior.

****
Lichuan Xiang,Quan Nguyen-Tri,Lan-Cuong Nguyen,Hoang Pham,Khoat Than,Long Tran-Thanh,Hongkai Wen
DPaI: Differentiable Pruning at Initialization with Node-Path Balance Principle
Pruning at Initialization (PaI) is a technique in neural network optimization characterized by the proactive elimination of weights before the network's training on designated tasks. This innovative strategy potentially reduces the costs for training and inference, significantly advancing computational efficiency. A key factor leading to PaI's effectiveness is that it considers the saliency of weights in an untrained network, and prioritizes the trainability and optimization potential of the pruned subnetworks. Recent methods can effectively prevent the formation of hard-to-optimize networks, e.g. through iterative adjustments at each network layer. However, this way often results in large-scale discrete optimization problems, which could make PaI further challenging. This paper introduces a novel method, called DPaI, that involves a differentiable optimization of the pruning mask. DPaI adopts a dynamic and adaptable pruning process, allowing easier optimization processes and better solutions. More importantly, our differentiable formulation enables readily use of the existing rich body of efficient gradient-based methods for PaI. Our empirical results demonstrate that DPaI significantly outperforms current state-of-the-art PaI methods on various architectures, such as Convolutional Neural Networks and Vision-Transformers. Code is available at https://github.com/QuanNguyen-Tri/DPaI.git
****
Tristan Thrush,Christopher Potts,Tatsunori Hashimoto
Improving Pretraining Data Using Perplexity Correlations
Quality pretraining data is often seen as the key to high-performance language models. However, progress in understanding pretraining data has been slow due to the costly pretraining runs required for data selection experiments. We present a framework that avoids these costs and selects high-quality pretraining data without any LLM training of our own. Our work is based on a simple observation: LLM losses on many pretraining texts are correlated with downstream benchmark performance, and selecting high-correlation documents is an effective pretraining data selection method. We build a new statistical framework for data selection centered around estimates of perplexity-benchmark correlations and perform data selection using a sample of 90 LLMs taken from the Open LLM Leaderboard on texts from tens of thousands of web domains. In controlled pretraining experiments at the 160M parameter scale on 8 benchmarks, our approach outperforms DSIR on every benchmark, while matching the best data selector found in DataComp-LM, a hand-engineered bigram classifier. We have now also updated this paper to include results from preregistered experiments with new pretraining data on an aggregation of 22 benchmarks up to the 1.4B scale, showing increasing improvements of our method over others with more scale. A pip package with full documentation can be found here: https://github.com/TristanThrush/perplexity-correlations.
****
Qizhou Wang,Jin Peng Zhou,Zhanke Zhou,Saebyeol Shin,Bo Han,Kilian Q Weinberger
Rethinking LLM Unlearning Objectives: A Gradient Perspective and Go Beyond
Large language models (LLMs) should undergo rigorous audits to identify potential risks, such as copyright and privacy infringements. Once these risks emerge, timely updates are crucial to remove undesirable responses, ensuring legal and safe model usage. It has spurred recent research into LLM unlearning, focusing on erasing targeted undesirable knowledge without compromising the integrity of other, non-targeted responses. Existing studies have introduced various unlearning objectives to pursue LLM unlearning without necessitating complete retraining. However, each of these objectives has unique properties, and no unified framework is currently available to comprehend them thoroughly. To fill the gap, we propose the metric of the G-effect, quantifying the impacts of unlearning objectives on model performance from a gradient lens. A significant advantage of our metric is its broad ability to detail the unlearning impacts from various aspects across instances, updating steps, and LLM layers. Accordingly, the G-effect offers new insights into identifying drawbacks of existing unlearning objectives, furthe

r motivating us to explore a series of candidate solutions for their mitigation and improvements. Finally, we outline promising directions that merit further st udies, aiming at contributing to the community to advance this critical field.
****

Md Farhamdur Reza,Richeng Jin,Tianfu Wu,Huaiyu Dai

GSBA^K: top-K Geometric Score-based Black-box Attack

Existing score-based adversarial attacks mainly focus on crafting top-1 advers arial examples against classifiers with single-label classification. Their attac k success rate and query efficiency are often less than satisfactory, particular ly under small perturbation requirements; moreover, the vulnerability of classif iers with multi-label learning is yet to be studied. In this paper, we propose a comprehensive surrogate free score-based attack, named \b geometric \b score-ba sed \b black-box \b attack (GSBA^K), to craft adversarial examples in an aggre ssive top-K setting for both untargeted and targeted attacks, where the goal is to change the top-K predictions of the target classifier. We introduce n ovel gradient-based methods to find a good initial boundary point to attack. Our iterative method employs novel gradient estimation techniques, particularly eff ective in top-K setting, on the decision boundary to effectively exploit the geometry of the decision boundary. Additionally, GSBA^K can be used to attack against classifiers with top-K multi-label learning. Extensive experiential results on ImageNet and PASCAL VOC datasets validate the effectiveness of GSBA$ ^K$ in crafting top-K adversarial examples.
****

Xiaolei Wang,Xinyu Tang,Junyi Li,Xin Zhao,Ji-Rong Wen

Investigating the Pre-Training Dynamics of In-Context Learning: Task Recognition vs. Task Learning

The emergence of in-context learning (ICL) is potentially attributed to two majo r abilities: task recognition (TR) for recognizing the task from demonstrations and utilizing pre-trained priors, and task learning (TL) for learning from demon strations. However, relationships between the two abilities and how such relatio nships affect the emergence of ICL is unclear. In this paper, we take the first step by examining the pre-training dynamics of the emergence of ICL. With carefu lly designed metrics, we find that these two abilities are, in fact, competitive during pre-training. Moreover, we observe a negative correlation between the co mpetition and the performance of ICL. Further analysis of common pre-training fa ctors (i.e., model size, dataset size, and data curriculum) demonstrates possibl e ways to regulate the competition. Based on these insights, we propose a simple yet effective method to better integrate these two abilities for ICL at inferen ce time. Through adaptive ensemble learning, the performance of ICL can be signi ficantly boosted, enabling two small models to outperform a larger one with more than twice the parameters.
****

Patrik Reizinger,Alice Bizeul,Attila Juhos,Julia E Vogt,Randall Balestriero,Wiel and Brendel,David Klindt

Cross-Entropy Is All You Need To Invert the Data Generating Process

Supervised learning has become a cornerstone of modern machine learning, yet a c omprehensive theory explaining its effectiveness remains elusive. Empirical phen omena, such as neural analogy-making and the linear representation hypothesis, s uggest that supervised models can learn interpretable factors of variation in a linear fashion. Recent advances in self-supervised learning, particularly nonlin ear Independent Component Analysis, have shown that these methods can recover la tent structures by inverting the data generating process. We extend these identi fiability results to parametric instance discrimination,
then show how insights transfer to the ubiquitous setting of supervised learning with cross-entropy minimization. We prove that even in standard classification tasks, models learn representations of ground-truth factors of variation up to a linear transformation under a certain DGP. We corroborate our theoretical contr ibution with a series of empirical studies. First, using simulated data matching our theoretical assumptions, we demonstrate successful disentanglement of laten t factors. Second, we show that on DisLib, a widely-used disentanglement benchma

rk, simple classification tasks recover latent structures up to linear transform
ations. Finally, we reveal that models trained on ImageNet encode representation
s that permit linear decoding of proxy factors of variation.
Together, our theoretical findings and experiments offer a compelling explanatio
n for recent observations of linear representations, such as superposition in ne
ural networks. This work takes a significant step toward a cohesive theory that
accounts for the unreasonable effectiveness of supervised learning.
****
Shane Bergsma,Nolan Simran Dey,Gurpreet Gosal,Gavia Gray,Daria Soboleva,Joel Hes
tness
Straight to Zero: Why Linearly Decaying the Learning Rate to Zero Works Best for
 LLMs
LLMs are commonly trained with a learning rate (LR) warmup, followed by cosine d
ecay to 10% of the maximum (10x decay). In a large-scale empirical study, we sho
w that under an optimal peak LR, a simple linear decay-to-zero (D2Z) schedule co
nsistently outperforms other schedules when training at compute-optimal dataset
sizes. D2Z is superior across a range of model sizes, batch sizes, datasets, and
 vocabularies. Benefits increase as dataset size increases. Leveraging a novel i
nterpretation of AdamW as an exponential moving average of weight updates, we sh
ow how linear D2Z optimally balances the demands of early training (moving away
from initial conditions) and late training (averaging over more updates in order
 to mitigate gradient noise). In experiments, a 610M-parameter model trained for
 80 tokens-per-parameter (TPP) using D2Z achieves lower loss than when trained f
or 200 TPP using 10x decay, corresponding to an astonishing 60% compute savings.
 Models such as Llama2-7B, trained for 286 TPP with 10x decay, could likely have
 saved a majority of compute by training with D2Z.
****
Zhihao Zhu,Yi Yang,Defu Lian
TDDBench: A Benchmark for Training data detection
Training Data Detection (TDD) is a task aimed at determining whether a specific
data instance is used to train a machine learning model. In the computer securi
ty literature, TDD is also referred to as Membership Inference Attack (MIA). Giv
en its potential to assess the risks of training data breaches, ensure copyright
 authentication, and verify model unlearning, TDD has garnered significant atten
tion in recent years, leading to the development of numerous methods. Despite th
ese advancements, there is no comprehensive benchmark to thoroughly evaluate the
 effectiveness of TDD methods.
In this work, we introduce TDDBench, which consists of 13 datasets spanning thre
e data modalities: image, tabular, and text. We benchmark 21 different TDD metho
ds across four detection paradigms and evaluate their performance from five pers
pectives: average detection performance, best detection performance, memory cons
umption, and computational efficiency in both time and memory. With TDDBench, re
searchers can identify bottlenecks and areas for improvement in TDD algorithms,
while practitioners can make informed trade-offs between effectiveness and effic
iency when selecting TDD algorithms for specific use cases. Our extensive experi
ments also reveal the generally unsatisfactory performance of TDD algorithms acr
oss different datasets. To enhance accessibility and reproducibility, we open-so
urce TDDBench for the research community at https://github.com/zzh9568/TDDBench.
****
Jinjie Ni,Yifan Song,Deepanway Ghosal,Bo Li,David Junhao Zhang,Xiang Yue,Fuzhao
Xue,Yuntian Deng,Zian Zheng,Kaichen Zhang,Mahir Shah,Kabir Jain,Yang You,Michael
 Shieh
MixEval-X: Any-to-any Evaluations from Real-world Data Mixture
Perceiving and generating diverse modalities are crucial for AI models to effect
ively learn from and engage with real-world signals, necessitating reliable eval
uations for their development. We identify two major issues in current evaluatio
ns: (1) inconsistent standards, shaped by different communities with varying pro
tocols and maturity levels; and (2) significant query, grading, and generalizati
on biases. To address these, we introduce MixEval-X, the first any-to-any, real-
world benchmark designed to optimize and standardize evaluations across diverse

input and output modalities. We propose multi-modal benchmark mixture and adapta
tion-rectification pipelines to reconstruct real-world task distributions, ensur
ing evaluations generalize effectively to real-world use cases. Extensive meta-e
valuations show our approach effectively aligns benchmark samples with real-worl
d task distributions. Meanwhile, MixEval-X's model rankings correlate strongly w
ith that of crowd-sourced real-world evaluations (up to 0.98) while being much m
ore efficient. We provide comprehensive leaderboards to rerank existing models a
nd organizations and offer insights to enhance understanding of multi-modal eval
uations and inform future research.
****
Zhikai Chen,Han Xie,Jian Zhang,Xiang song,Jiliang Tang,Huzefa Rangwala,George Ka
rypis
AutoG: Towards automatic graph construction from tabular data
Recent years have witnessed significant advancements in graph machine learning (
GML), with its applications spanning numerous domains. However, the focus of GML
 has predominantly been on developing powerful models, often overlooking a cruci
al initial step: constructing suitable graphs from common data formats, such as
tabular data.
This construction process is fundamental to applying graph-based models, yet it
remains largely understudied and lacks formalization.
Our research aims to address this gap by formalizing the graph construction prob
lem and proposing an effective solution. We identify two critical challenges to
achieve this goal: 1. The absence of dedicated datasets to formalize and evaluat
e the effectiveness of graph construction methods, and 2. Existing automatic con
struction methods can only be applied to some specific cases, while tedious huma
n engineering is required to generate high-quality graphs.
To tackle these challenges, we present a two-fold contribution.
First, we introduce a set of datasets to formalize and evaluate graph constructi
on methods.
Second, we propose an LLM-based solution, AutoG, automatically generating high-q
uality graph schemas without human intervention.
The experimental results demonstrate that the quality of constructed graphs is c
ritical to downstream task performance, and AutoG can generate high-quality grap
hs that rival those produced by human experts. Our code can be accessible from h
ttps://github.com/amazon-science/Automatic-Table-to-Graph-Generation.
****
Yutong Wang,Jiali Zeng,Xuebo Liu,Derek F. Wong,Fandong Meng,Jie Zhou,Min Zhang
DelTA: An Online Document-Level Translation Agent Based on Multi-Level Memory
Large language models (LLMs) have achieved reasonable quality improvements in ma
chine translation (MT).
However, most current research on MT-LLMs still faces significant challenges in
maintaining translation consistency and accuracy when processing entire document
s.
In this paper, we introduce DelTA, a Document-levEL Translation Agent designed t
o overcome these limitations.
DelTA features a multi-level memory structure that stores information across var
ious granularities and spans, including Proper Noun Records, Bilingual Summary,
Long-Term Memory, and Short-Term Memory, which are continuously retrieved and up
dated by auxiliary LLM-based components.
Experimental results indicate that DelTA significantly outperforms strong baseli
nes in terms of translation consistency and quality across four open/closed-sour
ce LLMs and two representative document translation datasets, achieving an incre
ase in consistency scores by up to 4.58 percentage points and in COMET scores by
 up to 3.16 points on average.
DelTA employs a sentence-by-sentence translation strategy, ensuring no sentence
omissions and offering a memory-efficient solution compared to the mainstream me
thod.
Furthermore, DelTA improves pronoun and context-dependent translation accuracy,
and the summary component of the agent also shows promise as a tool for query-ba
sed summarization tasks.

The code and data of our approach are released at https://github.com/YutongWang1216/DocMTAgent.
****
Mengxuan Hu,Zihan Guan,Yi Zeng,Junfeng Guo,Zhongliang Zhou,Jielu Zhang,Ruoxi Jia,Anil Kumar Vullikanti,Sheng Li
Mind Control through Causal Inference: Predicting Clean Images from Poisoned Data
Anti-backdoor learning, aiming to train clean models directly from poisoned datasets, serves as an important defense method for backdoor attack. However, existing methods usually fail to recover backdoored samples to their original, correct labels and suffer from poor generalization to large pre-trained models due to its non end-to end training, making them unsuitable for protecting the increasingly prevalent large pre-trained models. To bridge the gap, we first revisit the anti-backdoor learning problem from a causal perspective. Our theoretical causal analysis reveals that incorporating \emph{\textbf{both}} images and the associated attack indicators preserves the model's integrity. Building on the theoretical analysis, we introduce an end-to-end method, Mind Control through Causal Inference (MCCI), to train clean models directly from poisoned datasets. This approach leverages both the image and the attack indicator to train the model. Based on this training paradigm, the model’s perception of whether an input is clean or backdoored can be controlled. Typically, by introducing fake non-attack indicators, the model perceives all inputs as clean and makes correct predictions, even for poisoned samples. Extensive experiments demonstrate that our method achieves state-of-the-art performance, efficiently recovering the original correct predictions for poisoned samples and enhancing accuracy on clean samples.
****
Baixiang Huang,Canyu Chen,Xiongxiao Xu,Ali Payani,Kai Shu
Can Knowledge Editing Really Correct Hallucinations?
Large Language Models (LLMs) suffer from hallucinations, referring to the non-factual information in generated content, despite their superior capacities across tasks. Meanwhile, knowledge editing has been developed as a new popular paradigm to correct erroneous factual knowledge encoded in LLMs with the advantage of avoiding retraining from scratch. However, a common issue of existing evaluation datasets for knowledge editing is that they do not ensure that LLMs actually generate hallucinated answers to the evaluation questions before editing. When LLMs are evaluated on such datasets after being edited by different techniques, it is hard to directly adopt the performance to assess the effectiveness of different knowledge editing methods in correcting hallucinations. Thus, the fundamental question remains insufficiently validated: Can knowledge editing really correct hallucinations in LLMs? We proposed HalluEditBench to holistically benchmark knowledge editing methods in correcting real-world hallucinations. First, we rigorously construct a massive hallucination dataset with 9 domains, 26 topics and more than 6,000 hallucinations. Then, we assess the performance of knowledge editing methods in a holistic way on five dimensions including Efficacy, Generalization, Portability, Locality, and Robustness. Through HalluEditBench, we have provided new insights into the potentials and limitations of different knowledge editing methods in correcting hallucinations, which could inspire future improvements and facilitate progress in the field of knowledge editing.
****
Tobias Gessler,Tin Dizdarevic,Ani Calinescu,Benjamin Ellis,Andrei Lupu,Jakob Nicolaus Foerster
OvercookedV2: Rethinking Overcooked for Zero-Shot Coordination
AI agents hold the potential to transform everyday life by helping humans achieve their goals.
To do this successfully, agents need to be able to coordinate with novel partners without prior interaction, a setting known as zero-shot coordination (ZSC).
Overcooked has become one of the most popular benchmarks for evaluating coordination capabilities of AI agents and learning algorithms.
In this work, we investigate the origins of ZSC challenges in Overcooked.
We introduce a state augmentation mechanism which mixes states that might be enc

ountered when paired with unknown partners into the training distribution, reduc
ing the out-of-distribution challenge associated with ZSC.
We show that independently trained agents under this algorithm coordinate succes
sfully in Overcooked.
Our results suggest that ZSC failure can largely be attributed to poor state cov
erage under self-play rather than more sophisticated coordination challenges. Th
e Overcooked environment is therefore not suitable as a ZSC benchmark.
To address these shortcomings, we introduce OvercookedV2, a new version of the b
enchmark, which includes asymmetric information and stochasticity, facilitating
the creation of interesting ZSC scenarios.
To validate OvercookedV2, we conduct experiments demonstrating that mere exhaust
ive state coverage is insufficient to coordinate well. Finally, we use Overcooke
dV2 to build a new range of coordination challenges, including ones that require
 test time protocol formation, and we demonstrate the need for new coordination
algorithms that can adapt online.
We hope that OvercookedV2 will help benchmark the next generation of ZSC algorit
hms and advance collaboration between AI agents and humans.
****
Fredrik Nestaas,Edoardo Debenedetti,Florian Tramèr
Adversarial Search Engine Optimization for Large Language Models
Large Language Models (LLMs) are increasingly used in applications where the mod
el selects from competing third-party content, such as in LLM-powered search eng
ines or chatbot plugins.
In this paper, we introduce _Preference Manipulation Attacks_, a new class of at
tacks that manipulate an LLM's selections to favor the attacker. We demonstrate
that carefully crafted website content or plugin documentations can trick an LLM
 to promote the attacker products and discredit competitors, thereby increasing
user traffic and monetization (a form of adversarial Search Engine Optimization)
.
We show this can lead to a _prisoner's dilemma_, where all parties are incentivi
zed to launch attacks, but this collectively degrades the LLM's outputs for eve
ryone. We demonstrate our attacks on production LLM search engines (Bing and Per
plexity) and plugin APIs (for GPT-4 and Claude). As LLMs are increasingly used t
o rank third-party content, we expect Preference Manipulation Attacks to emerge
as a significant threat.
****
Yuewen Sun,Lingjing Kong,Guangyi Chen,Loka Li,Gongxu Luo,Zijian Li,Yixuan Zhang,
Yujia Zheng,Mengyue Yang,Petar Stojanov,Eran Segal,Eric P. Xing,Kun Zhang
Causal Representation Learning from Multimodal Biomedical Observations
Prevalent in biomedical applications (e.g., human phenotype research), multimoda
l datasets can provide valuable insights into the underlying physiological mecha
nisms. However, current machine learning (ML) models designed to analyze these d
atasets often lack interpretability and identifiability guarantees, which are es
sential for biomedical research. Recent advances in causal representation learni
ng have shown promise in identifying interpretable latent causal variables with
formal theoretical guarantees. Unfortunately, most current work on multimodal di
stributions either relies on restrictive parametric assumptions or yields only c
oarse identification results, limiting their applicability to biomedical researc
h that favors a detailed understanding of the mechanisms.

In this work, we aim to develop flexible identification conditions for multimoda
l data and principled methods to facilitate the understanding of biomedical data
sets. Theoretically, we consider a nonparametric latent distribution (c.f., para
metric assumptions in previous work) that allows for causal relationships across
 potentially different modalities. We establish identifiability guarantees for e
ach latent component, extending the subspace identification results from previou
s work. Our key theoretical contribution is the structural sparsity of causal co
nnections between modalities, which, as we will discuss, is natural for a large
collection of biomedical systems.
Empirically, we present a practical framework to instantiate our theoretical ins

ights. We demonstrate the effectiveness of our approach through extensive experi
ments on both numerical and synthetic datasets. Results on a real-world human ph
enotype dataset are consistent with established biomedical research, validating
our theoretical and methodological framework.
****

Hao Dong,Eleni Chatzi,Olga Fink

Towards Robust Multimodal Open-set Test-time Adaptation via Adaptive Entropy-awa
re Optimization

Test-time adaptation (TTA) has demonstrated significant potential in addressing
distribution shifts between training and testing data. Open-set test-time adapta
tion (OSTTA) aims to adapt a source pre-trained model online to an unlabeled tar
get domain that contains unknown classes. This task becomes more challenging whe
n multiple modalities are involved. Existing methods have primarily focused on u
nimodal OSTTA, often filtering out low-confidence samples without addressing the
 complexities of multimodal data. In this work, we present Adaptive Entropy-awar
e Optimization (AEO), a novel framework specifically designed to tackle Multimod
al Open-set Test-time Adaptation (MM-OSTTA) for the first time. Our analysis sho
ws that the entropy difference between known and unknown samples in the target d
omain strongly correlates with MM-OSTTA performance. To leverage this, we propos
e two key components: Unknown-aware Adaptive Entropy Optimization (UAE) and Adap
tive Modality Prediction Discrepancy Optimization (AMP). These components enhanc
e the model’s ability to distinguish unknown class samples during online adaptat
ion by amplifying the entropy difference between known and unknown samples. To t
horoughly evaluate our proposed methods in the MM-OSTTA setting, we establish a
new benchmark derived from existing datasets. This benchmark includes two downst
ream tasks – action recognition and 3D semantic segmentation – and incorporates
five modalities: video, audio, and optical flow for action recognition, as well
as LiDAR and camera for 3D semantic segmentation. Extensive experiments across v
arious domain shift situations demonstrate the efficacy and versatility of the A
EO framework. Additionally, we highlight the strong performance of AEO in long-t
erm and continual MM-OSTTA settings, both of which are challenging and highly re
levant to real-world applications. This underscores AEO’s robustness and adaptab
ility in dynamic environments. Our source code and benchmarks are available at h
ttps://github.com/donghao51/AEO.
****

Wei Huang,Yue Liao,Jianhui Liu,Ruifei He,Haoru Tan,Shiming Zhang,Hongsheng Li,Si
 Liu,XIAOJUAN QI

Mixture Compressor for Mixture-of-Experts LLMs Gains More

Mixture-of-Experts large language models (MoE-LLMs) marks a significant step for
ward of language models, however, they encounter two critical challenges in prac
tice: 1) expert parameters lead to considerable memory consumption and loading l
atency; and 2) the current activated experts are redundant, as many tokens may o
nly require a single expert. Motivated by these issues, we investigate the MoE-L
LMs and make two key observations: a) different experts exhibit varying behavior
s on activation reconstruction error, routing scores, and activated frequencies,
 highlighting their differing importance, and b) not all tokens are equally impo
rtant-- only a small subset is critical. Building on these insights, we propose
MC, a training-free Mixture-Compressor for MoE-LLMs, which leverages the signifi
cance of both experts and tokens to achieve an extreme compression. First, to mi
tigate storage and loading overheads, we introduce Pre-Loading Mixed-Precision Q
uantization (PMQ), which formulates the adaptive bit-width allocation as a Linea
r Programming (LP) problem, where the objective function balances multi-factors
reflecting the importance of each expert. Additionally, we develop Online Dynami
c Pruning (ODP), which identifies important tokens to retain and dynamically se
lect activated experts for other tokens during inference to optimize efficiency
while maintaining performance. Our MC integrates static quantization and dynamic
 pruning to collaboratively achieve extreme compression for MoE-LLMs with less a
ccuracy loss, ensuring an optimal trade-off between performance and efficiency E
xtensive experiments confirm the effectiveness of our approach. For instance, at
 2.54 bits, MC compresses 76.6% of the model, with only a 3.8% average accuracy

loss. During dynamic inference, we further reduce activated parameters by 15%, w
ith a performance drop of less than 0.6%. Remarkably, MC even surpasses floating
-point 13b dense LLMs with significantly smaller parameter sizes, suggesting tha
t mixture compression in MoE-LLMs has the potential to outperform both comparabl
e and larger dense LLMs. Our code is
available at https://github.com/Aaronhuang-778/MC-MoE
****
Anthony GX-Chen,Kenneth Marino,Rob Fergus
Efficient Exploration and Discriminative World Model Learning with an Object-Cen
tric Abstraction
In the face of difficult exploration problems in reinforcement learning, we stud
y whether giving an agent an object-centric mapping (describing a set of items a
nd their attributes) allow for more efficient learning. We found this problem is
 best solved hierarchically by modelling items at a higher level of state abstra
ction to pixels, and attribute change at a higher level of temporal abstraction
to primitive actions. This abstraction simplifies the transition dynamic by maki
ng specific future states easier to predict. We make use of this to propose a fu
lly model-based algorithm that learns a discriminative world model, plans to exp
lore efficiently with only a count-based intrinsic reward, and can subsequently
plan to reach any discovered (abstract) states.

We demonstrate the model's ability to (i) efficiently solve single tasks, (ii) t
ransfer zero-shot and few-shot across item types and environments, and (iii) pla
n across long horizons. Across a suite of 2D crafting and MiniHack environments,
 we empirically show our model significantly out-performs state-of-the-art low-l
evel methods (without abstraction), as well as performant model-free and model-b
ased methods using the same abstraction. Finally, we show how to learn low level
 object-perturbing policies via reinforcement learning, and the object mapping i
tself by supervised learning.
****
Jiangxing Wang,Zongqing Lu
Watch Less, Do More: Implicit Skill Discovery for Video-Conditioned Policy
In this paper, we study the problem of video-conditioned policy learning. While
previous works mostly focus on learning policies that perform a single skill spe
cified by the given video, we take a step further and aim to learn a policy that
 can perform multiple skills according to the given video, and generalize to uns
een videos by recombining these skills. To solve this problem, we propose our al
gorithm, Watch-Less-Do-More, an information bottleneck-based imitation learning
framework for implicit skill discovery and video-conditioned policy learning. In
 our method, an information bottleneck objective is employed to control the info
rmation contained in the video representation, ensuring that it only encodes inf
ormation relevant to the current skill (Watch-Less). By discovering potential sk
ills from training videos, the learned policy is able to recombine them and gene
ralize to unseen videos to achieve compositional generalization (Do-More). To ev
aluate our method, we perform extensive experiments in various environments and
show that our algorithm substantially outperforms baselines (up to 2x) in terms
of compositional generalization ability.
****
Jaehong Yoon,Shoubin Yu,Vaidehi Patil,Huaxiu Yao,Mohit Bansal
SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Genera
tion
Recent advances in diffusion models have significantly enhanced their ability to
 generate high-quality images and videos, but they have also increased the risk
of producing unsafe content. Existing unlearning/editing-based methods for safe
generation remove harmful concepts from models but face several challenges: (1)
They cannot instantly remove harmful or undesirable concepts (e.g., artist style
s) without additional training. (2) Their safe generation capabilities depend on
 collected training data. (3) They alter model weights, risking degradation in q
uality for content unrelated to the targeted toxic concepts. To address these ch
allenges, we propose SAFREE, a novel, training-free approach for safe text-to-im

age and video generation, that does not alter the model's weights. Specifically, we detect a subspace corresponding to a set of toxic concepts in the text embedding space and steer prompt token embeddings away from this subspace, thereby filtering out harmful content while preserving intended semantics. To balance the trade-off between filtering toxicity and preserving safe concepts, SAFREE incorporates a novel self-validating filtering mechanism that dynamically adjusts the denoising steps when applying the filtered embeddings. Additionally, we incorporate adaptive re-attention mechanisms within the diffusion latent space to selectively diminish the influence of features related to toxic concepts at the pixel level. By integrating filtering across both textual embedding and visual latent spaces, SAFREE ensures coherent safety checking, preserving the fidelity, quality, and safety of the generated outputs. Empirically, SAFREE achieves state-of-the-art performance in suppressing unsafe content in T2I generation (reducing it by 22% across 5 datasets) compared to other training-free methods and effectively filters targeted concepts, e.g., specific artist styles, while maintaining high-quality output. It also shows competitive results against training-based methods. We further extend SAFREE to various T2I backbones and T2V tasks, showcasing its flexibility and generalization. As generative AI rapidly evolves, SAFREE provides a robust and adaptable safeguard for ensuring safe visual generation.
****
Fiorenzo Parascandolo,Nicholas Moratelli,Enver Sangineto,Lorenzo Baraldi,Rita Cucchiara
Causal Graphical Models for Vision-Language Compositional Understanding
Recent work has empirically shown that Vision-Language Models (VLMs) struggle
to fully understand the compositional properties of the human language, usually
modeling an image caption as a “bag of words”. As a result, they perform
poorly on compositional tasks, which require a deeper understanding of the different
entities of a sentence (subject, verb, etc.) jointly with their mutual relationships
in order to be solved. In this paper, we model the dependency relations
among textual and visual tokens using a Causal Graphical Model (CGM), built using
a dependency parser, and we train a decoder conditioned by the VLM visual
encoder. Differently from standard autoregressive or parallel predictions, our decoder’s
generative process is partially-ordered following the CGM structure. This
structure encourages the decoder to learn only the main causal dependencies in
a sentence discarding spurious correlations. Using extensive experiments on five
compositional benchmarks, we show that our method significantly outperforms
all the state-of-the-art compositional approaches by a large margin, and it also improves
over methods trained using much larger datasets.
Our model weights and code are publicly available.
****
Audrey Huang,Wenhao Zhan,Tengyang Xie,Jason D. Lee,Wen Sun,Akshay Krishnamurthy,Dylan J Foster
Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization
Language model alignment methods such as reinforcement learning from human feedback (RLHF) have led to impressive advances in language model capabilities, but are limited by a widely observed phenomenon known as *overoptimization*, where the quality of the language model degrades over the course of the alignment process. As the model optimizes performance on an offline reward model, it overfits to inaccuracies and drifts away from preferred responses covered by the data. To discourage such distribution shift, KL-regularization is widely employed in existing offline alignment methods, but overoptimization continues to harm performance. Lending theoretical insight into the source of these empirical observations, we first show that the KL-regularization is too weak to prevent overfitting, then ask: is it possible to design an efficient algorithm that is provably robust t

o overoptimization?

In this paper, we advance theoretical understanding of sample-efficient offline alignment and introduce a new algorithm called χ^2-Preference Optimization (χPO). χPO is a one-line change to Direct Preference Optimization (DPO ; Rafailov et al. 2023), that modifies only the logarithmic link function in the DPO objective. Despite this minimal change, χPO implicitly implements the principle of *pessimism in the face of uncertainty* via regularization with the χ^2-divergence---which quantifies uncertainty more effectively than KL-regu larization---and provably alleviates overoptimization, achieving sample-complexi ty guarantees based on *single-policy concentrability*, the gold standard in off line reinforcement learning. This guarantee makes χPO the first simple, yet general-purpose offline alignment algorithm that is provably robust to overopti mization.
****
Maksym Andriushchenko,Francesco Croce,Nicolas Flammarion
Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks
We show that even the most recent safety-aligned LLMs are not robust to simple * adaptive* jailbreaking attacks. First, we demonstrate how to successfully levera ge access to *logprobs* for jailbreaking: we initially design an adversarial pro mpt template (sometimes adapted to the target LLM), and then we apply random sea rch on a suffix to maximize a target logprob (e.g., of the token *``Sure''*), po tentially with multiple restarts. In this way, we achieve 100\% attack success r ate---according to GPT-4 as a judge---on Vicuna-13B, Mistral-7B, Phi-3-Mini, Nem otron-4-340B, Llama-2-Chat-7B/13B/70B, Llama-3-Instruct-8B, Gemma-7B, GPT-3.5, G PT-4o, and R2D2 from HarmBench that was adversarially trained against the GCG at tack. We also show how to jailbreak *all* Claude models---that do not expose log probs---via either a transfer or prefilling attack with a *100\% success rate*. In addition, we show how to use random search on a restricted set of tokens for finding trojan strings in poisoned models---a task that shares many similarities with jailbreaking---which is the algorithm that brought us the *first place* in a recent trojan detection competition. The common theme behind these attacks is that *adaptivity* is crucial: different models are vulnerable to different prom pting templates (e.g., R2D2 is very sensitive to in-context learning prompts), s ome models have unique vulnerabilities based on their APIs (e.g., prefilling for Claude), and in some settings, it is crucial to restrict the token search space based on prior knowledge (e.g., for trojan detection). For reproducibility purp oses, we provide the code, logs, and jailbreak artifacts in the JailbreakBench f ormat at https://github.com/tml-epfl/llm-adaptive-attacks.
****
Song Li,Yang Tan,Song Ke,Liang Hong,Bingxin Zhou
Immunogenicity Prediction with Dual Attention Enables Vaccine Target Selection
Immunogenicity prediction is a central topic in reverse vaccinology for finding candidate vaccines that can trigger protective immune responses. Existing approa ches typically rely on highly compressed features and simple model architectures , leading to limited prediction accuracy and poor generalizability. To address t hese challenges, we introduce VenusVaccine, a novel deep learning solution with a dual attention mechanism that integrates pre-trained latent vector representat ions of protein sequences and structures. We also compile the most comprehensive immunogenicity dataset to date, encompassing over 7000 antigen sequences, struc tures, and immunogenicity labels from bacteria, viruses, and tumors. Extensive e xperiments demonstrate that VenusVaccine outperforms existing methods across a w ide range of evaluation metrics. Furthermore, we establish a post-hoc validation protocol to assess the practical significance of deep learning models in tackli ng vaccine design challenges. Our work provides an effective tool for vaccine de sign and sets valuable benchmarks for future research. The implementation is at \url{https://github.com/songleee/VenusVaccine}.
****
Matthew Joseph,Mónica Ribero,Alexander Yu
Privately Counting Partially Ordered Data

We consider differentially private counting when each data point consists of d bits satisfying a partial order. Our main technical contribution is a problem-specific K-norm mechanism that runs in time $O(d^2)$. Experiments show that, depending on the partial order in question, our solution dominates existing pure differentially private mechanisms and can reduce their error by an order of magnitude or more.
****

Saurav Jha,Shiqi Yang,Masato Ishii,Mengjie Zhao,christian simon,Muhammad Jehanzeb Mirza,Dong Gong,Lina Yao,Shusuke Takahashi,Yuki Mitsufuji
Mining your own secrets: Diffusion Classifier Scores for Continual Personalization of Text-to-Image Diffusion Models
Personalized text-to-image diffusion models have grown popular for their ability to efficiently acquire a new concept from user-defined text descriptions and a few images. However, in the real world, a user may wish to personalize a model on multiple concepts but one at a time, with no access to the data from previous concepts due to storage/privacy concerns. When faced with this continual learning (CL) setup, most personalization methods fail to find a balance between acquiring new concepts and retaining previous ones -- a challenge that *continual personalization* (CP) aims to solve.
Inspired by the successful CL methods that rely on class-specific information for regularization, we resort to the inherent class-conditioned density estimates, also known as *diffusion classifier* (DC) scores, for CP of text-to-image diffusion models.
Namely, we propose using DC scores for regularizing the parameter-space and function-space of text-to-image diffusion models.
Using several diverse evaluation setups, datasets, and metrics, we show that our proposed regularization-based CP methods outperform the state-of-the-art C-LoRA, and other baselines. Finally, by operating in the replay-free CL setup and on low-rank adapters, our method incurs zero storage and parameter overhead, respectively, over the state-of-the-art.
****

Qi Liu,Xinhao Zheng,Xudong Lu,Qinxiang Cao,Junchi Yan
Rethinking and Improving Autoformalization: Towards a Faithful Metric and a Dependency Retrieval-based Approach
As a central component in formal verification, statement autoformalization has been widely studied including the recent efforts from machine learning community, but still remains a widely-recognized difficult and open problem. In this paper, we delve into two critical yet under-explored gaps: 1) absence of faithful and universal automated evaluation for autoformalization results; 2) agnosia of contextual information, inducing severe hallucination of formal definitions and theorems. To address the first issue, we propose **BEq** (_**B**idirectional **E**xtended Definitional E**q**uivalence_), an automated neuro-symbolic method to determine the equivalence between two formal statements, which is formal-grounded and well-aligned with human intuition. For the second, we propose **RAutoformalizer** (_**R**etrieval-augmented **Autoformalizer**_), augmenting statement autoformalization by _Dependency Retrieval_, retrieving potentially dependent objects from formal libraries. We parse the dependencies of libraries and propose to _structurally informalise_ formal objects by the topological order of dependencies. To evaluate OOD generalization and research-level capabilities, we build a novel benchmark, _Con-NF_, consisting of 961 informal-formal statement pairs from frontier mathematical researches. Experiments validate the effectiveness of our approaches: BEq is evaluated on 200 diverse formal statement pairs with expert-annotated equivalence label, exhibiting significantly improved accuracy ($82.50\\% \mapsto 90.50\\%$) and precision ($70.59\\% \mapsto 100.0\\%$). For dependency retrieval, a strong baseline is devised. Our RAutoformalizer substantially outperforms SOTA baselines in both in-distribution ProofNet benchmark ($12.83\\% \mapsto 18.18\\%$, BEq@8) and OOD Con-NF scenario ($4.58\\%\mapsto 16.86\\%$, BEq@8).
****

Ruochen Wang,Si Si,Felix Yu,Dorothea Wiesmann Rothuizen,Cho-Jui Hsieh,Inderjit S Dhillon

Large Language Models are Interpretable Learners
The trade-off between expressiveness and interpretability remains a core challen ge when building human-centric models for classification and decision-making. Wh ile symbolic rules offer interpretability, they often lack expressiveness, where as neural networks excel in performance but are known for being black boxes. Thi s paper shows a combination of Large Language Models (LLMs) and symbolic program s can bridge this gap. In the proposed LLM-based Symbolic Programs (LSPs), the p retrained LLM with natural language prompts provides a massive set of interpreta ble modules that can transform raw input into natural language concepts. Symboli c programs then integrate these modules into interpretable decision rules. To tr ain LSPs, we develop a divide-and-conquer approach to incrementally build the pr ogram from scratch, where the learning process of each step is guided by LLMs. T o evaluate the effectiveness of LSPs in extracting interpretable and accurate kn owledge from data, we introduce IL-Bench, a collection of diverse tasks, includi ng both synthetic and real-world scenarios across different modalities. Empirica l results demonstrate LSP's superior performance compared to traditional neurosy mbolic programs and vanilla automatic prompt tuning methods. Moreover, as the kn owledge learned by LSP is a combination of natural language descriptions and sym bolic rules, it is easily transferable to humans (interpretable), and other LLMs , and generalizes well to out-of-distribution samples. Our code and benchmark wi ll be released for future research.
****
Daniel Cai,Randall Balestriero
No Location Left Behind: Measuring and Improving the Fairness of Implicit Repres entations for Earth Data
Implicit neural representations (INRs) exhibit growing promise in addressing Ear th representation challenges, ranging from emissions monitoring to climate model ing. However, existing methods disproportionately prioritize global average perf ormance, whereas practitioners require fine-grained insights to understand biase s and variations in these models. To bridge this gap, we introduce FAIR-Earth: a first-of-its-kind dataset explicitly crafted to challenge and examine inequitie s in Earth representations. FAIR-Earth comprises various high-resolution Earth s ignals, and uniquely aggregates extensive metadata along stratifications like la ndmass size and population density to assess the fairness of models. Evaluating state-of-the-art INRs across the various modalities of FAIR-Earth, we uncover st riking performance disparities. Certain subgroups, especially those associated w ith high-frequency signals (e.g., islands, coastlines), are consistently poorly modeled by existing methods. In response, we propose spherical wavelet encodings , building on previous spatial encoding research for INRs. Leveraging the multi-resolution analysis capabilities of wavelets, our encodings yield more consisten t performance over various scales and locations, offering more accurate and robu st representations of the biased subgroups. These open-source contributions repr esent a crucial step towards facilitating the equitable assessment and deploymen t of implicit Earth representations.
****
Yuxin Jiang,Bo Huang,Yufei Wang,Xingshan Zeng,Liangyou Li,Yasheng Wang,Xin Jiang ,Lifeng Shang,Ruiming Tang,Wei Wang
Bridging and Modeling Correlations in Pairwise Data for Direct Preference Optimi zation
Direct preference optimization (DPO), a widely adopted offline preference optimi zation algorithm, aims to align large language models (LLMs) with human-desired behaviors using pairwise preference data. However, the generation of the winning response and the losing response within pairwise data are typically isolated, l eading to weak correlations between them as well as suboptimal alignment perform ance. To address this issue, we propose an effective framework for Bridging and Modeling Correlations in pairwise data, named BMC. Firstly, we increase the cons istency and informativeness of the pairwise preference signals through targeted modifications, synthesizing a pseudo-winning response by improving the losing re sponse with the winning response as a reference. Secondly, we identify that DPO alone is insufficient to model these correlations and capture nuanced variations

. Therefore, we propose learning token-level correlations by dynamically leverag
ing the policy model's confidence during training. Comprehensive experiments on
QA, math, and instruction-following tasks demonstrate the effectiveness of our a
pproach, significantly surpassing competitive baselines, including DPO. Addition
ally, our in-depth quantitative analysis reveals the reasons behind our method's
 superior performance over DPO and showcases its versatility to other DPO varian
ts.
****
Keon Lee,Dong Won Kim,Jaehyeon Kim,Seungjun Chung,Jaewoong Cho
DiTTo-TTS: Diffusion Transformers for Scalable Text-to-Speech without Domain-Spe
cific Factors
Large-scale latent diffusion models (LDMs) excel in content generation across va
rious modalities, but their reliance on phonemes and durations in text-to-speech
 (TTS) limits scalability and access from other fields. While recent studies sho
w potential in removing these domain-specific factors, performance remains subop
timal. In this work, we introduce DiTTo-TTS, a Diffusion Transformer (DiT)-based
 TTS model, to investigate whether LDM-based TTS can achieve state-of-the-art pe
rformance without domain-specific factors. Through rigorous analysis and empiric
al exploration, we find that (1) DiT with minimal modifications outperforms U-Ne
t, (2) variable-length modeling with a speech length predictor significantly imp
roves results over fixed-length approaches, and (3) conditions like semantic ali
gnment in speech latent representations are key to further enhancement. By scali
ng our training data to 82K hours and the model size to 790M parameters, we achi
eve superior or comparable zero-shot performance to state-of-the-art TTS models
in naturalness, intelligibility, and speaker similarity, all without relying on
domain-specific factors. Speech samples are available at https://ditto-tts.githu
b.io.
****
Nico Daheim,Clara Meister,Thomas Möllenhoff,Iryna Gurevych
Uncertainty-Aware Decoding with Minimum Bayes Risk
Despite their outstanding performance in the majority of scenarios, contemporary
 language models still occasionally generate undesirable outputs, for example, h
allucinated text. While such behaviors have previously been linked to uncertaint
y, there is a notable lack of methods that actively consider uncertainty during
text generation. In this work, we show how Minimum Bayes Risk (MBR) decoding, wh
ich selects model generations according to an expected risk, can be generalized
into a principled uncertainty-aware decoding method. In short, we account for mo
del uncertainty during decoding by incorporating a posterior over model paramete
rs into MBR’s computation of expected risk. We show that this modified expected
risk is useful for both choosing outputs and deciding when to abstain from gener
ation and can provide improvements without incurring overhead. We benchmark diff
erent methods for learning posteriors and show that performance improves with pr
ediction diversity. We release our code publicly.
****
Hongyin Zhang,Pengxiang Ding,Shangke Lyu,Ying Peng,Donglin Wang
GEVRM: Goal-Expressive Video Generation Model For Robust Visual Manipulation
With the rapid development of embodied artificial intelligence, significant prog
ress has been made in vision-language-action (VLA) models for general robot deci
sion-making. However, the majority of existing VLAs fail to account for the inev
itable external perturbations encountered during deployment. These perturbations
 introduce unforeseen state information to the VLA, resulting in inaccurate acti
ons and consequently, a significant decline in generalization performance. The c
lassic internal model control (IMC) principle demonstrates that a closed-loop sy
stem with an internal model that includes external input signals can accurately
track the reference input and effectively offset the disturbance. We propose a n
ovel closed-loop VLA method GEVRM that integrates the IMC principle to enhance t
he robustness of robot visual manipulation. The text-guided video generation mod
el in GEVRM can generate highly expressive future visual planning goals. Simulta
neously, we evaluate perturbations by simulating responses, which are called int
ernal embeddings and optimized through prototype contrastive learning. This allo

ws the model to implicitly infer and distinguish perturbations from the external environment. The proposed GEVRM achieves state-of-the-art performance on both standard and perturbed CALVIN benchmarks and shows significant improvements in realistic robot tasks.
****
Guy Ohayon,Tomer Michaeli,Michael Elad
Posterior-Mean Rectified Flow: Towards Minimum MSE Photo-Realistic Image Restoration
Photo-realistic image restoration algorithms are typically evaluated by distortion measures (e.g., PSNR, SSIM) and by perceptual quality measures (e.g., FID, NIQE), where the desire is to attain the lowest possible distortion without compromising on perceptual quality. To achieve this goal, current methods commonly attempt to sample from the posterior distribution, or to optimize a weighted sum of a distortion loss (e.g., MSE) and a perceptual quality loss (e.g., GAN). Unlike previous works, this paper is concerned specifically with the *optimal* estimator that minimizes the MSE under a constraint of perfect perceptual index, namely where the distribution of the reconstructed images is equal to that of the ground-truth ones. A recent theoretical result shows that such an estimator can be constructed by optimally transporting the posterior mean prediction (MMSE estimate) to the distribution of the ground-truth images. Inspired by this result, we introduce Posterior-Mean Rectified Flow (PMRF), a simple yet highly effective algorithm that approximates this optimal estimator. In particular, PMRF first predicts the posterior mean, and then transports the result to a high-quality image using a rectified flow model that approximates the desired optimal transport map. We investigate the theoretical utility of PMRF and demonstrate that it consistently outperforms previous methods on a variety of image restoration tasks.
****
Ignat Georgiev,Varun Giridhar,Nicklas Hansen,Animesh Garg
PWM: Policy Learning with Multi-Task World Models
Reinforcement Learning (RL) has made significant strides in complex tasks but struggles in multi-task settings with different embodiments. World model methods offer scalability by learning a simulation of the environment but often rely on inefficient gradient-free optimization methods for policy extraction. In contrast, gradient-based methods exhibit lower variance but fail to handle discontinuities. Our work reveals that well-regularized world models can generate smoother optimization landscapes than the actual dynamics, facilitating more effective first-order optimization. We introduce Policy learning with multi-task World Models (PWM), a novel model-based RL algorithm for continuous control. Initially, the world model is pre-trained on offline data, and then policies are extracted from it using first-order optimization in less than 10 minutes per task. PWM effectively solves tasks with up to 152 action dimensions and outperforms methods that use ground-truth dynamics. Additionally, PWM scales to an 80-task setting, achieving up to 27\% higher rewards than existing baselines without relying on costly online planning. Visualizations and code are available at [imgeorgiev.com/pwm](https://www.imgeorgiev.com/pwm/).
****
ZHAOYANG LI,Minghao Han,Xunyuan Yin
MamKO: Mamba-based Koopman operator for modeling and predictive control
The Koopman theory, which enables the transformation of nonlinear systems into linear representations, is a powerful and efficient tool to model and control nonlinear systems. However, the ability of the Koopman operator to model complex systems, particularly time-varying systems, is limited by the fixed linear state-space representation. To address the limitation, the large language model, Mamba, is considered a promising strategy for enhancing modeling capabilities while preserving the linear state-space structure.
In this paper, we propose a new framework, the Mamba-based Koopman operator (MamKO), which provides enhanced model prediction capability and adaptability, as compared to Koopman models with constant Koopman operators. Inspired by the Mamba structure, MamKO generates Koopman operators from online data; this enables the model to effectively capture the dynamic behaviors of the nonlinear system over

time. A model predictive control system is then developed based on the proposed MamKO model. The modeling and control performance of the proposed method is evaluated through experiments on benchmark time-invariant and time-varying systems. The experimental results demonstrate the superiority of the proposed approach. Additionally, we perform ablation experiments to test the effectiveness of individual components of MamKO. This approach unlocks new possibilities for integrating large language models with control frameworks, and it achieves a good balance between advanced modeling capabilities and real-time control implementation efficiency.
****

Jingjie Zhang,Hanqun CAO,Zijun Gao,Xiaorui Wang,Chunbin Gu
SAGEPhos: Sage Bio-Coupled and Augmented Fusion for Phosphorylation Site Detection
Phosphorylation site prediction based on kinase-substrate interaction plays a vital role in understanding cellular signaling pathways and disease mechanisms. Computational methods for this task can be categorized into kinase-family-focused and individual kinase-targeted approaches. Individual kinase-targeted methods have gained prominence for their ability to explore a broader protein space and provide more precise target information for kinase inhibitors. However, most existing individual kinase-based approaches focus solely on sequence inputs, neglecting crucial structural information. To address this limitation, we introduce SAGEPhos (Structure-aware kinAse-substrate bio-coupled and bio-auGmented nEtwork for Phosphorylation site prediction), a novel framework that modifies the semantic space of main protein inputs using auxiliary inputs at two distinct modality levels. At the inter-modality level, SAGEPhos introduces a Bio-Coupled Modal Fusion method, distilling essential kinase sequence information to refine task-oriented local substrate feature space, creating a shared semantic space that captures crucial kinase-substrate interaction patterns. Within the substrate's intra-modality domain, it focuses on Bio-Augmented Fusion, emphasizing 2D local sequence information while selectively incorporating 3D spatial information from predicted structures to complement the sequence space. Moreover, to address the lack of structural information in current datasets, we contribute a new, refined phosphorylation site prediction dataset, which incorporates crucial structural elements and will serve as a new benchmark for the field. Experimental results demonstrate that SAGEPhos significantly outperforms baseline methods, notably achieving almost 10\% and 12\% improvements in prediction accuracy and AUC-ROC, respectively. We further demonstrate our algorithm's robustness and generalization through stable results across varied data partitions and significant improvements in zero-shot scenarios. These results underscore the effectiveness of constructing a larger and more precise protein space in advancing the state-of-the-art in phosphorylation site prediction. We release the SAGEPhos models and code at https://github.com/ZhangJJ26/SAGEPhos.
****

Zijian Chen,tingzhu chen,Wenjun Zhang,Guangtao Zhai
OBI-Bench: Can LMMs Aid in Study of Ancient Script on Oracle Bones?
We introduce OBI-Bench, a holistic benchmark crafted to systematically evaluate large multi-modal models (LMMs) on whole-process oracle bone inscriptions (OBI) processing tasks demanding expert-level domain knowledge and deliberate cognition. OBI-Bench includes 5,523 meticulously collected diverse-sourced images, covering five key domain problems: recognition, rejoining, classification, retrieval, and deciphering. These images span centuries of archaeological findings and years of research by front-line scholars, comprising multi-stage font appearances from excavation to synthesis, such as original oracle bone, inked rubbings, oracle bone fragments, cropped single characters, and handprinted characters. Unlike existing benchmarks, OBI-Bench focuses on advanced visual perception and reasoning with OBI-specific knowledge, challenging LMMs to perform tasks akin to those faced by experts. The evaluation of 6 proprietary LMMs as well as 17 open-source LMMs highlights the substantial challenges and demands posed by OBI-Bench. Even the latest versions of GPT-4o, Gemini 1.5 Pro, and Qwen-VL-Max are still far from public-level humans in some fine-grained perception tasks. However, they perf

orm at a level comparable to untrained humans in deciphering tasks, indicating r
emarkable capabilities in offering new interpretative perspectives and generatin
g creative guesses. We hope OBI-Bench can facilitate the community to develop do
main-specific multi-modal foundation models towards ancient language research an
d delve deeper to discover and enhance these untapped potentials of LMMs.
****

David Robinson,Marius Miron,Masato Hagiwara,Olivier Pietquin
NatureLM-audio: an Audio-Language Foundation Model for Bioacoustics
Large language models (LLMs) prompted with text and audio have achieved state-of
-the-art performance across various auditory tasks, including speech, music, and
 general audio, showing emergent abilities on unseen tasks. However, their poten
tial has yet to be fully demonstrated in bioacoustics tasks, such as detecting a
nimal vocalizations in large recordings, classifying rare and endangered species
, and labeling context and behavior—tasks that are crucial for conservation, bio
diversity monitoring, and animal behavior studies. In this work, we present Natu
reLM-audio, the first audio-language foundation model specifically designed for
bioacoustics. Our training dataset consists of carefully curated text-audio pair
s spanning bioacoustics, speech, and music, designed to address the field's limi
ted availability of annotated data. We demonstrate successful transfer of learne
d representations from music and speech to bioacoustics, and our model shows pro
mising generalization to unseen taxa and tasks. We evaluate NatureLM-audio on a
novel benchmark (BEANS-Zero) and it sets a new state of the art on several bioac
oustics tasks, including zero-shot classification of unseen species. To advance
bioacoustics research, we release our model weights, benchmark data, and open-so
urce the code for training and benchmark data generation and model training.
****

HONGZE CHEN,Zehong Lin,Jun Zhang
GI-GS: Global Illumination Decomposition on Gaussian Splatting for Inverse Rende
ring
We present GI-GS, a novel inverse rendering framework that leverages 3D Gaussian
 Splatting (3DGS) and deferred shading to achieve photo-realistic novel view syn
thesis and relighting. In inverse rendering, accurately modeling the shading pro
cesses of objects is essential for achieving high-fidelity results. Therefore, i
t is critical to incorporate global illumination to account for indirect lightin
g that reaches an object after multiple bounces across the scene. Previous 3DGS-
based methods have attempted to model indirect lighting by characterizing indire
ct illumination as learnable lighting volumes or additional attributes of each G
aussian, while using baked occlusion to represent shadow effects. These methods,
 however, fail to accurately model the complex physical interactions between lig
ht and objects, making it impossible to construct realistic indirect illuminatio
n during relighting. To address this limitation, we propose to calculate indirec
t lighting using efficient path tracing with deferred shading. In our framework,
 we first render a G-buffer to capture the detailed geometry and material proper
ties of the scene. Then, we perform physically-based rendering (PBR) only for di
rect lighting. With the G-buffer and previous rendering results, the indirect li
ghting can be calculated through a lightweight path tracing. Our method effectiv
ely models indirect lighting under any given lighting conditions, thereby achiev
ing better novel view synthesis and competitive relighting. Quantitative and qua
litative results show that our GI-GS outperforms existing baselines in both rend
ering quality and efficiency. Project page: https://stopaimme.github.io/GI-GS-si
te/.
****

Zhanfeng Mo,Haosen Shi,Sinno Jialin Pan
Probabilistic Neural Pruning via Sparsity Evolutionary Fokker-Planck-Kolmogorov
Equation
Neural pruning aims to compress and accelerate deep neural networks by identifyi
ng the optimal subnetwork within a specified sparsity budget. In this work, we s
tudy how to gradually sparsify the unpruned dense model to the target sparsity l
evel with minimal performance drop. Specifically, we analyze the evolution of th
e population of optimal subnetworks under continuous sparsity increments from a

thermodynamic perspective. We first reformulate neural pruning as an expected loss minimization problem over the mask distributions. Then, we establish an effective approximation for the sparsity evolution of the optimal mask distribution, termed the **S**parsity Evolutionary **F**okker-**P**lanck-**K**olmogorov Equation (**SFPK**), which provides closed-form, mathematically tractable guidance on distributional transitions for minimizing the expected loss under an infinitesimal sparsity increment. On top of that, we propose SFPK-pruner, a particle simulation-based probabilistic pruning method, to sample performant masks with desired sparsity from the destination distribution of SFPK. In theory, we establish the convergence guarantee for the proposed SFPK-pruner. Our SFPK-pruner exhibits competitive performance in various pruning scenarios. The code is available on https://github.com/mzf666/SFPK-main.
****
Xuanlei Zhao,Xiaolong Jin,Kai Wang,Yang You
Real-Time Video Generation with Pyramid Attention Broadcast
We present Pyramid Attention Broadcast (PAB), a real-time, high quality and training-free approach for DiT-based video generation. Our method is founded on the observation that attention difference in the diffusion process exhibits a U-shaped pattern, indicating significant redundancy. We mitigate this by broadcasting attention outputs to subsequent steps in a pyramid style. It applies different broadcast strategies to each attention based on their variance for best efficiency. We further introduce broadcast sequence parallel for more efficient distributed inference. PAB demonstrates up to 10.5x speedup across three models compared to baselines, achieving real-time generation for up to 720p videos. We anticipate that our simple yet effective method will serve as a robust baseline and facilitate future research and application for video generation.
****
Yeongmin Kim,Kwanghyeon Lee,Minsang Park,Byeonghu Na,Il-chul Moon
Diffusion Bridge AutoEncoders for Unsupervised Representation Learning
Diffusion-based representation learning has achieved substantial attention due to its promising capabilities in latent representation and sample generation. Recent studies have employed an auxiliary encoder to identify a corresponding representation from data and to adjust the dimensionality of a latent variable $\mathbf{z}$. Meanwhile, this auxiliary structure invokes an *information split problem*; the information of each data instance $\mathbf{x}_0$ is divided into diffusion endpoint $\mathbf{x}_T$ and encoded $\mathbf{z}$ because there exist two inference paths starting from the data. The latent variable modeled by diffusion endpoint $\mathbf{x}_T$ has some disadvantages. The diffusion endpoint $\mathbf{x}_T$ is computationally expensive to obtain and inflexible in dimensionality. To address this problem, we introduce Diffusion Bridge AuteEncoders (DBAE), which enables $\mathbf{z}$-dependent endpoint $\mathbf{x}_T$ inference through a feed-forward architecture. This structure creates an information bottleneck at $\mathbf{z}$, so $\mathbf{x}_T$ becomes dependent on $\mathbf{z}$ in its generation. This results in $\mathbf{z}$ holding the full information of data. We propose an objective function for DBAE to enable both reconstruction and generative modeling, with their theoretical justification. Empirical evidence supports the effectiveness of the intended design in DBAE, which notably enhances downstream inference quality, reconstruction, and disentanglement. Additionally, DBAE generates high-fidelity samples in the unconditional generation. Our code is available at https://github.com/aailab-kaist/DBAE.
****
Jacopo Iollo,Christophe Heinkelé,Pierre Alliez,Florence Forbes
Bayesian Experimental Design Via Contrastive Diffusions
Bayesian Optimal Experimental Design (BOED) is a powerful tool to reduce the cost of running a sequence of experiments.
When based on the Expected Information Gain (EIG), design optimization corresponds to the maximization of some intractable expected *contrast* between prior and posterior distributions.
Scaling this maximization to high dimensional and complex settings has been an issue due to BOED inherent computational complexity.

In this work, we introduce an *pooled posterior* distribution with cost-effectiv
e sampling properties and provide a tractable access to the EIG contrast maximiz
ation via a new EIG gradient expression. Diffusion-based samplers are used to co
mpute the dynamics of the pooled posterior and ideas from bi-level optimization
are leveraged to derive an efficient joint sampling-optimization loop, without r
esorting to lower bound approximations of the EIG. The resulting efficiency gain
 allows to extend BOED to the well-tested generative capabilities of diffusion m
odels.
By incorporating generative models into the BOED framework, we expand its scope
and its use in scenarios that were previously impractical. Numerical experiments
 and comparison with state-of-the-art methods show the potential of the approach
.
****
Yi Li,Yuquan Deng,Jesse Zhang,Joel Jang,Marius Memmel,Caelan Reed Garrett,Fabio
Ramos,Dieter Fox,Anqi Li,Abhishek Gupta,Ankit Goyal
HAMSTER: Hierarchical Action Models for Open-World Robot Manipulation
Large foundation models have shown strong open-world generalization to complex p
roblems in vision and language, but similar levels of generalization have yet to
 be achieved in robotics. One fundamental challenge is the lack of robotic data,
 which are typically obtained through expensive on-robot operation. A promising
remedy is to leverage cheaper, *off-domain* data such as action-free videos, han
d-drawn sketches, or simulation data. In this work, we posit that *hierarchical*
 vision-language-action (VLA) models can be more effective in utilizing off-doma
in data than standard monolithic VLA models that directly finetune vision-langua
ge models (VLMs) to predict actions.
In particular, we study a class of hierarchical VLA models, where the high-level
 VLM is finetuned to produce a coarse 2D path indicating the desired robot end-e
ffector trajectory given an RGB image and a task description. The intermediate 2
D path prediction is then served as guidance to the low-level, 3D-aware control
policy capable of precise manipulation. Doing so alleviates the high-level VLM f
rom fine-grained action prediction, while reducing the low-level policy's burden
 on complex task-level reasoning.
We show that, with the hierarchical design, the high-level VLM can transfer acro
ss significant domain gaps between the off-domain finetuning data and real-robot
 testing scenarios, including differences in embodiments, dynamics, visual appea
rances, and task semantics, etc.
In the real-robot experiments, we observe an average of 20% improvement in succe
ss rate across seven different axes of generalization over OpenVLA, representing
 a 50% relative gain.
Visual results are provided at: [https://hamster-robot.github.io/](https://hamst
er-robot.github.io/)
****
Sinan Fan,Liang Xie,Chen Shen,Ge Teng,Xiaosong Yuan,Xiaofeng Zhang,Chenxi Huang,
Wenxiao Wang,Xiaofei He,Jieping Ye
Improving Complex Reasoning with Dynamic Prompt Corruption: A Soft Prompt Optimi
zation Approach
Prompt Tuning (PT) has emerged as a promising Parameter-Efficient Fine-Tuning (P
EFT) approach by appending trainable continuous prompt vectors to the input, mai
ntaining competitive performance with significantly fewer trainable parameters.
While PT has shown effectiveness in enhancing task performance, particularly for
 classification tasks, its application to complex reasoning tasks has been large
ly overlooked. Our investigation reveals that PT provides limited improvement an
d may even degrade performance in reasoning tasks. This phenomenon suggests that
 soft prompts can positively impact certain instances while negatively affecting
 others, particularly during the latter stages of reasoning.
To address these challenges, we propose a novel method called Dynamic Prompt Cor
ruption (DPC), which seeks to optimize the use of soft prompts in reasoning task
s. DPC dynamically adjusts the influence of soft prompts based on their impact o
n the reasoning process. Specifically, it involves two key components: Dynamic T
rigger and Dynamic Corruption. Dynamic Trigger measures the influence of soft pr

ompts, determining whether their impact is beneficial or detrimental. Dynamic Co
rruption mitigates the negative effects of soft prompts by selectively masking k
ey tokens that interfere with the reasoning process.
We validate our approach through extensive experiments on various large language
 models (LLMs) and reasoning tasks, including GSM8K, MATH, and AQuA. The results
 demonstrate that Dynamic Prompt Corruption consistently improves the performanc
e of LLMs, achieving 4\%-8\% accuracy gains compared to standard prompt tuning.
 These findings highlight the effectiveness of our approach and its potential to
 enhance complex reasoning in LLMs.
****
Hossein Taheri,Christos Thrampoulidis,Arya Mazumdar
Sharper Guarantees for Learning Neural Network Classifiers with Gradient Methods
In this paper, we study the data-dependent convergence and generalization behavi
or of gradient methods for neural networks with smooth activation. Our first res
ult is a novel bound on the excess risk of deep networks trained by the logistic
 loss via an alogirthmic stability analysis. Compared to previous works, our res
ults improve upon the shortcomings of the well-established Rademacher complexity
-based bounds. Importantly, the bounds we derive in this paper are tighter, hold
 even for neural networks of small width, do not scale unfavorably with width, a
re algorithm-dependent, and consequently capture the role of initialization on t
he sample complexity of gradient descent for deep nets. Specialized to noiseless
 data separable with margin γ by neural tangent kernel (NTK) features of
a network of width $\Omega(poly(\log(n)))$, we show the test-error rate $e^{O(L)
}/{\gamma^2 n}$, where n is the training set size and L denotes the number o
f hidden layers. This results in an improvement in the test loss bound compared
to previous works while maintaining the poly-logarithmic width conditions. We fu
rther investigate excess risk bounds for deep nets trained with noisy data, esta
blishing that under a polynomial condition on the network width, gradient descen
t can achieve the optimal excess risk. Finally, we show that a large step-size s
ignificantly improves upon the NTK regime's results in classifying the XOR distr
ibution. In particular, we show for a one-hidden layer neural network of constan
t width m with quadratic activation and standard Gaussian initialization that
SGD with linear sample complexity and with a large step-size $\eta=m$ reaches th
e perfect test accuracy after only $\lceil\log(d)\rceil$ iterations, where d i
s the data dimension.
****
Feiyang Ye,Yueming Lyu,Xuehao Wang,Masashi Sugiyama,Yu Zhang,Ivor Tsang
Sharpness-Aware Black-Box Optimization
Black-box optimization algorithms have been widely used in various machine learn
ing problems, including reinforcement learning and prompt fine-tuning. However,
directly optimizing the training loss value, as commonly done in existing black-
box optimization methods, could lead to suboptimal model quality and generalizat
ion performance. To address those problems in black-box optimization, we propose
 a novel Sharpness-Aware Black-box Optimization (SABO) algorithm, which applies
a sharpness-aware minimization strategy to improve the model generalization. Spe
cifically, the proposed SABO method first reparameterizes the objective function
 by its expectation over a Gaussian distribution. Then it iteratively updates th
e parameterized distribution by approximated stochastic gradients of the maximum
 objective value within a small neighborhood around the current solution in the
Gaussian distribution space. Theoretically, we prove the convergence rate and ge
neralization bound of the proposed SABO algorithm. Empirically, extensive experi
ments on the black-box prompt fine-tuning tasks demonstrate the effectiveness of
 the proposed SABO method in improving model generalization performance.
****
Aldo Pacchiano
Second Order Bounds for Contextual Bandits with Function Approximation
Many works have developed no-regret algorithms for contextual bandits with funct
ion approximation, where the mean rewards over context-action pairs belong to a
function class $\mathcal{F}$. Although there are many approaches to this problem
, algorithms based on the principle of optimism, such as optimistic least square

s have gained in importance. It can be shown the regret of this algorithm scales as $\widetilde{\mathcal{O}}\left(\sqrt{d_{\mathrm{eluder}}(\mathcal{F}) \log(\mathcal{F}) T }\right)$ where $d_{\mathrm{eluder}}(\mathcal{F})$ is a statistical measure of the complexity of the function class $\mathcal{F}$ known as eluder dimension. Unfortunately, even if the variance of the measurement noise of the rewards at time t equals σ_t^2 and these are close to zero, the optimistic least squares algorithm’s regret scales with $\sqrt{T}$. In this work we are the first to develop algorithms that satisfy regret bounds for contextual bandits with function approximation of the form $\widetilde{\mathcal{O}}\left(\sigma \sqrt{\log(\mathcal{F})d_{\mathrm{eluder}}(\mathcal{F}) T } + d_{\mathrm{eluder}}(\mathcal{F}) \cdot \log(|\mathcal{F}|)\right) $ when the variances are unknown and satisfy $\sigma_t^2 = \sigma$ for all t and $\widetilde{\mathcal{O}}\left(d_{\mathrm{eluder}}(\mathcal{F})\sqrt{\log(\mathcal{F})\sum_{t=1}^T \sigma_t^2 } + d_{\mathrm{eluder}}(\mathcal{F}) \cdot \log(|\mathcal{F}|)\right) $ when the variances change every time-step. These bounds generalize existing techniques for deriving second order bounds in contextual linear problems.
****

Gwangpyo Yoo,Honguk Woo
Model Risk-sensitive Offline Reinforcement Learning
Offline reinforcement learning (RL) is becoming critical in risk-sensitive areas such as finance and autonomous driving, where incorrect decisions can lead to substantial financial loss or compromised safety. However, traditional risk-sensitive offline RL methods often struggle with accurately assessing risk, with minor errors in the estimated return potentially causing significant inaccuracies of risk estimation. These challenges are intensified by distribution shifts inherent in offline RL. To mitigate these issues, we propose a model risk-sensitive offline RL framework designed to minimize the worst-case of risks across a set of plausible alternative scenarios rather than solely focusing on minimizing estimated risk. We present a critic-ensemble criterion method that identifies the plausible alternative scenarios without introducing additional hyperparameters. We also incorporate the learned Fourier feature framework and the IQN framework to address spectral bias in neural networks, which can otherwise lead to severe errors in calculating model risk. Our experiments in finance and self-driving scenarios demonstrate that the proposed framework significantly reduces risk, by 11.2% to 18.5%, compared to the most outperforming risk-sensitive offline RL baseline, particularly in highly uncertain environments.
****

Fangxin Wang,Kay Liu,Sourav Medya,Philip S. Yu
BANGS: Game-theoretic Node Selection for Graph Self-Training
Graph self-training is a semi-supervised learning method that iteratively selects a set of unlabeled data to retrain the underlying graph neural network (GNN) model and improve its prediction performance. While selecting highly confident nodes has proven effective for self-training, this pseudo-labeling strategy ignores the combinatorial dependencies between nodes and suffers from a local view of the distribution.
To overcome these issues, we propose BANGS, a novel framework that unifies the labeling strategy with conditional mutual information as the objective of node selection. Our approach---grounded in game theory---selects nodes in a combinatorial fashion and provides theoretical guarantees for robustness under noisy objective. More specifically, unlike traditional methods that rank and select nodes independently, BANGS considers nodes as a collective set in the self-training process. Our method demonstrates superior performance and robustness across various datasets, base models, and hyperparameter settings, outperforming existing techniques. The codebase is available on https://anonymous.4open.science/r/BANGS-3EA4.
****

Kaiyue Wen,Xingyu Dang,Kaifeng Lyu
RNNs are not Transformers (Yet): The Key Bottleneck on In-Context Retrieval
This paper investigates the gap in representation powers of Transformers and Recurrent Neural Networks (RNNs), which are more memory efficient than Transformers

. We aim to understand whether RNNs can match the performance of Transformers, p
articularly when enhanced with Chain-of-Thought (CoT) prompting. Our theoretical
 analysis reveals that CoT improves RNNs but is insufficient to close the gap wi
th Transformers. A key bottleneck lies in the inability of RNNs to perfectly ret
rieve information from the context, even with CoT:
for several tasks that explicitly or implicitly require this capability, such as
 associative recall and determining if a graph is a tree, we prove that RNNs are
 not expressive enough to solve the tasks while Transformers can solve them with
 ease.
Conversely, we prove that adopting techniques to enhance the in-context retrieva
l capability of RNNs, including Retrieval-Augmented Generation (RAG) and adding
a single Transformer layer, can elevate RNNs to be capable of solving all polyno
mial-time solvable problems with CoT, hence closing the representation gap with
Transformers. We validate our theory on synthetic and natural language experimen
ts.
****
Jinhao Jiang,Junyi Li,Xin Zhao,Yang Song,Tao Zhang,Ji-Rong Wen
Mix-CPT: A Domain Adaptation Framework via Decoupling Knowledge Learning and For
mat Alignment
Adapting large language models (LLMs) to specialized domains typically requires
domain-specific corpora for continual pre-training to facilitate knowledge memor
ization and related instructions for fine-tuning to apply this knowledge.
However, this method may lead to inefficient knowledge memorization due to a lac
k of awareness of knowledge utilization during the continual pre-training and de
mands LLMs to simultaneously learn knowledge utilization and format alignment wi
th divergent training objectives during the fine-tuning.
To enhance the domain adaptation of LLMs, we revise this process and propose a n
ew domain adaptation framework including domain knowledge learning and general f
ormat alignment, called \emph{Mix-CPT}. Specifically, we first conduct a knowled
ge mixture continual pre-training that concurrently focuses on knowledge memoriz
ation and utilization. To avoid catastrophic forgetting, we further propose a lo
git swap self-distillation constraint. By leveraging the knowledge and capabilit
ies acquired during continual pre-training, we then efficiently perform instruct
ion tuning and alignment with a few general training samples to achieve format a
lignment.
Extensive experiments show that our proposed \emph{Mix-CPT} framework can simult
aneously improve the task-solving capabilities of LLMs on the target and general
 domains.
****
Arhaan Ahmad,Tanay Vineet Tayal,Ashutosh Gupta,S. Akshay
Sensitivity Verification for Additive Decision Tree Ensembles
Tree ensemble models, such as Gradient Boosted Decision Trees (GBDTs) and random
 forests, are widely popular models for a variety of machine learning tasks. The
 power of these models comes from the ensemble of decision trees, which makes an
alysis of such models significantly harder than for single trees. As a result, r
ecent work has focused on developing exact and approximate techniques for questi
ons such as robustness verification, fairness and explainability for such models
 of tree ensembles.

In this paper, we focus on a specific problem of feature sensitivity for additiv
e decision tree ensembles and build a formal verification framework for a parame
trized variant of it, where we also take into account the confidence of the tree
 ensemble in its output. We start by showing theoretical (NP-)hardness of the pr
oblem and explain how it relates to other verification problems. Next, we provid
e a novel encoding of the problem using pseudo-Boolean constraints. Based on thi
s encoding, we develop a tunable algorithm to perform sensitivity analysis, whic
h can trade off precision for running time. We implement our algorithm and study
 its performance on a suite of GBDT benchmarks from the literature. Our experime
nts show the practical utility of our approach and its improved performance comp
ared to existing approaches.

****
Leo Richter,Xuanli He,Pasquale Minervini,Matt Kusner
An Auditing Test to Detect Behavioral Shift in Language Models
As language models (LMs) approach human-level performance, a comprehensive under standing of their behavior becomes crucial.
This includes evaluating capabilities, biases, task performance, and alignment w ith societal values. Extensive initial evaluations, including red teaming and di verse benchmarking, can establish a model’s behavioral profile. However, subsequ ent fine-tuning or deployment modifications may alter these behaviors in uninten ded ways. We present an efficient statistical test to tackle Behavioral Shift Au diting (BSA) in LMs, which we define as detecting distribution shifts in qualita tive properties of the output distributions of LMs. Our test compares model gene rations from a baseline model to those of the model under scrutiny and provides theoretical guarantees for change detection while controlling false positives. T he test features a configurable tolerance parameter that adjusts sensitivity to behavioral changes for different use cases. We evaluate our approach using two c ase studies: monitoring changes in (a) toxicity and (b) translation performance. We find that the test is able to detect meaningful changes in behavior distribu tions using just hundreds of examples.
****
Junlin Wang,Jue WANG,Ben Athiwaratkun,Ce Zhang,James Zou
Mixture-of-Agents Enhances Large Language Model Capabilities
Recent advances in large language models (LLMs) demonstrate substantial capabili ties in natural language understanding and generation tasks. With the growing nu mber of LLMs, how to harness the collective expertise of multiple LLMs is an exc iting open direction. Toward this goal, we propose a new approach that leverages the collective strengths of multiple LLMs through a Mixture-of-Agents (MoA) met hodology. In our approach, we construct a layered MoA architecture wherein each layer comprises multiple LLM agents. Each agent takes all the outputs from agent s in the previous layer as auxiliary information in generating its response. MoA models achieves state-of-art performance on AlpacaEval 2.0, Arena-Hard, MT-Benc h, and FLASK, surpassing GPT-4 Omni. For example, our MoA using only open-source LLMs achieves a score of 65.1% on AlpacaEval 2.0 compared to 57.5% by GPT-4 Omn i.
****
Zhenhong Zhou,Haiyang Yu,Xinghua Zhang,Rongwu Xu,Fei Huang,Kun Wang,Yang Liu,Jun feng Fang,Yongbin Li
On the Role of Attention Heads in Large Language Model Safety
Large language models (LLMs) achieve state-of-the-art performance on multiple la nguage tasks, yet their safety guardrails can be circumvented, leading to harmfu l generations. In light of this, recent research on safety mechanisms has emerge d, revealing that when safety representations or component are suppressed, the s afety capability of LLMs are compromised. However, existing research tends to ov erlook the safety impact of multi-head attention mechanisms, despite their cruci al role in various model functionalities. Hence, in this paper, we aim to explor e the connection between standard attention mechanisms and safety capability to fill this gap in the safety-related mechanistic interpretability. We propose an novel metric which tailored for multi-head attention, the Safety Head ImPortant Score (Ships), to assess the individual heads' contributions to model safety. Ba se on this, we generalize Ships to the dataset level and further introduce the S afety Attention Head AttRibution Algorithm (Sahara) to attribute the critical sa fety attention heads inside the model. Our findings show that special attention head has a significant impact on safety. Ablating a single safety head allows al igned model (e.g., Llama-2-7b-chat) to respond to **16$\times\uparrow$** more ha rmful queries, while only modifying **0.006\%** $\downarrow$ of the parameters, in contrast to the $\sim$ **5\%** modification required in previous studies. Mor e importantly, we demonstrate that attention heads primarily function as feature extractors for safety and models fine-tuned from the same base model exhibit ov erlapping safety heads through comprehensive experiments. Together, our attribut ion approach and findings provide a novel perspective for unpacking the black bo

x of safety mechanisms in large models.
****
Yusei Ito,Tatsunori Taniai,Ryo Igarashi,Yoshitaka Ushiku,Kanta Ono
Rethinking the role of frames for SE(3)-invariant crystal structure modeling
Crystal structure modeling with graph neural networks is essential for various a
pplications in materials informatics, and capturing SE(3)-invariant geometric fe
atures is a fundamental requirement for these networks. A straightforward approa
ch is to model with orientation-standardized structures through structure-aligne
d coordinate systems, or “frames.” However, unlike molecules, determining frame
s for crystal structures is challenging due to their infinite and highly symmetr
ic nature. In particular, existing methods rely on a statically fixed frame for
each structure, determined solely by its structural information, regardless of t
he task under consideration. Here, we rethink the role of frames, *questioning w
hether such simplistic alignment with the structure is sufficient*, and propose
the concept of *dynamic frames*. While accommodating the infinite and symmetric
nature of crystals, these frames provide each atom with a dynamic view of its lo
cal environment, focusing on actively interacting atoms. We demonstrate this con
cept by utilizing the attention mechanism in a recent transformer-based crystal
encoder, resulting in a new architecture called **CrystalFramer**. Extensive exp
eriments show that CrystalFramer outperforms conventional frames and existing cr
ystal encoders in various crystal property prediction tasks.
****
Sijia Zhang,Shuli Zeng,Shaoang Li,Feng Wu,Xiangyang Li
Learning to Select Nodes in Branch and Bound with Sufficient Tree Representation
Branch-and-bound methods are pivotal in solving Mixed Integer Linear Programming
 (MILP), where the challenge of node selection arises, necessitating the priorit
ization of different regions of the space for subsequent exploration. While mach
ine learning techniques have been proposed to address this, two crucial problems
 concerning \textbf{(P1)} how to sufficiently extract features from the branch-a
nd-bound tree, and \textbf{(P2)} how to assess the node quality comprehensively
based on the features remain open. To tackle these challenges, we propose to tac
kle the node selection problem employing a novel Tripartite graph representation
 and Reinforcement learning with a Graph Neural Network model (TRGNN). The tripa
rtite graph is theoretically proved to encompass sufficient information for tree
 representation in information theory. We learn node selection via reinforcement
 learning for learning delay rewards and give more comprehensive node metrics. E
xperiments show that TRGNN significantly improves the efficiency of solving MILP
s compared to human-designed and learning-based node selection methods on both s
ynthetic and large-scale real-world MILPs. Moreover, experiments demonstrate tha
t TRGNN well generalizes to MILPs that are significantly larger than those seen
during training.
****
Erik Jones,Arjun Patrawala,Jacob Steinhardt
Uncovering Gaps in How Humans and LLMs Interpret Subjective Language
Humans often rely on subjective natural language to direct language models (LLMs
); for example, users might instruct the LLM to write an *enthusiastic* blogpost
, while developers might train models to be *helpful* and *harmless* using LLM-b
ased edits. The LLM’s *operational semantics* of such subjective phrases---how i
t adjusts its behavior when each phrase is included in the prompt---thus dictate
s how aligned it is with human intent. In this work, we uncover instances of *mi
salignment* between LLMs' actual operational semantics and what humans expect. O
ur method, TED (thesaurus error detector), first constructs a thesaurus that cap
tures whether two phrases have similar operational semantics according to the LL
M. It then elicits failures by unearthing disagreements between this thesaurus a
nd a human-constructed reference. TED routinely produces surprising instances of
 misalignment; for example, Mistral 7B Instruct produces more *harassing* output
s when it edits text to be *witty*, and Llama 3 8B Instruct produces *dishonest*
 articles when instructed to make the articles *enthusiastic*. Our results demon
strate that humans can uncover unexpected LLM behavior by scrutinizing relations
hips between abstract concepts, without supervising outputs directly.

****
Rana Shahroz,Pingzhi Li,Sukwon Yun,Zhenyu Wang,Shahriar Nirjon,Chau-Wai Wong,Tianlong Chen
PortLLM: Personalizing Evolving Large Language Models with Training-Free and Portable Model Patches
As large language models (LLMs) increasingly shape the AI landscape, fine-tuning pretrained models has become more popular than in the pre-LLM era for achieving optimal performance in domain-specific tasks. However, pretrained LLMs such as ChatGPT are periodically evolved (i.e., model parameters are frequently updated), making it challenging for downstream users with limited resources to keep up with fine-tuning the newest LLMs for their domain application. Even though fine-tuning costs have nowadays been reduced thanks to the innovations of parameter-efficient fine-tuning such as LoRA, not all downstream users have adequate computing for frequent personalization. Moreover, access to fine-tuning datasets, particularly in sensitive domains such as healthcare, could be time-restrictive, making it crucial to retain the knowledge encoded in earlier fine-tuned rounds for future adaptation. In this paper, we present PORTLLM, a training-free framework that (i) creates an initial lightweight model update patch to capture domain-specific knowledge, and (ii) allows a subsequent seamless plugging for the continual personalization of evolved LLM at minimal cost. Our extensive experiments cover seven representative datasets, from easier question-answering tasks {BoolQ, SST2} to harder reasoning tasks {WinoGrande, GSM8K}, and models including {Mistral-7B,Llama2, Llama3.1, and Gemma2}, validating the portability of our designed model patches and showcasing the effectiveness of our proposed framework. For instance, PORTLLM achieves comparable performance to LoRA fine-tuning with reductions of up to 12.2× in GPU memory usage. Finally, we provide theoretical justifications to understand the portability of our model update patches, which offers new insights into the theoretical dimension of LLMs' personalization.
****
Hugo Thimonier,José Lucas De Melo Costa,Fabrice Popineau,Arpad Rimmel,Bich-Liên DOAN
T-JEPA: Augmentation-Free Self-Supervised Learning for Tabular Data
Self-supervision is often used for pre-training to foster performance on a downstream task by constructing meaningful representations of samples. Self-supervised learning (SSL) generally involves generating different views of the same sample and thus requires data augmentations that are challenging to construct for tabular data. This constitutes one of the main challenges of self-supervision for structured data. In the present work, we propose a novel augmentation-free SSL method for tabular data. Our approach, T-JEPA, relies on a Joint Embedding Predictive Architecture (JEPA) and is akin to mask reconstruction in the latent space. It involves predicting the latent representation of one subset of features from the latent representation of a different subset within the same sample, thereby learning rich representations without augmentations. We use our method as a pre-training technique and train several deep classifiers on the obtained representation. Our experimental results demonstrate a substantial improvement in both classification and regression tasks, outperforming models trained directly on samples in their original data space. Moreover, T-JEPA enables some methods to consistently outperform or match the performance of traditional methods likes Gradient Boosted Decision Trees. To understand why, we extensively characterize the obtained representations and show that T-JEPA effectively identifies relevant features for downstream tasks without access to the labels. Additionally, we introduce regularization tokens, a novel regularization method critical for training of JEPA-based models on structured data.
****
Taishi Nakamura,Takuya Akiba,Kazuki Fujii,Yusuke Oda,Rio Yokota,Jun Suzuki
Drop-Upcycling: Training Sparse Mixture of Experts with Partial Re-initialization
The Mixture of Experts (MoE) architecture reduces the training and inference cost significantly compared to a dense model of equivalent capacity. Upcycling is an approach that initializes and trains an MoE model using a pre-trained dense mo

del. While upcycling leads to initial performance gains, the training progresses slower than when trained from scratch, leading to suboptimal performance in the long term. We propose Drop-Upcycling - a method that effectively addresses this problem. Drop-Upcycling combines two seemingly contradictory approaches: utilizing the knowledge of pre-trained dense models while statistically re-initializing some parts of the weights. This approach strategically promotes expert specialization, significantly enhancing the MoE model's efficiency in knowledge acquisition.
Extensive large-scale experiments demonstrate that Drop-Upcycling significantly outperforms previous MoE construction methods in the long term, specifically when training on hundreds of billions of tokens or more.
As a result, our MoE model with 5.9B active parameters achieves comparable performance to a 13B dense model in the same model family, while requiring approximately 1/4 of the training FLOPs.
All experimental resources, including source code, training data, model checkpoints and logs, are publicly available to promote reproducibility and future research on MoE.
****
Yujin Han,Lei Xu,Sirui Chen,Difan Zou,Chaochao Lu
Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension ability
Large language models (LLMs) have shown remarkable capability in natural language tasks, yet debate persists on whether they truly comprehend deep structure (i.e., core semantics) or merely rely on surface structure (e.g., presentation format). Prior studies observe that LLMs' performance declines when intervening on surface structure, arguing their success relies on surface structure recognition. However, surface structure sensitivity does not prevent deep structure comprehension. Rigorously evaluating LLMs' capability requires analyzing both, yet deep structure is often overlooked. To this end, we assess LLMs' comprehension ability using causal mediation analysis, aiming to fully discover the capability of using both deep and surface structures. Specifically, we formulate the comprehension of deep structure as direct causal effect (DCE) and that of surface structure as indirect causal effect (ICE), respectively. To address the non-estimability of original DCE and ICE --- stemming from the infeasibility of isolating mutual influences of deep and surface structures, we develop the corresponding quantifiable surrogates, including approximated DCE (ADCE) and approximated ICE (AICE). We further apply the ADCE to evaluate a series of mainstream LLMs (and the one with random weights), showing that most of them exhibit deep structure comprehension ability, which grows along with the prediction accuracy. Comparing ADCE and AICE demonstrates closed-source LLMs (e.g., GPT) rely more on deep structure, while open-source LLMs (e.g., Llama) are more surface-sensitive, which decreases with model scale. Theoretically, ADCE is a bidirectional evaluation, which measures both the sufficiency and necessity of deep structure changes in causing output variations, thus offering a more comprehensive assessment than accuracy, a common evaluation in LLMs. Our work provides new insights into LLMs' deep structure comprehension and offers novel methods for LLMs evaluation. The code for our project is available at [ADCE Project](https://github.com/OpenCausaLab/ADCE).
****
Yang Li,Jiale Ma,Wenzheng Pan,Runzhong Wang,Haoyu Geng,Nianzu Yang,Junchi Yan
Unify ML4TSP: Drawing Methodological Principles for TSP and Beyond from Streamlined Design Space of Learning and Search
Despite the rich works on machine learning (ML) for combinatorial optimization (CO), a unified, principled framework remains lacking. This study utilizes the Travelling Salesman Problem (TSP) as a major case study, with adaptations demonstrated for other CO problems, dissecting established mainstream learning-based solvers to outline a comprehensive design space. We present ML4TSPBench, which advances a unified modular streamline incorporating existing technologies in both learning and search for transparent ablation, aiming to reassess the role of learning and discern which parts of existing techniques are genuinely beneficial and which are not. This further leads to the investigation of desirable principles of learning designs and the exploration of concepts guiding method designs. We de

monstrate the desirability of principles such as joint probability estimation, s
ymmetry solution representation, and online optimization for learning-based desi
gns. Leveraging the findings, we propose enhancements to existing methods to com
pensate for their missing attributes, thereby advancing performance and enrichin
g the technique library. From a higher viewpoint, we also uncover a performance
advantage in non-autoregressive and supervised paradigms compared to their count
erparts. The strategic decoupling and organic recompositions yield a factory of
new TSP solvers, where we investigate synergies across various method combinatio
ns and pinpoint the optimal design choices to create more powerful ML4TSP solver
s, thereby facilitating and offering a reference for future research and enginee
ring endeavors.
****

Theo Bourdais,Houman Owhadi

Minimal Variance Model Aggregation: A principled, non-intrusive, and versatile i
ntegration of black box models

Whether deterministic or stochastic, models can be viewed as functions designed
to approximate a specific quantity of interest.
We introduce Minimal Empirical Variance Aggregation (MEVA), a data-driven framew
ork that integrates predictions from various models, enhancing overall accuracy
by leveraging the individual strengths of each. This non-intrusive, model-agnost
ic approach treats the contributing models as black boxes and accommodates outpu
ts from diverse methodologies, including machine learning algorithms and traditi
onal numerical solvers.
We advocate for a point-wise linear aggregation process and consider two methods
 for optimizing this aggregate: Minimal Error Aggregation (MEA), which minimizes
 the prediction error, and Minimal Variance Aggregation (MVA), which focuses on
reducing variance. We prove a theorem showing that MVA can be more robustly esti
mated from data than MEA, making MEVA superior to Minimal Empirical Error Aggreg
ation (MEEA). Unlike MEEA, which interpolates target values directly, MEVA formu
lates aggregation as an error estimation problem, which can be performed
using any backbone learning paradigm. We demonstrate the versatility and effecti
veness of our framework across various applications, including data science and
partial differential equations, illustrating its ability to significantly enhanc
e both robustness and accuracy.
****

Bartlomiej Sobieski,Jakub Grzywaczewski,Bart■omiej Sadlej,Matthew Tivnan,Przemys
law Biecek

Rethinking Visual Counterfactual Explanations Through Region Constraint

Visual counterfactual explanations (VCEs) have recently gained immense popularit
y as a tool for clarifying the decision-making process of image classifiers. Thi
s trend is largely motivated by what these explanations promise to deliver -- in
dicate semantically meaningful factors that change the classifier's decision. Ho
wever, we argue that current state-of-the-art approaches lack a crucial componen
t -- the region constraint -- whose absence prevents from drawing explicit concl
usions, and may even lead to faulty reasoning due to phenomenons like confirmati
on bias. To address the issue of previous methods, which modify images in a very
 entangled and widely dispersed manner, we propose region-constrained VCEs (RVCE
s), which assume that only a predefined image region can be modified to influenc
e the model's prediction. To effectively sample from this subclass of VCEs, we p
ropose Region-Constrained Counterfactual Schrödinger Bridge (RCSB), an adaptatio
n of a tractable subclass of Schrödinger Bridges to the problem of conditional i
npainting, where the conditioning signal originates from the classifier of inter
est. In addition to setting a new state-of-the-art by a large margin, we extend
RCSB to allow for exact counterfactual reasoning, where the predefined region co
ntains only the factor of interest, and incorporating the user to actively inter
act with the RVCE by predefining the regions manually.
****

Boyang Sun,Yu Yao,Guang-Yuan Hao,Yumou Qiu,Kun Zhang

A Conditional Independence Test in the Presence of Discretization

Testing conditional independence (CI) has many important applications, such as B

ayesian network learning and causal discovery. Although several approaches have been developed for learning CI structures for observed variables, those existing methods generally fail to work when the variables of interest can not be directly observed and only discretized values of those variables are available. For example, if X_1, $\tilde{X}_2$ and X_3 are the observed variables, where $\tilde{X}_2$ is a discretization of the latent variable X_2, applying the existing methods to the observations of X_1, $\tilde{X}_2$ and X_3 would lead to a false conclusion about the underlying CI of variables X_1, X_2 and X_3.
Motivated by this, we propose a CI test specifically designed to accommodate the presence of discretization. To achieve this, a bridge equation and nodewise regression are used to recover the precision coefficients reflecting the conditional dependence of the latent continuous variables under the nonparanormal model. An appropriate test statistic has been proposed, and its asymptotic distribution under the null hypothesis of CI has been derived.
Theoretical analysis, along with empirical validation on various datasets, rigorously demonstrates the effectiveness of our testing methods.
****

Jingcun Wang,Yu-Guang Chen,Ing-Chao Lin,Bing Li,Grace Li Zhang
Basis Sharing: Cross-Layer Parameter Sharing for Large Language Model Compression
Large Language Models (LLMs) have achieved remarkable breakthroughs. However, the huge number of parameters in LLMs require significant amount of memory storage in inference, which prevents their practical deployment in many applications. To reduce memory storage of LLMs, singular value decomposition (SVD) provides a promising solution to approximate weight matrices for compressing LLMs. In this paper, we take a step further to explore parameter sharing across different layers with SVD to achieve more effective compression for LLMs. Specifically, weight matrices in different layers are decomposed and represented with a linear combination of a set of shared basis vectors and unique coefficients. The types of weight matrices and the layer selection for basis sharing are examined when compressing LLMs to maintain the performance. Comprehensive experiments demonstrate that Basis-Sharing outperforms state-of-the-art SVD-based compression approaches, especially at large compression ratios.
****

Veeramakali Vignesh Manivannan,Yasaman Jafari,Srikar Eranky,Spencer Ho,Rose Yu,Duncan Watson-Parris,Yian Ma,Leon Bergen,Taylor Berg-Kirkpatrick
ClimaQA: An Automated Evaluation Framework for Climate Question Answering Models
The use of Large Language Models (LLMs) in climate science has recently gained significant attention. However, a critical issue remains: the lack of a comprehensive evaluation framework capable of assessing the quality and scientific validity of model outputs. To address this issue, we develop *ClimaGen* (Climate QA Generator), an adaptive learning framework that generates question-answer pairs from graduate textbooks with climate scientists in the loop. As a result, we present *ClimaQA-Gold*, an expert-annotated benchmark dataset alongside *ClimaQA-Silver*, a large-scale, comprehensive synthetic QA dataset for climate science. Finally, we develop evaluation strategies and compare different LLMs on our benchmarks. Our results offer novel insights into various approaches used to enhance knowledge of climate LLMs. ClimaQA’s source code is publicly available at https://github.com/Rose-STL-Lab/genie-climaqa
****

Chengyue Huang,Junjiao Tian,Brisa Maneechotesuwan,Shivang Chopra,Zsolt Kira
Directional Gradient Projection for Robust Fine-Tuning of Foundation Models
Robust fine-tuning aims to adapt large foundation models to downstream tasks while preserving their robustness to distribution shifts. Existing methods primarily focus on constraining and projecting current model towards the pre-trained initialization based on the magnitudes between fine-tuned and pre-trained weights, which often require extensive hyper-parameter tuning and can sometimes result in underfitting. In this work, we propose $\textbf{Di}$rectional $\textbf{Gra}$dient $\textbf{P}$rojection (DiGraP), a novel layer-wise trainable method that incorporates directional information from gradients to bridge regularization and mul

ti-objective optimization. Besides demonstrating our method on image classificat
ion, as another contribution we generalize this area to the multi-modal evaluati
on settings for robust fine-tuning. Specifically, we first bridge the uni-modal
and multi-modal gap by performing analysis on Image Classification reformulated
Visual Question Answering (VQA) benchmarks and further categorize ten out-of-dis
tribution (OOD) VQA datasets by distribution shift types and degree (i.e. near v
ersus far OOD). Experimental results show that DiGraP consistently outperforms e
xisting baselines across Image Classfication and VQA tasks with discriminative a
nd generative backbones, improving both in-distribution (ID) generalization and
OOD robustness.
****
YUCHENG LI,Huiqiang Jiang,Qianhui Wu,Xufang Luo,Surin Ahn,Chengruidong Zhang,Ami
r H. Abdi,Dongsheng Li,Jianfeng Gao,Yuqing Yang,Lili Qiu
SCBench: A KV Cache-Centric Analysis of Long-Context Methods
Long-context Large Language Models (LLMs) have enabled numerous downstream appli
cations but also introduced significant challenges related to computational and
memory efficiency. To address these challenges, optimizations for long-context i
nference have been developed, centered around the KV cache. However, existing be
nchmarks often evaluate in single-request, neglecting the full lifecycle of the
KV cache in real-world use. This oversight is particularly critical, as KV cache
 reuse has become widely adopted in LLMs inference frameworks, such as vLLM and
SGLang, as well as by LLM providers, including OpenAI, Microsoft, Google, and An
thropic. To address this gap, we introduce SCBENCH (SharedContextBENCH), a compr
ehensive benchmark for evaluating long-context methods from a KV cache centric p
erspective: 1) KV cache generation, 2) KV cache compression, 3) KV cache retriev
al, and 4) KV cache loading. Specifically, SCBench uses test examples with share
d context, ranging 12 tasks with two shared context modes, covering four categor
ies of long-context capabilities: string retrieval, semantic retrieval, global i
nformation, and multi-task. With SCBench, we provide an extensive KV cache-centr
ic analysis of eight categories long-context solutions, including Gated Linear R
NNs (Codestal-Mamba), Mamba-Attention hybrids (Jamba-1.5-Mini), and efficient me
thods such as sparse attention, KV cache dropping, quantization, retrieval, load
ing, and prompt compression. The evaluation is conducted on six Transformer-base
d long-context LLMs: Llama-3.1-8B/70B, Qwen2.5-72B/32B, Llama-3-8B-262K, and GLM
-4-9B. Our findings show that sub-O(n) memory methods suffer in multi-turn scena
rios, while sparse encoding with O(n) memory and sub-O(n^2) pre-filling computat
ion perform robustly. Dynamic sparsity yields more expressive KV caches than sta
tic patterns, and layer-level sparsity in hybrid architectures reduces memory us
age with strong performance. Additionally, we identify attention distribution sh
ift issues in long-generation scenarios.
****
Tianqi Chen,Shujian Zhang,Mingyuan Zhou
Score Forgetting Distillation: A Swift, Data-Free Method for Machine Unlearning
in Diffusion Models
The machine learning community is increasingly recognizing the importance of fos
tering trust and safety in modern generative AI (GenAI) models. We posit machine
 unlearning (MU) as a crucial foundation for developing safe, secure, and trustw
orthy GenAI models. Traditional MU methods often rely on stringent assumptions a
nd require access to real data. This paper introduces Score Forgetting Distillat
ion (SFD), an innovative MU approach that promotes the forgetting of undesirable
 information in diffusion models by aligning the conditional scores of "unsafe"
classes or concepts with those of "safe" ones. To eliminate the need for real da
ta, our SFD framework incorporates a score-based MU loss into the score distilla
tion objective of a pretrained diffusion model. This serves as a regularization
term that preserves desired generation capabilities while enabling the productio
n of synthetic data through a one-step generator. Our experiments on pretrained
label-conditional and text-to-image diffusion models demonstrate that our method
 effectively accelerates the forgetting of target classes or concepts during gen
eration, while preserving the quality of other classes or concepts. This unlearn
ed and distilled diffusion not only pioneers a novel concept in MU but also acce

lerates the generation speed of diffusion models. Our experiments and studies on a range of diffusion models and datasets confirm that our approach is generalizable, effective, and advantageous for MU in diffusion models. Code is available at https://github.com/tqch/score-forgetting-distillation. (**Warning:** This paper contains sexually explicit imagery, discussions of pornography, racially-charged terminology, and other content that some readers may find disturbing, distressing, and/or offensive.)
****
Li Ju,Xingyi Yang,Qi Li,Xinchao Wang
GraphBridge: Towards Arbitrary Transfer Learning in GNNs
Graph neural networks (GNNs) are conventionally trained on a per-domain, per-task basis. It creates a significant barrier in transferring the acquired knowledge to different, heterogeneous data setups. This paper introduces **GraphBridge**, a novel framework to enable knowledge transfer across disparate tasks and domains in GNNs, circumventing the need for modifications to task configurations or graph structures. Specifically, GraphBridge allows for the augmentation of any pre-trained GNN with prediction heads and a bridging network that connects the input to the output layer. This architecture not only preserves the intrinsic knowledge of the original model but also supports outputs of arbitrary dimensions. To mitigate the negative transfer problem, GraphBridge merges the source model with a concurrently trained model, thereby reducing the source bias when applied to the target domain. Our method is thoroughly evaluated across diverse transfer learning scenarios, including Graph2Graph, Node2Node, Graph2Node, and graph2point-cloud. Empirical validation, conducted over 16 datasets representative of these scenarios, confirms the framework's capacity for task- and domain-agnostic transfer learning within graph-like data, marking a significant advancement in the field of GNNs. Code is available at https://github.com/jujulili888/GraphBridge.
****
Abdelrahman Eldesokey,Peter Wonka
Build-A-Scene: Interactive 3D Layout Control for Diffusion-Based Image Generation
We propose a diffusion-based approach for Text-to-Image (T2I) generation with interactive 3D layout control.
Layout control has been widely studied to alleviate the shortcomings of T2I diffusion models in understanding objects' placement and relationships from text descriptions.
Nevertheless, existing approaches for layout control are limited to 2D layouts, require the user to provide a static layout beforehand, and fail to preserve generated images under layout changes.
This makes these approaches unsuitable for applications that require 3D object-wise control and iterative refinements, e.g., interior design and complex scene generation.
To this end, we leverage the recent advancements in depth-conditioned T2I models and propose a novel approach for interactive 3D layout control.
We replace the traditional 2D boxes used in layout control with 3D boxes.
Furthermore, we revamp the T2I task as a multi-stage generation process, where at each stage, the user can insert, change, and move an object in 3D while preserving objects from earlier stages.
We achieve this through a novel Dynamic Self-Attention (DSA) module and a consistent 3D object translation strategy.
To evaluate our approach, we establish a benchmark and an evaluation protocol for interactive 3D layout control.
Experiments show that our approach can generate complicated scenes based on 3D layouts, outperforming the standard depth-conditioned T2I methods by two-folds on object generation success rate.
Moreover, it outperforms all methods in comparison on preserving objects under layout changes.
Project Page: https://abdo-eldesokey.github.io/build-a-scene/
****

Jonathan Light,Min Cai,Weiqin Chen,Guanzhi Wang,Xiusi Chen,Wei Cheng,Yisong Yue, Ziniu Hu
Strategist: Self-improvement of LLM Decision Making via Bi-Level Tree Search
Traditional reinforcement learning and planning require a lot of data and training to develop effective strategies. On the other hand, large language models (LLMs) can generalize well and perform tasks without prior training but struggle with complex planning and decision-making. We introduce **STRATEGIST**, a new approach that combines the strengths of both methods. It uses LLMs to generate and update high-level strategies in text form, while a Monte Carlo Tree Search (MCTS) algorithm refines and executes them. STRATEGIST is a general framework that optimizes strategies through self-play simulations without requiring any training data. We test STRATEGIST in competitive, multi-turn games with partial information, such as **Game of Pure Strategy (GOPS)** and **The Resistance: Avalon**, a multi-agent hidden-identity discussion game. Our results show that STRATEGIST-based agents outperform traditional reinforcement learning models, other LLM-based methods, and existing LLM agents while achieving performance levels comparable to human players.
****
■ukasz Staniszewski,Bartosz Cywi■ski,Franziska Boenisch,Kamil Deja,Adam Dziedzic
Precise Parameter Localization for Textual Generation in Diffusion Models
Novel diffusion models can synthesize photo-realistic images with integrated high-quality text. Surprisingly, we demonstrate through
attention activation patching that only less than 1\% of diffusion models' parameters, all contained in attention layers, influence the generation of textual content within the images. Building on this observation, we improve textual generation efficiency and performance by targeting cross and joint attention layers of diffusion models. We introduce several applications that benefit from localizing the layers responsible for textual content generation. We first show that a LoRA-based fine-tuning solely of the localized layers enhances, even more, the general text-generation capabilities of large diffusion models while preserving the quality and diversity of the diffusion models' generations. Then, we demonstrate how we can use the localized layers to edit textual content in generated images. Finally, we extend this idea to the practical use case of preventing the generation of toxic text in a cost-free manner. In contrast to prior work, our localization approach is broadly applicable across various diffusion model architectures, including U-Net (e.g., SDXL and DeepFloyd IF) and transformer-based (e.g., Stable Diffusion 3), utilizing diverse text encoders (e.g., from CLIP to the large language models like T5). Project page available at https://t2i-text-loc.github.io/.
****
Jiangrong Shen,Qi Xu,Gang Pan,Badong Chen
Improving the Sparse Structure Learning of Spiking Neural Networks from the View of Compression Efficiency
The human brain utilizes spikes for information transmission and dynamically reorganizes its network structure to boost energy efficiency and cognitive capabilities throughout its lifespan. Drawing inspiration from this spike-based computation, Spiking Neural Networks (SNNs) have been developed to construct event-driven models that emulate this efficiency. Despite these advances, deep SNNs continue to suffer from over-parameterization during training and inference, a stark contrast to the brain’s ability to self-organize. Furthermore, existing sparse SNNs are challenged by maintaining optimal pruning levels due to a static pruning ratio, resulting in either under or over-pruning.
In this paper, we propose a novel two-stage dynamic structure learning approach for deep SNNs, aimed at maintaining effective sparse training from scratch while optimizing compression efficiency.
The first stage evaluates the compressibility of existing sparse subnetworks within SNNs using the PQ index, which facilitates an adaptive determination of the rewiring ratio for synaptic connections based on data compression insights. In the second stage, this rewiring ratio critically informs the dynamic synaptic connection rewiring process, including both pruning and regrowth. This approach sig

nificantly improves the exploration of sparse structures training in deep SNNs, adapting sparsity dynamically from the point view of compression efficiency. Our experiments demonstrate that this sparse training approach not only aligns w ith the performance of current deep SNNs models but also significantly improves the efficiency of compressing sparse SNNs. Crucially, it preserves the advantage s of initiating training with sparse models and offers a promising solution for implementing Edge AI on neuromorphic hardware.
****
Gangwei Jiang,Caigao JIANG,Zhaoyi Li,Siqiao Xue,JUN ZHOU,Linqi Song,Defu Lian,Yi ng Wei
Unlocking the Power of Function Vectors for Characterizing and Mitigating Catast rophic Forgetting in Continual Instruction Tuning
Catastrophic forgetting (CF) poses a significant challenge in machine learning, where a model forgets previously learned information upon learning new tasks. Despite the advanced capabilities of Large Language Models (LLMs), they continue to face challenges with CF during continual learning. The majority of existing research focuses on analyzing forgetting patterns through a singular training se quence, thereby overlooking the intricate effects that diverse tasks have on mod el behavior.
Our study explores CF across various settings, discovering that model forgetting is influenced by both the specific training tasks and the models themselves. To this end, we interpret forgetting by examining the function vector (FV), a comp act representation of functions in LLMs, offering a model-dependent indicator fo r the occurrence of CF. Through theoretical and empirical analyses, we demonstra ted that CF in LLMs primarily stems from biases in function activation rather th an the overwriting of task processing functions.
Leveraging these insights, we propose a novel function vector guided training me thodology, incorporating a regularization technique to stabilize the FV and miti gate forgetting. Empirical tests on four benchmarks confirm the effectiveness of our proposed training method, substantiating our theoretical framework concerni ng CF and model function dynamics.
****
Yuxin Wang,Xiaomeng Zhu,Weimin Lyu,Saeed Hassanpour,Soroush Vosoughi
ImpScore: A Learnable Metric For Quantifying The Implicitness Level of Sentences
Handling implicit language is essential for natural language processing systems to achieve precise text understanding and facilitate natural interactions with u sers. Despite its importance, the absence of a metric for accurately measuring t he implicitness of language significantly constrains the depth of analysis possi ble in evaluating models' comprehension capabilities. This paper addresses this gap by developing a scalar metric that quantifies the implicitness level of lang uage without relying on external references. Drawing on principles from traditio nal linguistics, we define "implicitness" as the divergence between semantic mea ning and pragmatic interpretation. To operationalize this definition, we introdu ce ImpScore, a reference-free metric formulated through an interpretable regress ion model. This model is trained using pairwise contrastive learning on a specia lly curated dataset consisting of (*implicit sentence*, *explicit sentence*) pai rs. We validate ImpScore through a user study that compares its assessments with human evaluations on out-of-distribution data, demonstrating its accuracy and s trong correlation with human judgments. Additionally, we apply ImpScore to hate speech detection datasets, illustrating its utility and highlighting significant limitations in current large language models' ability to understand highly impl icit content. Our metric is publicly available at https://github.com/audreycs/Im pScore.
****
Jing Wen,Alex Schwing,Shenlong Wang
LIFe-GoM: Generalizable Human Rendering with Learned Iterative Feedback Over Mul ti-Resolution Gaussians-on-Mesh
Generalizable rendering of an animatable human avatar from sparse inputs relies on data priors and inductive biases extracted from training on large data to avo id scene-specific optimization and to enable fast reconstruction. This raises tw

o main challenges: First, unlike iterative gradient-based adjustment in scene-sp
ecific optimization, generalizable methods must reconstruct the human shape repr
esentation in a single pass at inference time.
Second, rendering is preferably computationally efficient yet of high resolution
.
To address both challenges we augment the recently proposed
dual shape representation, which combines the benefits of a mesh and Gaussian po
ints, in two ways.
To improve reconstruction, we propose an iterative feedback update framework, wh
ich successively improves the canonical human shape representation during recons
truction.
To achieve computationally efficient yet high-resolution rendering, we study a c
oupled-multi-resolution Gaussians-on-Mesh representation.
We evaluate the proposed approach on the challenging THuman2.0, XHuman and AIST+
+ data. Our approach reconstructs an animatable representation from sparse inpu
ts in less than 1s, renders views with 95.1FPS at 1024×1024, and achiev
es PSNR/LPIPS*/FID of 24.65/110.82/51.27 on THuman2.0, outperforming the state-
of-the-art in rendering quality.
****
Yanzhe Chen,Zhiwen Yang,Jinglin Xu,Yuxin Peng
MAI: A Multi-turn Aggregation-Iteration Model for Composed Image Retrieval
Multi-Turn Composed Image Retrieval (MTCIR) addresses a real-world scenario wher
e users iteratively refine retrieval results by providing additional information
 until a target meeting all their requirements is found. Existing methods primar
ily achieve MTCIR through a "multiple single-turn" paradigm, wherein methods inc
orrectly converge on shortcuts that only utilize the most recent turn's image, i
gnoring attributes from historical turns. Consequently, retrieval failures occur
 when modification requests involve historical information. We argue that explic
itly incorporating historical information into the modified text is crucial to a
ddressing this issue. To this end, we build a new retrospective-based MTCIR data
set, **FashionMT**, wherein modification demands are highly associated with hist
orical turns. We also propose a Multi-turn Aggregation-Iteration (**MAI**) model
, emphasizing efficient aggregation of multimodal semantics and optimization of
information propagation in multi-turn retrieval. Specifically, we propose a new
Two-stage Semantic Aggregation (TSA) paradigm coupled with a Cyclic Combination
Loss (CCL), achieving improved semantic consistency and modality alignment by pr
ogressively interacting the reference image with its caption and the modified te
xt. In addition, we design a Multi-turn Iterative Optimization (MIO) mechanism t
hat dynamically selects representative tokens and reduces redundancy during mult
i-turn iterations. Extensive experiments demonstrate that the proposed MAI model
 achieves substantial improvements over state-of-the-art methods.
****
Mete Kemertas,Amir-massoud Farahmand,Allan Douglas Jepson
A Truncated Newton Method for Optimal Transport
Developing a contemporary optimal transport (OT) solver requires navigating trad
e-offs among several critical requirements: GPU parallelization, scalability to
high-dimensional problems, theoretical convergence guarantees, empirical perform
ance in terms of precision versus runtime, and numerical stability in practice.
With these challenges in mind, we introduce a specialized truncated Newton algor
ithm for entropic-regularized OT. In addition to proving that locally quadratic
convergence is possible without assuming a Lipschitz Hessian, we provide strateg
ies to maximally exploit the high rate of local convergence in practice. Our GPU
-parallel algorithm exhibits exceptionally favorable runtime performance, achiev
ing high precision orders of magnitude faster than many existing alternatives. T
his is evidenced by wall-clock time experiments on 24 problem sets (12 datasets
$\times$ 2 cost functions). The scalability of the algorithm is showcased on an
extremely large OT problem with $n \approx 10^6$, solved approximately under wea
k entopric regularization.
****
Honghui Yang,Di Huang,Wei Yin,Chunhua Shen,Haifeng Liu,Xiaofei He,Binbin Lin,Wan

li Ouyang,Tong He
Depth Any Video with Scalable Synthetic Data
Video depth estimation has long been hindered by the scarcity of consistent and scalable ground truth data, leading to inconsistent and unreliable results. In this paper, we introduce Depth Any Video, a model that tackles the challenge through two key innovations. First, we develop a scalable synthetic data pipeline, capturing real-time video depth data from diverse virtual environments, yielding 40,000 video clips of 5-second duration, each with precise depth annotations. Second, we leverage the powerful priors of generative video diffusion models to handle real-world videos effectively, integrating advanced techniques such as rotary position encoding and flow matching to further enhance flexibility and efficiency. Unlike previous models, which are limited to fixed-length video sequences, our approach introduces a novel mixed-duration training strategy that handles videos of varying lengths and performs robustly across different frame rates—even on single frames. At inference, we propose a depth interpolation method that enables our model to infer high-resolution video depth across sequences of up to 150 frames. Our model outperforms all previous generative depth models in terms of spatial accuracy and temporal consistency. The code and model weights are open-sourced.
****
Anh-Dung Dinh,Daochang Liu,Chang Xu
Representative Guidance: Diffusion Model Sampling with Coherence
The diffusion sampling process faces a persistent challenge stemming from its incoherence, attributable to varying noise directions across different timesteps.
Our Representative Guidance (RepG) offers a new perspective to handle this issue by reformulating the sampling process with a coherent direction towards a representative target.
In this formulation, while the classic classifier guidance improves feature discernment by steering the model away from ambiguous features, it fails to provide a favourable representative target since the class label is overly compact and leads to sacrificed diversity and the adversarial generation problem.
In contrast, we leverage self-supervised representations as the coherent target and treat sampling as a downstream task, which refines image details and corrects errors rather than settling for simpler samples.
Our representative guidance achieves superior performance and illustrates the potential of pre-trained self-supervised models in image sampling. Our findings demonstrate that RepG not only substantially enhances vanilla diffusion sampling but also surpasses state-of-the-art benchmarks when combined with classifier-free guidance.
****
Pengcheng Wang,Chenran Li,Catherine Weaver,Kenta Kawamoto,Masayoshi Tomizuka,Chen Tang,Wei Zhan
Residual-MPPI: Online Policy Customization for Continuous Control
Policies developed through Reinforcement Learning (RL) and Imitation Learning (IL) have shown great potential in continuous control tasks, but real-world applications often require adapting trained policies to unforeseen requirements. While fine-tuning can address such needs, it typically requires additional data and access to the original training metrics and parameters.
In contrast, an online planning algorithm, if capable of meeting the additional requirements, can eliminate the necessity for extensive training phases and customize the policy without knowledge of the original training scheme or task. In this work, we propose a generic online planning algorithm for customizing continuous-control policies at the execution time, which we call Residual-MPPI. It can customize a given prior policy on new performance metrics in few-shot and even zero-shot online settings, given access to the prior action distribution alone. Through our experiments, we demonstrate that the proposed Residual-MPPI algorithm can accomplish the few-shot/zero-shot online policy customization task effectively, including customizing the champion-level racing agent, Gran Turismo Sophy (GT Sophy) 1.0, in the challenging car racing scenario, Gran Turismo Sport (GTS) environment. Code for MuJoCo experiments is included in the supplementary and wi

ll be open-sourced upon acceptance. Demo videos are available on our website: ht
tps://sites.google.com/view/residual-mppi.
****
Ihor Neporozhnii,Julien Roy,Emmanuel Bengio,Jason Hartford
Efficient Biological Data Acquisition through Inference Set Design
In drug discovery, highly automated high-throughput laboratories are used to scr
een a large number of compounds in search of effective drugs. These experiments
are expensive, so one might hope to reduce their cost by only experimenting on a
 subset of the compounds, and predicting the outcomes of the remaining experimen
ts. In this work, we model this scenario as a sequential subset selection proble
m: we aim to select the smallest set of candidates in order to achieve some desi
red level of accuracy for the system as a whole. Our key observation is that, if
 there is heterogeneity in the difficulty of the prediction problem across the i
nput space, selectively obtaining the labels for the hardest examples in the acq
uisition pool will leave only the relatively easy examples to remain in the infe
rence set, leading to better overall system performance. We call this mechanism
inference set design, and propose the use of a confidence-based active learning
solution to prune out these challenging examples. Our algorithm includes an expl
icit stopping criterion that interrupts the acquisition loop when it is sufficie
ntly confident that the system has reached the target performance. Our empirical
 studies on image and molecular datasets, as well as a real-world large-scale bi
ological assay, show that active learning for inference set design leads to sign
ificant reduction in experimental cost while retaining high system performance.
****
Daouda Sow,Herbert Woisetschläger,Saikiran Bulusu,Shiqiang Wang,Hans Arno Jacobs
en,Yingbin Liang
Dynamic Loss-Based Sample Reweighting for Improved Large Language Model Pretrain
ing
Pretraining large language models (LLMs) on vast and heterogeneous datasets is c
rucial for achieving state-of-the-art performance across diverse downstream task
s. However, current training paradigms treat all samples equally, overlooking th
e importance or relevance of individual samples throughout the training process.
 Existing reweighting strategies, which primarily focus on group-level data impo
rtance, fail to leverage fine-grained instance-level information and do not adap
t dynamically to individual sample importance as training progresses. In this pa
per, we introduce novel algorithms for dynamic, instance-level data reweighting
aimed at improving both the efficiency and effectiveness of LLM pretraining. Our
 methods adjust the weight of each training sample based on its loss value in an
 online fashion, allowing the model to dynamically focus on more informative or
important samples at the current training stage. In particular, our framework al
lows us to systematically devise reweighting strategies deprioritizing redundant
 or uninformative data, which we find tend to work best.
Furthermore, we develop a new theoretical framework for analyzing the impact of
loss-based reweighting on the convergence of gradient-based optimization, provid
ing the first formal characterization of how these strategies affect convergence
 bounds. We empirically validate our approach across a spectrum of tasks, from p
retraining 7B and 1.4B parameter LLMs to smaller-scale language models and linea
r regression problems, demonstrating that our loss-based reweighting approach ca
n lead to faster convergence and significantly improved performance.
****
Zhengbo Wang,Jian Liang,Ran He,Zilei Wang,Tieniu Tan
LoRA-Pro: Are Low-Rank Adapters Properly Optimized?
Low-rank adaptation, also known as LoRA, has emerged as a prominent method for p
arameter-efficient fine-tuning of foundation models.
Despite its computational efficiency, LoRA still yields inferior performance com
pared to full fine-tuning.
In this paper, we first uncover a fundamental connection between the optimizatio
n processes of LoRA and full fine-tuning: using LoRA for optimization is mathema
tically equivalent to full fine-tuning using a low-rank gradient for parameter u
pdates.

And this low-rank gradient can be expressed in terms of the gradients of the two low-rank matrices in LoRA.
Leveraging this insight, we introduce LoRA-Pro, a method that enhances LoRA's performance by strategically adjusting the gradients of these low-rank matrices.
This adjustment allows the low-rank gradient to more accurately approximate the full fine-tuning gradient, thereby narrowing the performance gap between LoRA and full fine-tuning.
Furthermore, we theoretically derive the optimal solutions for adjusting the gradients of the low-rank matrices, applying them during fine-tuning in LoRA-Pro.
We conduct extensive experiments across natural language understanding, dialogue generation, mathematical reasoning, code generation, and image classification tasks, demonstrating that LoRA-Pro substantially improves LoRA's performance, effectively narrowing the gap with full fine-tuning.
Our code is publicly available at https://github.com/mrflogs/LoRA-Pro.
****

Peng Liu,Dongyang Dai,Zhiyong Wu
RFWave: Multi-band Rectified Flow for Audio Waveform Reconstruction
Recent advancements in generative modeling have significantly enhanced the reconstruction of audio waveforms from various representations. While diffusion models are adept at this task, they are hindered by latency issues due to their operation at the individual sample point level and the need for numerous sampling steps. In this study, we introduce RFWave, a cutting-edge multi-band Rectified Flow approach designed to reconstruct high-fidelity audio waveforms from Mel-spectrograms or discrete acoustic tokens. RFWave uniquely generates complex spectrograms and operates at the frame level, processing all subbands simultaneously to boost efficiency. Leveraging Rectified Flow, which targets a straight transport trajectory, RFWave achieves reconstruction with just 10 sampling steps. Our empirical evaluations show that RFWave not only provides outstanding reconstruction quality but also offers vastly superior computational efficiency, enabling audio generation at speeds up to 160 times faster than real-time on a GPU. Both an online demonstration and the source code are accessible.
****

Charles Dawson,Van Tran,Max Z. Li,Chuchu Fan
Rare event modeling with self-regularized normalizing flows: what can we learn from a single failure?
Increased deployment of autonomous systems in fields like transportation and robotics have seen a corresponding increase in safety-critical failures. These failures can be difficult to model and debug due to the relative lack of data: compared to tens of thousands of examples from normal operations, we may have only seconds of data leading up to the failure. This scarcity makes it challenging to train generative models of rare failure events, as existing methods risk either overfitting to noise in the limited failure dataset or underfitting due to an overly strong prior. We address this challenge with CalNF, or calibrated normalizing flows, a self-regularized framework for posterior learning from limited data.
CalNF achieves state-of-the-art performance on data-limited failure modeling and inverse problems and enables a first-of-a-kind case study into the root causes of the 2022 Southwest Airlines scheduling crisis.
****

Zhenyi Zhang,Tiejun Li,Peijie Zhou
Learning stochastic dynamics from snapshots through regularized unbalanced optimal transport
Reconstructing dynamics using samples from sparsely time-resolved snapshots is an important problem in both natural sciences and machine learning. Here, we introduce a new deep learning approach for solving regularized unbalanced optimal transport (RUOT) and inferring continuous unbalanced stochastic dynamics from observed snapshots. Based on the RUOT form, our method models these dynamics without requiring prior knowledge of growth and death processes or additional information, allowing them to be learned directly from data. Theoretically, we explore the connections between the RUOT and Schrödinger bridge problem and discuss the key challenges and potential solutions. The effectiveness of our method is demons

trated with a synthetic gene regulatory network, high-dimensional Gaussian Mixtu
re Model, and single-cell RNA-seq data from blood development. Compared with oth
er methods, our approach accurately identifies growth and transition patterns, e
liminates false transitions, and constructs the Waddington developmental landsca
pe. Our code is available at: [https://github.com/zhenyiizhang/DeepRUOT](https:/
/github.com/zhenyiizhang/DeepRUOT).
****

Khanh Nguyen,Raouf Kerkouche,Mario Fritz,Dimosthenis Karatzas
DocMIA: Document-Level Membership Inference Attacks against DocVQA Models
Document Visual Question Answering (DocVQA) has introduced a new paradigm for en
d-to-end document understanding, and quickly became one of the standard benchmar
ks for multimodal LLMs. Automating document processing workflows, driven by DocV
QA models, presents significant potential for many business sectors. However, do
cuments tend to contain highly sensitive information, raising concerns about pri
vacy risks associated with training such DocVQA models. One significant privacy
vulnerability, exploited by the membership inference attack, is the possibility
for an adversary to determine if a particular record was part of the model's tra
ining data. In this paper, we introduce two novel membership inference attacks t
ailored specifically to DocVQA models. These attacks are designed for two differ
ent adversarial scenarios: a white-box setting, where the attacker has full acce
ss to the model architecture and parameters, and a black-box setting, where only
 the model's outputs are available. Notably, our attacks assume the adversary la
cks access to auxiliary datasets, which is more realistic in practice but also m
ore challenging. Our unsupervised methods outperform existing state-of-the-art m
embership inference attacks across a variety of DocVQA models and datasets, demo
nstrating their effectiveness and highlighting the privacy risks in this domain.
****

Tyler A. Chang,Dheeraj Rajagopal,Tolga Bolukbasi,Lucas Dixon,Ian Tenney
Scalable Influence and Fact Tracing for Large Language Model Pretraining
Training data attribution (TDA) methods aim to attribute model outputs back to s
pecific training examples, and the application of these methods to large languag
e model (LLM) outputs could significantly advance model transparency and data cu
ration. However, it has been challenging to date to apply these methods to the f
ull scale of LLM pretraining. In this paper, we refine existing gradient-based m
ethods to work effectively at scale, allowing us to retrieve influential example
s for an 8B-parameter language model from a pretraining corpus of over 160B toke
ns with no need for subsampling or pre-filtering. Our method combines several te
chniques, including optimizer state correction, a task-specific Hessian approxim
ation, and normalized encodings, which we find to be critical for performance at
 scale. In quantitative evaluations on a fact tracing task, our method performs
best at identifying examples that influence model predictions, but classical, mo
del-agnostic retrieval methods such as BM25 still perform better at finding pass
ages which explicitly contain relevant facts. These results demonstrate a misali
gnment between factual *attribution* and causal *influence*. With increasing mod
el size and training tokens, we find that influence more closely aligns with fac
tual attribution. Finally, we examine different types of examples identified as
influential by our method, finding that while many directly entail a particular
fact, others support the same output by reinforcing priors on relation types, co
mmon entities, and names. We release our prompt set and model outputs, along wit
h a web-based visualization tool to explore influential examples for factual pre
dictions, commonsense reasoning, arithmetic, and open-ended generation for an 8B
-parameter LLM.
****

Zixuan Gong,Xiaolin Hu,Huayi Tang,Yong Liu
Towards Auto-Regressive Next-Token Prediction: In-context Learning Emerges from
Generalization
Large language models (LLMs) have demonstrated remarkable in-context learning (I
CL) abilities. However, existing theoretical analysis of ICL primarily exhibits
two limitations: \textbf{(a) Limited \textit{i.i.d.} Setting.} Most studies focu
s on supervised function learning tasks where prompts are constructed with \text

it{i.i.d.} input-label pairs. This \textit{i.i.d.} assumption diverges significa
ntly from real language learning scenarios where prompt tokens are interdependen
t. \textbf{(b) Lack of Emergence Explanation.} Most literature answers \textbf{\
textit{what}} ICL does from an implicit optimization perspective but falls short
 in elucidating \textbf{\textit{how}} ICL emerges and the impact of pre-training
 phase on ICL. In our paper, to extend (a), we adopt a more practical paradigm,
\textbf{\textit{auto-regressive next-token prediction (AR-NTP)}}, which closely
aligns with the actual training of language models. Specifically, within AR-NTP,
 we emphasize prompt token-dependency, which involves predicting each subsequent
 token based on the preceding sequence. To address (b), we formalize a systemati
c pre-training and ICL framework, highlighting the layer-wise structure of seque
nces and topics, alongside a two-level expectation. In conclusion, we present da
ta-dependent, topic-dependent and optimization-dependent PAC-Bayesian generaliza
tion bounds for pre-trained LLMs, investigating that \textbf{\textit{ICL emerges
 from the generalization of sequences and topics}}. Our theory is supported by e
xperiments on numerical linear dynamic systems, synthetic GINC and real-world la
nguage datasets.
****

Kaustubh Mani,Vincent Mai,Charlie Gauthier,Annie S Chen,Samer B. Nashed,Liam Pau
ll
Safety Representations for Safer Policy Learning
Reinforcement learning algorithms typically necessitate extensive exploration of
 the state space to find optimal policies. However, in safety-critical applicati
ons, the risks associated with such exploration can lead to catastrophic consequ
ences. Existing safe exploration methods attempt to mitigate this by imposing co
nstraints, which often result in overly conservative behaviours and inefficient
learning. Heavy penalties for early constraint violations can trap agents in loc
al optima, deterring exploration of risky yet high-reward regions of the state s
pace. To address this, we introduce a method that explicitly learns state-condit
ioned safety representations. By augmenting the state features with these safety
 representations, our approach naturally encourages safer exploration without be
ing excessively cautious, resulting in more efficient and safer policy learning
in safety-critical scenarios. Empirical evaluations across diverse environments
show that our method significantly improves task performance while reducing cons
traint violations during training, underscoring its effectiveness in balancing e
xploration with safety.
****

Michael T Pearce,Thomas Dooms,Alice Rigg,Jose Oramas,Lee Sharkey
Bilinear MLPs enable weight-based mechanistic interpretability
A mechanistic understanding of how MLPs do computation in deep neural net-
works remains elusive. Current interpretability work can extract features from
hidden activations over an input dataset but generally cannot explain how MLP
weights construct features. One challenge is that element-wise nonlinearities
introduce higher-order interactions and make it difficult to trace computations
through the MLP layer. In this paper, we analyze bilinear MLPs, a type of
Gated Linear Unit (GLU) without any element-wise nonlinearity that neverthe-
less achieves competitive performance. Bilinear MLPs can be fully expressed in
terms of linear operations using a third-order tensor, allowing flexible analysi
s of
the weights. Analyzing the spectra of bilinear MLP weights using eigendecom-
position reveals interpretable low-rank structure across toy tasks, image classi
fi-
cation, and language modeling. We use this understanding to craft adversarial
examples, uncover overfitting, and identify small language model circuits direct
ly
from the weights alone. Our results demonstrate that bilinear layers serve as an
interpretable drop-in replacement for current activation functions and that weig
ht-
based interpretability is viable for understanding deep-learning models.
****

Gabriele Corso,Vignesh Ram Somnath,Noah Getz,Regina Barzilay,Tommi Jaakkola,Andreas Krause
Composing Unbalanced Flows for Flexible Docking and Relaxation
Diffusion models have emerged as a successful approach for molecular docking, but they often cannot model protein flexibility or generate nonphysical poses. We argue that both these challenges can be tackled by framing the problem as a transport between distributions. Still, existing paradigms lack the flexibility to define effective maps between such complex distributions. To address this limitation we propose Unbalanced Flow Matching, a generalization of Flow Matching (FM) that allows trading off sample efficiency with approximation accuracy and enables more accurate transport. Empirically, we apply Unbalanced FM on flexible docking and structure relaxation, demonstrating our ability to model protein flexibility and generate energetically favorable poses. On the PDBBind docking benchmark, our method FlexDock improves the docking performance while increasing the proportion of energetically favorable poses from 30% to 73%.
****
Yi Cheng,Xiao Liang,Yeyun Gong,Wen Xiao,Song Wang,Yuji Zhang,Wenjun Hou,Kaishuai Xu,Wenge Liu,Wenjie Li,Jian Jiao,Qi Chen,Peng CHENG,Wayne Xiong
Integrative Decoding: Improving Factuality via Implicit Self-consistency
Self-consistency-based approaches, which involve repeatedly sampling multiple outputs and selecting the most consistent one as the final response, prove to be remarkably effective in improving the factual accuracy of large language models. Nonetheless, existing methods usually have strict constraints on the task format, largely limiting their applicability. In this paper, we present Integrative Decoding (ID), to unlock the potential of self-consistency in open-ended generation tasks. ID operates by constructing a set of inputs, each prepended with a previously sampled response, and then processes them concurrently, with the next token being selected by aggregating of all their corresponding predictions at each decoding step. In essence, this simple approach implicitly incorporates self-consistency in the decoding objective. Extensive evaluation shows that ID consistently enhances factuality over a wide range of language models, with substantial improvements on the TruthfulQA (+11.2%), Biographies (+15.4%) and LongFact (+8.5%) benchmarks. The performance gains amplify progressively as the number of sampled responses increases, indicating the potential of ID to scale up with repeated sampling.
****
Seong-Joon Park,Hee-Youl Kwak,Sang-Hyo Kim,Yongjune Kim,Jong-Seon No
CrossMPT: Cross-attention Message-passing Transformer for Error Correcting Codes
Error correcting codes (ECCs) are indispensable for reliable transmission in communication systems. Recent advancements in deep learning have catalyzed the exploration of ECC decoders based on neural networks. Among these, transformer-based neural decoders have achieved state-of-the-art decoding performance. In this paper, we propose a novel Cross-Attention Message-Passing Transformer (CrossMPT), which shares key operational principles with conventional message-passing decoders. While conventional transformer-based decoders employ a self-attention mechanism without distinguishing between magnitude and syndrome embeddings, CrossMPT updates these two types of embeddings separately and iteratively via two masked cross-attention blocks. The mask matrices are determined by the code's parity-check matrix, which explicitly captures and removes irrelevant relationships between the magnitude and syndrome embeddings. Our experimental results show that CrossMPT significantly outperforms existing neural network-based decoders for various code classes. Notably, CrossMPT achieves this decoding performance improvement while significantly reducing memory usage, computational complexity, inference time, and training time.
****
Sabrina Herbst,Sandeep Suresh Cranganore,Vincenzo De Maio,Ivona Brandi■
Exploring channel distinguishability in local neighborhoods of the model space in quantum neural networks
With the increasing interest in Quantum Machine Learning, Quantum Neural Networks (QNNs) have emerged and gained significant attention. These models have, howev

er, been shown to be notoriously difficult to train, which we hypothesize is par
tially due to the architectures, called ansatzes, that are hardly studied at thi
s point. Therefore, in this paper, we take a step back and analyze ansatzes. We
initially consider their expressivity, i.e., the space of operations they are ab
le to express, and show that the closeness to being a 2-design, the primarily us
ed measure, fails at capturing this property. Hence, we look for alternative way
s to characterize ansatzes, unrelated to expressivity, by considering the local
neighborhood of the model space, in particular, analyzing model distinguishabili
ty upon small perturbation of parameters. We derive an upper bound on their dist
inguishability, showcasing that QNNs using the Hardware Efficient Ansatz with fe
w parameters are hardly discriminable upon update. Our numerical experiments sup
port our bounds and further indicate that there is a significant degree of varia
bility, which stresses the need for warm-starting or clever initialization. Alto
gether, our work provides an ansatz-centric perspective on training dynamics and
 difficulties in QNNs, ultimately suggesting that iterative training of small qu
antum models may not be effective, which contrasts their initial motivation.
****

Yili Wang,Yixin Liu,Xu Shen,Chenyu Li,Rui Miao,Kaize Ding,Ying Wang,Shirui Pan,X
in Wang

Unifying Unsupervised Graph-Level Anomaly Detection and Out-of-Distribution Dete
ction: A Benchmark

To build safe and reliable graph machine learning systems, unsupervised graph-le
vel anomaly detection (GLAD) and unsupervised graph-level out-of-distribution (O
OD) detection (GLOD) have received significant attention in recent years. Though
 these two lines of research share the same objective, they have been studied in
dependently in the community due to distinct evaluation setups, creating a gap t
hat hinders the application and evaluation of methods from one to the other. To
bridge the gap, in this work, we present a Unified Benchmark for unsupervised Gr
aph-level OOD and anomaly Detection (UB-GOLD), a comprehensive evaluation framew
ork that unifies GLAD and GLOD under the concept of generalized graph-level OOD
detection. Our benchmark encompasses 35 datasets spanning four practical anomaly
 and OOD detection scenarios, facilitating the comparison of 18 representative G
LAD/GLOD methods. We conduct multi-dimensional analyses to explore the effective
ness, generalizability, robustness, and efficiency of existing methods, shedding
 light on their strengths and limitations. Furthermore, we provide an open-sourc
e codebase of UB-GOLD to foster reproducible research and outline potential dire
ctions for future investigations based on our insights.
****

Celia Rubio-Madrigal,Adarsh Jamadandi,Rebekka Burkholz

GNNs Getting ComFy: Community and Feature Similarity Guided Rewiring

Maximizing the spectral gap through graph rewiring has been proposed to enhance
the performance of message-passing graph neural networks (GNNs) by addressing ov
er-squashing. However, as we show, minimizing the spectral gap can also improve
generalization. To explain this, we analyze how rewiring can benefit GNNs within
 the context of stochastic block models. Since spectral gap optimization primari
ly influences community strength, it improves performance when the community str
ucture aligns with node labels. Building on this insight, we propose three disti
nct rewiring strategies that explicitly target community structure, node labels,
 and their alignment: (a) community structure-based rewiring (ComMa), a more com
putationally efficient alternative to spectral gap optimization that achieves si
milar goals; (b) feature similarity-based rewiring (FeaSt), which focuses on max
imizing global homophily; and (c) a hybrid approach (ComFy), which enhances loca
l feature similarity while preserving community structure to optimize label-comm
unity alignment. Extensive experiments confirm the effectiveness of these strate
gies and support our theoretical insights.
****

Satoki Ishikawa,Rio Yokota,Ryo Karakida

Local Loss Optimization in the Infinite Width: Stable Parameterization of Predic
tive Coding Networks and Target Propagation

Local learning, which trains a network through layer-wise local targets and loss

es, has been studied as an alternative to backpropagation (BP) in neural computa
tion. However, its algorithms often become more complex or require additional hy
perparameters due to the locality, making it challenging to identify desirable s
ettings where the algorithm progresses in a stable manner.
To provide theoretical and quantitative insights, we introduce maximal update p
arameterization (μP) in the infinite-width limit for two representative desi
gns of local targets: predictive coding (PC) and target propagation (TP). We ver
ify that μP enables hyperparameter transfer across models of different width
s.
Furthermore, our analysis reveals unique and intriguing properties of μP tha
t are not present in conventional BP. By analyzing deep linear networks, we find
that PC's gradients interpolate between first-order and Gauss-Newton-like gradi
ents, depending on the parameterization.
We demonstrate that, in specific standard settings, PC in the infinite-width lim
it behaves more similarly to the first-order gradient.
For TP, even with the standard scaling of the last layer differing from classica
l μP, its local loss optimization favors the feature learning regime over th
e kernel regime.
****
Qi Zhang,Yifei Wang,Jingyi Cui,Xiang Pan,Qi Lei,Stefanie Jegelka,Yisen Wang
Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustnes
s
Deep learning models often suffer from a lack of interpretability due to \emph{p
olysemanticity}, where individual neurons are activated by multiple unrelated se
mantics, resulting in unclear attributions of model behavior. Recent advances in
\emph{monosemanticity}, where neurons correspond to consistent and distinct sem
antics, have significantly improved interpretability but are commonly believed t
o compromise accuracy. In this work, we challenge the prevailing belief of the a
ccuracy-interpretability tradeoff, showing that monosemantic features not only e
nhance interpretability but also bring concrete gains in model performance of {\
color{black} robustness-related tasks}. Across multiple robust learning scenario
s—including input and label noise, few-shot learning, and out-of-domain generali
zation—our results show that models leveraging monosemantic features significant
ly outperform those relying on polysemantic features. Furthermore, we provide em
pirical and theoretical understandings on the robustness gains of feature monose
manticity. Our preliminary analysis suggests that monosemanticity, by promoting
better separation of feature representations, leads to more robust decision boun
daries {\color{black} under noise}. This diverse evidence highlights the \textbf
{generality} of monosemanticity in improving model robustness. As a first step i
n this new direction, we embark on exploring the learning benefits of monosemant
icity beyond interpretability, supporting the long-standing hypothesis of linkin
g interpretability and robustness. Code is available at \url{https://github.com/
PKU-ML/Monosemanticity-Robustness}.
****
Atsunobu Kotani,Ren Ng
A Computational Framework for Modeling Emergence of Color Vision in the Human Br
ain
It is a mystery how the brain decodes color vision purely from the optic nerve s
ignals it receives, with a core inferential challenge being how it disentangles
internal perception with the correct color dimensionality from the unknown encod
ing properties of the eye.
In this paper, we introduce a computational framework for modeling this emergenc
e of human color vision by simulating both the eye and the cortex. Existing rese
arch often overlooks how the cortex develops color vision or represents color sp
ace internally, assuming that the color dimensionality is known a priori; howeve
r, we argue that the visual cortex has the capability and the challenge of infer
ring the color dimensionality purely from fluctuations in the optic nerve signal
s. To validate our theory, we introduce a simulation engine for biological eyes
based on established vision science and generate optic nerve signals resulting f
rom looking at natural images. Further, we propose a bio-plausible model of cort

ical learning based on self-supervised prediction of optic nerve signal fluctuat
ions under natural eye motions. We show that this model naturally learns to gene
rate color vision by disentangling retinal invariants from the sensory signals.
When the retina contains N types of color photoreceptors, our simulation shows
 that N-dimensional color vision naturally emerges, verified through formal co
lorimetry. Using this framework, we also present the first simulation work that
successfully boosts the color dimensionality, as observed in gene therapy on squ
irrel monkeys, and demonstrates the possibility of enhancing human color vision
from 3D to 4D.
****

Arne Schneuing,Ilia Igashov,Adrian W. Dobbelstein,Thomas Castiglione,Michael M.
Bronstein,Bruno Correia
Multi-domain Distribution Learning for De Novo Drug Design
We introduce DrugFlow, a generative model for structure-based drug design that i
ntegrates continuous flow matching with discrete Markov bridges, demonstrating s
tate-of-the-art performance in learning chemical, geometric, and physical aspect
s of three-dimensional protein-ligand data. We endow DrugFlow with an uncertaint
y estimate that is able to detect out-of-distribution samples. To further enhanc
e the sampling process towards distribution regions with desirable metric values
, we propose a joint preference alignment scheme applicable to both flow matchin
g and Markov bridge frameworks. Furthermore, we extend our model to also explore
 the conformational landscape of the protein by jointly sampling side chain angl
es and molecules.
****

Xinyu Tian,Shu Zou,Zhaoyuan Yang,Mengqi He,Jing Zhang
Black Sheep in the Herd: Playing with Spuriously Correlated Attributes for Visio
n-Language Recognition
Few-shot adaptation for Vision-Language Models (VLMs) presents a dilemma: balanc
ing in-distribution accuracy with out-of-distribution generalization. Recent res
earch has utilized low-level concepts such as visual attributes to enhance gener
alization. However, this study reveals that VLMs overly rely on a small subset o
f attributes on decision-making, which co-occur with the category but are not in
herently part of it, termed spuriously correlated attributes. This biased nature
 of VLMs results in poor generalization. To address this, 1) we first propose Sp
urious Attribute Probing (SAP), identifying and filtering out these problematic
attributes to significantly enhance the generalization of existing attribute-bas
ed methods; 2) We introduce Spurious Attribute Shielding (SAS), a plug-and-play
module that mitigates the influence of these attributes on prediction, seamlessl
y integrating into various Parameter-Efficient Fine-Tuning (PEFT) methods. In ex
periments, SAP and SAS significantly enhance accuracy on distribution shifts acr
oss 11 datasets and 3 generalization tasks without compromising downstream perfo
rmance, establishing a new state-of-the-art benchmark.
****

Xiao Li,Wenxuan Sun,Huanran Chen,Qiongxiu Li,Yingzhe He,Jie Shi,Xiaolin Hu
ADBM: Adversarial Diffusion Bridge Model for Reliable Adversarial Purification
Recently Diffusion-based Purification (DiffPure) has been recognized as an effec
tive defense method against adversarial examples. However, we find DiffPure whic
h directly employs the original pre-trained diffusion models for adversarial pur
ification, to be suboptimal. This is due to an inherent trade-off between noise
purification performance and data recovery quality. Additionally, the reliabilit
y of existing evaluations for DiffPure is questionable, as they rely on weak ada
ptive attacks. In this work, we propose a novel Adversarial Diffusion Bridge Mod
el, termed ADBM. ADBM directly constructs a reverse bridge from the diffused adv
ersarial data back to its original clean examples, enhancing the purification ca
pabilities of the original diffusion models. Through theoretical analysis and ex
perimental validation across various scenarios, ADBM has proven to be a superior
 and robust defense mechanism, offering significant promise for practical applic
ations. Code is available at https://github.com/LixiaoTHU/ADBM.
****

Artem M Vysogorets,Kartik Ahuja,Julia Kempe

DRoP: Distributionally Robust Data Pruning
In the era of exceptionally data-hungry models, careful selection of the trainin g data is essential to mitigate the extensive costs of deep learning. Data pruni ng offers a solution by removing redundant or uninformative samples from the dat aset, which yields faster convergence and improved neural scaling laws. However, little is known about its impact on classification bias of the trained models. We conduct the first systematic study of this effect and reveal that existing da ta pruning algorithms can produce highly biased classifiers. We present theoreti cal analysis of the classification risk in a mixture of Gaussians to argue that choosing appropriate class pruning ratios, coupled with random pruning within cl asses has potential to improve worst-class performance. We thus propose DRoP, a distributionally robust approach to pruning and empirically demonstrate its perf ormance on standard computer vision benchmarks. In sharp contrast to existing al gorithms, our proposed method continues improving distributional robustness at a tolerable drop of average performance as we prune more from the datasets.
****
Ziqi Wang,Hanlin Zhang,Xiner Li,Kuan-Hao Huang,Chi Han,Shuiwang Ji,Sham M. Kakad e,Hao Peng,Heng Ji
Eliminating Position Bias of Language Models: A Mechanistic Approach
Position bias has proven to be a prevalent issue of modern language models (LMs) , where the models prioritize content based on its position within the given con text. This bias often leads to unexpected model failures and hurts performance, robustness, and reliability across various applications. A simple mechanistic an alysis attributes the position bias to two components employed in nearly all sta te-of-the-art LMs: causal attention and position embedding. Based on the analyse s, we propose to **eliminate** position bias (e.g., different retrieved document s' orders in QA affect performance) with a **training-free zero-shot** approach. Our method changes the causal attention to bidirectional attention between docu ments and utilizes model attention values to decide the relative orders of docum ents instead of using the order provided in input prompts, therefore enabling Po sition-INvariant inferencE (PINE) at the document level. By eliminating position bias, models achieve better performance and reliability in downstream tasks, in cluding LM-as-a-judge, retrieval-augmented QA, molecule generation, and math rea soning. Notably, PINE is especially useful when adapting LMs for evaluating reas oning pairs: it consistently provides 8 to 10 percentage points performance gains, making Llama-3-70B-Instruct perform even better than GPT-4-0125-preview a nd GPT-4o-2024-08-06 on the RewardBench reasoning set.
****
Gen Zhou,Sugitha Janarthanan,Yutong Lu,Pingzhao Hu
CL-MFAP: A Contrastive Learning-Based Multimodal Foundation Model for Molecular Property Prediction and Antibiotic Screening
Due to the rise in antimicrobial resistance, identifying novel compounds with an tibiotic potential is crucial for combatting this global health issue. However, traditional drug development methods are costly and inefficient. Recognizing the pressing need for more effective solutions, researchers have turned to machine learning techniques to streamline the prediction and development of novel antibi otic compounds. While foundation models have shown promise in antibiotic discove ry, current mainstream efforts still fall short of fully leveraging the potentia l of multimodal molecular data. Recent studies suggest that contrastive learning frameworks utilizing multimodal data exhibit excellent performance in represent ation learning across various domains. Building upon this, we introduce CL-MFAP, an unsupervised contrastive learning (CL)-based multimodal foundation (MF) mode l specifically tailored for discovering small molecules with potential antibioti c properties (AP) using three types of molecular data. This model employs 1.6 mi llion bioactive molecules with drug-like properties from the ChEMBL dataset to j ointly pretrain three encoders: (1) a transformer-based encoder with rotary posi tion embedding for processing SMILES strings; (2) another transformer-based enco der, incorporating a novel bi-level routing attention mechanism to handle molecu lar graph representations; and (3) a Morgan fingerprint encoder using a multilay er perceptron, to achieve the contrastive learning purpose. The CL-MFAP outperfo

rms baseline models in antibiotic property prediction by effectively utilizing different molecular modalities and demonstrates superior domain-specific performance when fine-tuned for antibiotic-related property prediction tasks.
****

Eric Ricardo Anschuetz
A Unified Theory of Quantum Neural Network Loss Landscapes
Classical neural networks with random initialization famously behave as Gaussian processes in the limit of many neurons, which allows one to completely characterize their training and generalization behavior. No such general understanding exists for quantum neural networks (QNNs), which—outside of certain special cases—are known to not behave as Gaussian processes when randomly initialized. We here prove that QNNs and their first two derivatives instead generally form what we call "Wishart processes," where certain algebraic properties of the network determine the hyperparameters of the process. This Wishart process description allows us to, for the first time: give necessary and sufficient conditions for a QNN architecture to have a Gaussian process limit; calculate the full gradient distribution, generalizing previously known barren plateau results; and calculate the local minima distribution of algebraically constrained QNNs. Our unified framework suggests a certain simple operational definition for the "trainability" of a given QNN model using a newly introduced, experimentally accessible quantity we call the "degrees of freedom" of the network architecture.
****

Chenyang Zhang,Xuran Meng,Yuan Cao
Transformer Learns Optimal Variable Selection in Group-Sparse Classification
Transformers have demonstrated remarkable success across various applications. However, the success of transformers have not been understood in theory. In this work, we give a case study of how transformers can be trained to learn a classic statistical model with "group sparsity", where the input variables form multiple groups, and the label only depends on the variables from one of the groups. We theoretically demonstrate that, a one-layer transformer trained by gradient descent can correctly leverage the attention mechanism to select variables, disregarding irrelevant ones and focusing on those beneficial for classification. We also demonstrate that a well-pretrained one-layer transformer can be adapted to new downstream tasks to achieve good prediction accuracy with a limited number of samples. Our study sheds light on how transformers effectively learn structured data.
****

Kennard Chan,Fayao Liu,Guosheng Lin,Chuan-Sheng Foo,Weisi Lin
Robust-PIFu: Robust Pixel-aligned Implicit Function for 3D Human Digitalization from a Single Image
Existing methods for 3D clothed human digitalization perform well when the input image is captured in ideal conditions that assume the lack of any occlusion. However, in reality, images may often have occlusion problems such as incomplete observation of the human subject's full body, self-occlusion by the human subject, and non-frontal body pose. When given such input images, these existing methods fail to perform adequately. Thus, we propose Robust-PIFu, a pixel-aligned implicit model that capitalized on large-scale, pretrained latent diffusion models to address the challenge of digitalizing human subjects from non-ideal images that suffer from occlusions.

Robust-PIfu offers four new contributions. Firstly, we propose a 'disentangling' latent diffusion model. This diffusion model, pretrained on billions of images, takes in any input image and removes external occlusions, such as inter-person occlusions, from that image. Secondly, Robust-PIFu addresses internal occlusions like self-occlusion by introducing a `penetrating' latent diffusion model. This diffusion model outputs multi-layered normal maps that by-pass occlusions caused by the human subject's own limbs or other body parts (i.e. self-occlusion). Thirdly, in order to incorporate such multi-layered normal maps into a pixel-aligned implicit model, we introduce our Layered-Normals Pixel-aligned Implicit Model, which improves the structural accuracy of predicted clothed human meshes. Last

ly, Robust-PIFu proposes an optional super-resolution mechanism for the multi-la
yered normal maps. This addresses scenarios where the input image is of low or i
nadequate resolution. Though not strictly related to occlusion, this is still an
 important subproblem. Our experiments show that Robust-PIFu outperforms current
 SOTA methods both qualitatively and quantitatively. Our code will be released t
o the public.
****
Luxi He,Yangsibo Huang,Weijia Shi,Tinghao Xie,Haotian Liu,Yue Wang,Luke Zettlemo
yer,Chiyuan Zhang,Danqi Chen,Peter Henderson
Fantastic Copyrighted Beasts and How (Not) to Generate Them
Recent studies show that image and video generation models can be prompted to re
produce copyrighted content from their training data, raising serious legal con-
 cerns about copyright infringement. Copyrighted characters (e.g., Mario, Batman
) present a significant challenge: at least one lawsuit has already awarded dama
ges based on the generation of such characters. Consequently, commercial service
s like DALL·E have started deploying interventions. However, little research has
 systematically examined these problems: (1) Can users easily prompt models to g
enerate copyrighted characters, even if it is unintentional?; (2) How effective
are the existing mitigation strategies? To address these questions, we introduce
 a novel evaluation framework with metrics that assess both the generated image'
s similarity to copyrighted characters and its consistency with user intent, gro
unded in a set of popular copyrighted characters from diverse studios and region
s. We show that state-of-the-art image and video generation models can still gen
erate characters even if characters' names are not explicitly mentioned, sometim
es with only two generic keywords (e.g., prompting with "videogame, plumber" con
sistently gener- ates Nintendo's Mario character). We also introduce semi-automa
tic techniques to identify such keywords or descriptions that trigger character
generation. Using this framework, we evaluate mitigation strategies, including p
rompt rewriting and new approaches we propose. Our findings reveal that common m
ethods, such as DALL·E's prompt rewriting, are insufficient alone and require su
pplementary strategies like negative prompting. Our work provides empirical grou
nding for discussions on copyright mitigation strategies and offers actionable i
nsights for model deployers implementing these safeguards.
****
Giordano Cicchetti,Eleonora Grassucci,Luigi Sigillo,Danilo Comminiello
Gramian Multimodal Representation Learning and Alignment
Human perception integrates multiple modalities—such as vision, hearing, and lan
guage—into a unified understanding of the surrounding reality. While recent mult
imodal models have achieved significant progress by aligning pairs of modalities
 via contrastive learning, their solutions are unsuitable when scaling to multip
le modalities. These models typically align each modality to a designated anchor
 without ensuring the alignment of all modalities with each other, leading to su
boptimal performance in tasks requiring a joint understanding of multiple modali
ties. In this paper, we structurally rethink the pairwise conventional approach
to multimodal learning and we present the novel Gramian Representation Alignment
 Measure (GRAM), which overcomes the above-mentioned limitations. GRAM learns an
d then aligns n modalities directly in the higher-dimensional space in which m
odality embeddings lie by minimizing the Gramian volume of the k-dimensional p
arallelotope spanned by the modality vectors, ensuring the geometric alignment o
f all modalities simultaneously. GRAM can replace cosine similarity in any downs
tream method, holding for 2 to n modalities and providing more meaningful alig
nment with respect to previous similarity measures. The novel GRAM-based contras
tive loss function enhances the alignment of multimodal models in the higher-dim
ensional embedding space, leading to new state-of-the-art performance in downstr
eam tasks such as video-audio-text retrieval and audio-video classification.
****
Pengwei Tang,Xiaolin Hu,Yong Liu
ADePT: Adaptive Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning
Prompt Tuning (PT) enables the adaptation of Pre-trained Large Language Models (
PLMs) to downstream tasks by optimizing a small amount of soft virtual tokens, w

hich are prepended to the input token embeddings. Recently, Decomposed Prompt Tuning (DePT) has demonstrated superior adaptation capabilities by decomposing the soft prompt into a shorter soft prompt and a pair of low-rank matrices. The product of the pair of low-rank matrices is added to the input token embeddings to offset them. Additionally, DePT achieves faster inference compared to PT due to the shorter soft prompt. However, in this paper, we find that the position-based token embedding offsets of DePT restricts its ability to generalize across diverse model inputs, and that the shared embedding offsets across many token embeddings result in sub-optimization. To tackle these issues, we introduce \textbf{A}daptive \textbf{De}composed \textbf{P}rompt \textbf{T}uning (ADePT), which is composed of a short soft prompt and a shallow token-shared feed-forward neural network. ADePT utilizes the token-shared feed-forward neural network to learn the embedding offsets for each token, enabling adaptive embedding offsets that vary according to the model input and better optimization of token embedding offsets. This enables ADePT to achieve superior adaptation performance without requiring more inference time or additional trainable parameters compared to vanilla PT and its variants. In comprehensive experiments across 23 natural language processing tasks and 4 typical PLMs of different scales, ADePT consistently surpasses the leading parameter-efficient fine-tuning methods, and even outperforms the full fine-tuning in certain scenarios. We also provide a theoretical analysis towards ADePT. Code is available at https://github.com/HungerPWAY/ADePT.
****
Xin Mao,Huimin Xu,Feng-Lin Li,Ziqi Jin,WANG CHEN,Wei Zhang,Anh Tuan Luu
As Simple as Fine-tuning: LLM Alignment via Bidirectional Negative Feedback Loss
Direct Preference Optimization (DPO) has emerged as a more computationally efficient alternative to Reinforcement Learning from Human Feedback (RLHF) with Proximal Policy Optimization (PPO), eliminating the need for reward models and online sampling. Despite these benefits, DPO and its variants remain sensitive to hyper-parameters and prone to instability, particularly on mathematical datasets. We argue that these issues arise from the unidirectional likelihood-derivative negative feedback inherent in the log-likelihood loss function.
To address this, we propose a novel LLM alignment loss that establishes a stable Bidirectional Negative Feedback (BNF) during optimization.
Our proposed BNF loss eliminates the need for pairwise contrastive losses and does not require any extra tunable hyper-parameters or pairwise preference data, streamlining the alignment pipeline to be as simple as supervised fine-tuning.
We conduct extensive experiments across two challenging QA benchmarks and four reasoning benchmarks.
The experimental results show that BNF achieves comparable performance to the best methods on QA benchmarks, while its performance decrease on the four reasoning benchmarks is significantly lower compared to the best methods, thus striking a better balance between value alignment and reasoning ability.
In addition, we further validate the performance of BNF on non-pairwise datasets, and conduct in-depth analysis of log-likelihood and logit shifts across different preference optimization methods.
We will release all the source code, checkpoints, and datasets on GitHub.
****
Zhicheng Yang,Yiwei Wang,Yinya Huang,Zhijiang Guo,Wei Shi,Xiongwei Han,Liang Feng,Linqi Song,Xiaodan Liang,Jing Tang
OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling
Large language models (LLMs) have exhibited their problem-solving abilities in mathematical reasoning. Solving realistic optimization (OPT) problems in application scenarios requires advanced and applied mathematics ability. However, current OPT benchmarks that merely solve linear programming are far from complex realistic situations. In this work, we propose **OptiBench**, a benchmark for End-to-end optimization problem-solving with human-readable inputs and outputs. **OptiBench** contains rich optimization problems, including linear and nonlinear programming with or without tabular data, which can comprehensively evaluate LLMs' solving ability. In our benchmark, LLMs are required to call a code solver to provide precise numerical answers.

Furthermore, to alleviate the data scarcity for optimization problems, and to br
idge the gap between open-source LLMs on a small scale (e.g., Llama-3-8b) and cl
osed-source LLMs (e.g., GPT-4), we further propose a data synthesis method namel
y ***ReSocratic***. Unlike general data synthesis methods that proceed from ques
tions to answers, \ReSocratic first incrementally synthesizes formatted optimiza
tion demonstration with mathematical formulations step by step and then back-tra
nslates the generated demonstrations into questions. Based on this, we synthesiz
e the ***ReSocratic-29k*** dataset. We further conduct supervised fine-tuning wi
th ***ReSocratic-29k*** on multiple open-source models. Experimental results sho
w that ***ReSocratic-29k*** significantly improves the performance of open-sourc
e models.
****
Yici Yan,Yichi Zhang,Xiangming Meng,Zhizhen Zhao
FIG: Flow with Interpolant Guidance for Linear Inverse Problems
Diffusion and flow matching models have recently been used to solve various line
ar inverse problems in image restoration, such as super-resolution and inpaintin
g. Using a pre-trained diffusion or flow-matching model as a prior, most existin
g methods modify the reverse-time sampling process by incorporating the likeliho
od information from the measurement. However, they struggle in challenging scena
rios, such as high measurement noise or severe ill-posedness. In this paper, we
propose Flow with Interpolant Guidance (FIG), an algorithm where reverse-time sa
mpling is efficiently guided with measurement interpolants through theoretically
 justified schemes. Experimentally, we demonstrate that FIG efficiently produces
 highly competitive results on a variety of linear image reconstruction tasks on
 natural image datasets, especially for challenging tasks. Our code is available
 at: https://riccizz.github.io/FIG/.
****
Ziteng Sun,Uri Mendlovic,Yaniv Leviathan,Asaf Aharoni,Jae Hun Ro,Ahmad Beirami,A
nanda Theertha Suresh
Block Verification Accelerates Speculative Decoding
Speculative decoding is an effective method for lossless acceleration of large
language models during inference. It uses a fast model to draft a block of token
s which are then verified in parallel by the target model, and provides a guaran
tee that the output is distributed identically to a sample from the target model
. In prior works, draft verification is performed independently token-by-token.
Surprisingly, we show that this approach is not optimal. We propose *Block Verif
ication*, a simple draft verification algorithm that verifies the entire block j
ointly and provides additional wall-clock speedup. We prove that the proposed me
chanism is optimal in the expected number of tokens produced each iteration and
specifically is never worse than the standard token-level verification.
Empirically, block verification provides modest but consistent wall-clock speedu
ps over the standard token verification algorithm of 5\%-8\% in a range of tasks
 and datasets.
Given that block verification does not increase code complexity, maintains the s
trong lossless guarantee of the standard speculative decoding verification algor
ithm, cannot deteriorate performance, and, in fact, consistently improves it, it
 can be used as a good default in speculative decoding implementations.
****
Yecheng Jason Ma,Joey Hejna,Chuyuan Fu,Dhruv Shah,Jacky Liang,Zhuo Xu,Sean Kirma
ni,Peng Xu,Danny Driess,Ted Xiao,Osbert Bastani,Dinesh Jayaraman,Wenhao Yu,Tingn
an Zhang,Dorsa Sadigh,Fei Xia
Vision Language Models are In-Context Value Learners
Predicting temporal progress from visual trajectories is important for intellige
nt robots that can learn, adapt, and improve. However, learning such progress es
timator, or temporal value function, across different tasks and domains requires
 both a large amount of diverse data and methods which can scale and generalize.
 To address these challenges, we present Generative Value Learning (GVL), a univ
ersal value function estimator that leverages the world knowledge embedded in vi
sion-language models (VLMs) to predict task progress. Naively asking a VLM to pr
edict values for a video sequence performs poorly due to the strong temporal cor

relation between successive frames. Instead, GVL poses value estimation as a temporal ordering problem over shuffled video frames; this seemingly more challenging task encourages VLMs to more fully exploit their underlying semantic and temporal grounding capabilities to differentiate frames based on their perceived task progress, consequently producing significantly better value predictions. Without any robot or task specific training, GVL can in-context zero-shot and few-shot predict effective values for more than 300 distinct real-world tasks across diverse robot platforms, including challenging bimanual manipulation tasks. Furthermore, we demonstrate that GVL permits flexible multi-modal in-context learning via examples from heterogeneous tasks and embodiments, such as human videos. The generality of GVL enables various downstream applications pertinent to visuomotor policy learning, including dataset filtering, success detection, and value-weighted regression -- all without any model training or finetuning.
****

Ziyi Yang,Fanqi Wan,Longguang Zhong,Tianyuan Shi,Xiaojun Quan
Weighted-Reward Preference Optimization for Implicit Model Fusion
While fusing heterogeneous open-source LLMs with varying architectures and sizes can potentially integrate the strengths of different models, existing fusion methods face significant challenges, such as vocabulary alignment and merging distribution matrices. These procedures are not only complex but also prone to introducing noise and errors. In this paper, we propose an implicit fusion method, Weighted-Reward Preference Optimization (WRPO), which leverages preference optimization between the source LLMs and the target LLM to transfer their capabilities effectively. WRPO eliminates the need for vocabulary alignment and matrix fusion and can be efficiently scaled to accommodate various LLMs. To address distributional deviations between the source and target LLMs, WRPO introduces a progressive adaptation strategy that gradually shifts reliance on preferred examples from the target LLM to the source LLMs. Extensive experiments on the MT-Bench, AlpacaEval-2, and Arena-Hard benchmarks demonstrate that WRPO consistently outperforms existing knowledge fusion methods and various fine-tuning baselines. When applied to LLaMA3-8B-Instruct as the target model, WRPO achieves a length-controlled win rate of 55.9\% against GPT-4-Preview-1106 on AlpacaEval-2 and a win rate of 46.2\% against GPT-4-0314 on Arena-Hard. Our code is available at https://github.com/SLIT-AI/WRPO.
****

Beomsu Kim,Yu-Guan Hsieh,Michal Klein,marco cuturi,Jong Chul Ye,Bahjat Kawar,James Thornton
Simple ReFlow: Improved Techniques for Fast Flow Models
Diffusion and flow-matching models achieve remarkable generative performance but at the cost of many neural function evaluations (NFE), which slows inference and limits applicability to time-critical tasks. The ReFlow procedure can accelerate sampling by straightening generation trajectories. But it is an iterative procedure, typically requiring training on simulated data, and results in reduced sample quality. To mitigate sample deterioration, we examine the design space of ReFlow and highlight potential pitfalls in prior heuristic practices. We then propose seven improvements for training dynamics, learning and inference, which are verified with thorough ablation studies on CIFAR10 32×32, AFHQv2 64×64, and FFHQ 64×64. Combining all our techniques, we achieve state-of-the-art FID scores (without / with guidance, resp.) for fast generation via neural ODEs: 2.23 / 1.98 on CIFAR10, 2.30 / 1.91 on AFHQv2, 2.84 / 2.67 on FFHQ, and 3.49 / 1.74 on ImageNet-64, all with merely 9 NFEs.
****

Davide Paglieri,Bart■omiej Cupia■,Samuel Coward,Ulyana Piterbarg,Maciej Wolczyk,Akbir Khan,Eduardo Pignatelli,■ukasz Kuci■ski,Lerrel Pinto,Rob Fergus,Jakob Nicolaus Foerster,Jack Parker-Holder,Tim Rocktäschel
BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games
Large Language Models (LLMs) and Vision Language Models (VLMs) possess extensive knowledge and exhibit promising reasoning abilities, however, they still struggle to perform well in complex, dynamic environments. Real-world tasks require handling intricate interactions, advanced spatial reasoning, long-term planning, a

nd continuous exploration of new strategies—areas in which we lack effective met
hodologies for comprehensively evaluating these capabilities. To address this ga
p, we introduce BALROG, a novel benchmark designed to assess the agentic capabil
ities of LLMs and VLMs through a diverse set of challenging games. Our benchmark
 incorporates a range of existing reinforcement learning environments with varyi
ng levels of difficulty, including tasks that are solvable by non-expert humans
in seconds to extremely challenging ones that may take years to master (e.g., th
e NetHack Learning Environment).
We devise fine-grained metrics to measure performance and conduct an extensive e
valuation of several popular open-source and closed-source LLMs and VLMs. Our fi
ndings indicate that while current models achieve partial success in the easier
games, they struggle significantly with more challenging tasks. Notably, we obse
rve severe deficiencies in vision-based decision-making, as several models perfo
rm worse when visual representations of the environments are provided. We releas
e BALROG as an open and user-friendly benchmark to facilitate future research an
d development in the agentic community. Code and Leaderboard at balrogai.com
****

Xin Xu,Jiaxin ZHANG,Tianhao Chen,Zitong Chao,Jishan Hu,Can Yang
UGMathBench: A Diverse and Dynamic Benchmark for Undergraduate-Level Mathematica
l Reasoning with Large Language Models
Large Language Models (LLMs) have made significant strides in mathematical reaso
ning, underscoring the need for a comprehensive and fair evaluation of their cap
abilities. However, existing benchmarks often fall short, either lacking extensi
ve coverage of undergraduate-level mathematical problems or probably suffering f
rom test-set contamination. To address these issues, we introduce UGMathBench, a
 diverse and dynamic benchmark specifically designed for evaluating undergraduat
e-level mathematical reasoning with LLMs. UGMathBench comprises 5,062 problems a
cross 16 subjects and 111 topics, featuring 10 distinct answer types. Each probl
em includes three randomized versions, with additional versions planned for rele
ase as the leading open-source LLMs become saturated in UGMathBench. Furthermore
, we propose two key metrics: effective accuracy (EAcc), which measures the perc
entage of correctly solved problems across all three versions, and reasoning gap
 (Δ), which assesses reasoning robustness by calculating the difference b
etween the average accuracy across all versions and EAcc. Our extensive evaluati
on of 23 leading LLMs reveals that the highest EAcc achieved is 56.3% by OpenAI-
o1-mini, with large Δ values observed across different models. This highl
ights the need for future research aimed at developing "large reasoning models"
with high EAcc and $\Delta = 0$. We anticipate that the release of UGMathBench,
along with its detailed evaluation codes, will serve as a valuable resource to a
dvance the development of LLMs in solving mathematical problems.
****

Anpeng Wu,Haiyi Qiu,Zhengming Chen,Zijian Li,Ruoxuan Xiong,Fei Wu,Kun Zhang
Causal Graph Transformer for Treatment Effect Estimation Under Unknown Interfere
nce
Networked interference, also known as the peer effect in social science and spil
lover effect in economics, has drawn increasing interest across various domains.
 This phenomenon arises when a unit’s treatment and outcome are influenced by th
e actions of its peers, posing significant challenges to causal inference, parti
cularly in treatment assignment and effect estimation in real applications, due
to the violation of the SUTVA assumption. While extensive graph models have been
 developed to identify treatment effects, these models often rely on structural
assumptions about networked interference, assuming it to be identical to the soc
ial network, which can lead to misspecification issues in real applications. To
address these challenges, we propose an Interference-Agnostic Causal Graph Trans
former (CauGramer), which aggregates peers information via L-order Graph Trans
former and employs cross-attention to infer aggregation function for learning in
terference representations. By integrating confounder balancing and minimax mome
nt constraints, CauGramer fully incorporates peer information, enabling robust t
reatment effect estimation. Extensive experiments on two widely-used benchmarks
demonstrate the effectiveness and superiority of CauGramer. The code is availabl

e at https://github.com/anpwu/CauGramer.
****
Nicolas Espinosa-Dice,Sanjiban Choudhury,Wen Sun,Gokul Swamy
Efficient Imitation under Misspecification
Interactive imitation learning (IL) is a powerful paradigm for learning to make sequences of decisions from an expert demonstrating how to perform a task. Prior work in efficient imitation learning has focused on the realizable setting, where the expert's policy lies within the learner's policy class (i.e. the learner can perfectly imitate the expert in all states). However, in practice, perfect imitation of the expert is often impossible due to differences in state information and action space expressiveness (e.g. morphological differences between robots and humans.) In this paper, we consider the more general misspecified setting, where no assumptions are made about the expert policy's realizability. We introduce a novel structural condition, reward-agnostic policy completeness, and prove that it is sufficient for interactive IL algorithms to efficiently avoid the quadratically compounding errors that stymie offline approaches like behavioral cloning. We address an additional practical constraint—the case of limited expert data—and propose a principled method for using additional offline data to further improve the sample-efficiency of interactive IL algorithms. Finally, we empirically investigate the optimal reset distribution in efficient IRL under misspecification with a suite of continuous control tasks.
****
Maxence Noble,Louis Grenioux,Marylou Gabrié,Alain Oliviero Durmus
Learned Reference-based Diffusion Sampler for multi-modal distributions
Over the past few years, several approaches utilizing score-based diffusion have been proposed to sample from probability distributions, that is without having access to exact samples and relying solely on evaluations of unnormalized densities. The resulting samplers approximate the time-reversal of a noising diffusion process, bridging the target distribution to an easy-to-sample base distribution. In practice, the performance of these methods heavily depends on key hyperparameters that require ground truth samples to be accurately tuned. Our work aims to highlight and address this fundamental issue, focusing in particular on multi-modal distributions, which pose significant challenges for existing sampling methods. Building on existing approaches, we introduce *Learned Reference-based Diffusion Sampler* (LRDS), a methodology specifically designed to leverage prior knowledge on the location of the target modes in order to bypass the obstacle of hyperparameter tuning. LRDS proceeds in two steps by (i) learning a *reference* diffusion model on samples located in high-density space regions and tailored for multimodality, and (ii) using this reference model to foster the training of a diffusion-based sampler. We experimentally demonstrate that LRDS best exploits prior knowledge on the target distribution compared to competing algorithms on a variety of challenging distributions.
****
Xianglu Wang,Hu Ding
Exploring The Forgetting in Adversarial Training: A Novel Method for Enhancing Robustness
In recent years, there has been an explosion of research into developing robust deep neural networks against adversarial examples. As one of the most successful methods, Adversarial Training (AT) has been widely studied before, but there is still a gap to achieve promising
clean and robust accuracy for many practical tasks. In this paper, we consider the AT problem from a new perspective which connects it to catastrophic forgetting in continual learning (CL). Catastrophic forgetting is a phenomenon in which neural networks forget old knowledge upon learning a new task. Although AT and CL are two different problems, we show that they actually share several key properties in their training processes. Specifically, we conduct an empirical study and find that this forgetting phenomenon indeed occurs in adversarial robust training across multiple datasets (SVHN, CIFAR-10, CIFAR-100, and TinyImageNet) and perturbation models (ℓ_{∞} and ℓ_{2}). Based on this observation, we propose a novel method called Adaptive Multi-teachers Self-distillation (AMS

), which leverages a carefully designed adaptive regularizer to mitigate the for
getting by aligning model outputs between new and old ``stages''. Moreover, our
approach can be used as a unified method to enhance multiple different AT algor
ithms. Our experiments demonstrate that our method can significantly enhance rob
ust accuracy and meanwhile preserve high clean accuracy, under several popular a
dversarial attacks (e.g., PGD, CW, and Auto Attacks). As another benefit of our
method, we discover that it can largely alleviate the robust overfitting issue o
f AT in our experiments.
****
Xiaoshuai Song,Muxi Diao,Guanting Dong,Zhengyang Wang,Yujia Fu,Runqi Qiao,Zhexu
Wang,Dayuan Fu,Huangxuan Wu,Bin Liang,Weihao Zeng,Yejie Wang,Zhuoma GongQue,Jian
ing Yu,Qiuna Tan,Weiran Xu
CS-Bench: A Comprehensive Benchmark for Large Language Models towards Computer S
cience Mastery
Large language models (LLMs) have demonstrated significant potential in advancin
g various fields of research and society. However, the current community of LLMs
 overly focuses on benchmarks for analyzing specific foundational skills (e.g. m
athematics and code generation), neglecting an all-round evaluation of the compu
ter science field. To bridge this gap, we introduce CS-Bench, the first multilin
gual (English, Chinese, French, German) benchmark dedicated to evaluating the pe
rformance of LLMs in computer science. CS-Bench comprises approximately 10K meti
culously curated test samples, covering 26 subfields across 4 key areas of compu
ter science, encompassing various task forms and divisions of knowledge and reas
oning. Utilizing CS-Bench, we conduct a comprehensive evaluation of over 30 main
stream LLMs, revealing the relationship between CS performance and model scales.
 We also quantitatively analyze the reasons for failures in existing LLMs and hi
ghlight directions for improvements, including knowledge supplementation and CS-
specific reasoning. Further cross-capability experiments show a high correlation
 between LLMs' capabilities in computer science and their abilities in mathemati
cs and coding. Moreover, expert LLMs specialized in mathematics and coding also
demonstrate strong performances in several CS subfields. Looking ahead, we envis
ion CS-Bench serving as a cornerstone for LLM applications in the CS field and p
aving new avenues in assessing LLMs' diverse reasoning capabilities. Our project
 homepage is available at https://csbench.github.io/.
****
Samuel Duffield,Kaelan Donatella,Johnathan Chiu,Phoebe Klett,Daniel Simpson
Scalable Bayesian Learning with posteriors
Although theoretically compelling, Bayesian learning with modern machine learnin
g models is computationally challenging since it requires approximating a high d
imensional posterior distribution. In this work, we (i) introduce **_posteriors_
**, an easily extensible PyTorch library hosting general-purpose implementations
 making Bayesian learning accessible and scalable to large data and parameter re
gimes; (ii) present a tempered framing of stochastic gradient Markov chain Monte
 Carlo, as implemented in posteriors, that transitions seamlessly into optimizat
ion and unveils a minor modification to deep ensembles to ensure they are asympt
otically unbiased for the Bayesian posterior, and (iii) demonstrate and compare
the utility of Bayesian approximations through experiments including an investig
ation into the cold posterior effect and applications with large language models
.

**posteriors** repository: https://github.com/normal-computing/posteriors
****
Wenye Li,Jiacai Liu,Ke Wei
ϕ-Update: A Class of Policy Update Methods with Policy Convergence Guarante
e
Inspired by the similar update pattern of softmax natural policy gradient and Ha
damard policy gradient, we propose to study a general policy update rule called
ϕ-update, where ϕ refers to a scaling function on advantage functions.
 Under very mild conditions on ϕ, the global asymptotic state value converg
ence of ϕ-update is firstly established. Then we show that the policy produ

ced by ϕ-update indeed converges, even when there are multiple optimal poli
cies. This is in stark contrast to existing results where explicit regulariza
tions are required to guarantee the convergence of the policy. Since softmax n
atural policy gradient is an instance of ϕ-update, it provides an affirmati
ve answer to the question whether the policy produced by softmax natural policy
gradient converges. The exact asymptotic convergence rate of state values is fur
ther established based on the policy convergence. Lastly, we establish the globa
l linear convergence of ϕ-update.
****
Xinran Liu,Yikun Bai,Rocio Diaz Martin,Kaiwen Shi,Ashkan Shahbazi,Bennett Allan
Landman,Catie Chang,Soheil Kolouri
Linear Spherical Sliced Optimal Transport: A Fast Metric for Comparing Spherical
 Data
Efficient comparison of spherical probability distributions becomes important in
 fields such as computer vision, geosciences, and medicine. Sliced optimal trans
port distances, such as spherical and stereographic spherical sliced Wasserstein
 distances, have recently been developed to address this need. These methods red
uce the computational burden of optimal transport by slicing hyperspheres into o
ne-dimensional projections, i.e., lines or circles. Concurrently, linear optimal
 transport has been proposed to embed distributions into L^2 spaces, where the
 L^2 distance approximates the optimal transport distance, thereby simplifying
 comparisons across multiple distributions. In this work, we introduce the Linea
r Spherical Sliced Optimal Transport (LSSOT) framework, which utilizes slicing t
o embed spherical distributions into L^2 spaces while preserving their intrins
ic geometry, offering a computationally efficient metric for spherical probabili
ty measures. We establish the metricity of LSSOT and demonstrate its superior co
mputational efficiency in applications such as cortical surface registration, 3D
 point cloud interpolation via gradient flow, and shape embedding. Our results d
emonstrate the significant computational benefits and high accuracy of LSSOT in
these applications.
****
Somesh Kumar Singh,Harini S I,Yaman Kumar Singla,Changyou Chen,Rajiv Ratn Shah,V
eeky Baths,Balaji Krishnamurthy
Teaching Human Behavior Improves Content Understanding Abilities Of VLMs
Communication is defined as "*Who* says *what* to *whom* with *what* effect." A
message from a communicator generates downstream receiver effects, also known as
 behavior. Receiver behavior, being a downstream effect of the message, carries
rich signals about it. Even after carrying signals about the message, the behavi
or signal is often ignored while training vision-language models. We show that t
raining VLMs on receiver behavior can actually help improve their content-unders
tanding abilities. We demonstrate that training VLMs to predict receiver behavio
rs, such as likes, comments, and replay graphs, which are available at scale, en
hances the VLM's performance across a broad range of downstream content understa
nding tasks. We show this performance increase over 6 types of behavior, 46 diff
erent tasks covering image, video, text, and audio over 26 benchmark datasets ac
ross both zero-shot and fine-tuning settings, outperforming many supervised base
lines on diverse tasks ranging from emotion recognition to captioning by up to 1
50%. We note that since receiver behavior, such as likes, comments, and replay g
raphs, is collected by default on the internet and does not need any human annot
ations to be useful, the performance improvement we get after training on this d
ata is essentially free lunch. We also release **BLIFT**, our **Behaviour-LLaVA
 IFT** dataset comprising 730k images and videos with their receiver behavior co
llected from multiple platforms on which we train our models to achieve this. Th
e dataset and code are available at [behavior-in-the-wild.github.io/behavior-lla
va](https://behavior-in-the-wild.github.io/behavior-llava).
****
Zeyu Yang,Zijie Pan,Chun Gu,Li Zhang
Diffusion^2: Dynamic 3D Content Generation via Score Composition of Video and
Multi-view Diffusion Models
Recent advancements in 3D generation are predominantly propelled by improvements

in 3D-aware image diffusion models. These models are pretrained on Internet-scale image data and fine-tuned on massive 3D data, offering the capability of producing highly consistent multi-view images. However, due to the scarcity of synchronized multi-view video data, it remains challenging to adapt this paradigm to 4D generation directly. Despite that, the available video and 3D data are adequate for training video and multi-view diffusion models separately that can provide satisfactory dynamic and geometric priors respectively. To take advantage of both, this paper presents Diffusion^2, a novel framework for dynamic 3D content creation that reconciles the knowledge about geometric consistency and temporal smoothness from these models to directly sample dense multi-view multi-frame images which can be employed to optimize continuous 4D representation. Specifically, we design a simple yet effective denoising strategy via score composition of pretrained video and multi-view diffusion models based on the probability structure of the target image array. To alleviate the potential conflicts between two heterogeneous scores, we further introduce variance-reducing sampling via interpolated steps, facilitating smooth and stable generation. Owing to the high parallelism of the proposed image generation process and the efficiency of the modern 4D reconstruction pipeline, our framework can generate 4D content within few minutes. Notably, our method circumvents the reliance on expensive and hard-to-scale 4D data, thereby having the potential to benefit from the scaling of the foundation video and multi-view diffusion models. Extensive experiments demonstrate the efficacy of our proposed framework in generating highly seamless and consistent 4D assets under various types of conditions.
****
Yingyan Li,Lue Fan,Jiawei He,Yuqi Wang,Yuntao Chen,Zhaoxiang Zhang,Tieniu Tan
Enhancing End-to-End Autonomous Driving with Latent World Model
In autonomous driving, end-to-end planners directly utilize raw sensor data, enabling them to extract richer scene features and reduce information loss compared to traditional planners. This raises a crucial research question: how can we develop better scene feature representations to fully leverage sensor data in end-to-end driving? Self-supervised learning methods show great success in learning rich feature representations in NLP and computer vision. Inspired by this, we propose a novel self-supervised learning approach using the LAtent World model (LAW) for end-to-end driving. LAW predicts future latent scene features based on current features and ego trajectories. This self-supervised task can be seamlessly integrated into perception-free and perception-based frameworks, improving scene feature learning while optimizing trajectory prediction. LAW achieves state-of-the-art performance across multiple benchmarks, including real-world open-loop benchmark nuScenes, NAVSIM, and simulator-based closed-loop benchmark CARLA. The code will be released.
****
Zongrui Zou,Jingcheng Liu,Jalaj Upadhyay
Optimality of Matrix Mechanism on ℓ_p^p-metric
In this paper, we introduce the ℓ_p^p-error metric (for $p \geq 2$) when answering linear queries under the constraint of differential privacy. We characterize such an error under (ϵ,δ)-differential privacy in the natural add/remove model. Before this paper, tight characterization in the hardness of privately answering linear queries was known under ℓ_2^2-error metric (Edmonds et al. 2020) and ℓ_p^2-error metric for unbiased mechanisms in the substitution model (Nikolov et al. 2024). As a direct consequence of our results, we give tight bounds on answering prefix sum and parity queries under differential privacy for all constant p in terms of the ℓ_p^p error, generalizing the bounds in Hhenzinger et al. for $p=2$.
****
Hong Yang,Qi Yu,Travis Desell
Can We Ignore Labels in Out of Distribution Detection?
Out-of-distribution (OOD) detection methods have recently become more prominent, serving as a core element in safety-critical autonomous systems. One major purpose of OOD detection is to reject invalid inputs that could lead to unpredictable errors and compromise safety. Due to the cost of labeled data, recent works ha

ve investigated the feasibility of self-supervised learning (SSL) OOD detection, unlabled OOD detection, and zero shot OOD detection. In this work, we identify a set of conditions for a theoretical guarantee of failure in unlabeled OOD detection algorithms from an information-theoretic perspective. These conditions are present in all OOD tasks dealing with real world data: I) we provide theoretical proof of unlabeled OOD detection failure when there exists zero mutual information between the learning objective and the in-distribution labels, a.k.a. ‘label blindness’, II) we define a new OOD task – Adjacent OOD detection – that tests for label blindness and accounts for a previously ignored safety gap in all OOD detection benchmarks, and III) we perform experiments demonstrating that existing unlabeled OOD methods fail under conditions suggested by our label blindness theory and analyze the implications for future research in unlabeled OOD methods.

****

Huy Nguyen,Pedram Akbarian,Huyen Trang Pham,Thien Trang Nguyen Vu,Shujian Zhang, Nhat Ho

Statistical Advantages of Perturbing Cosine Router in Mixture of Experts

The cosine router in Mixture of Experts (MoE) has recently emerged as an attractive alternative to the conventional linear router. Indeed, the cosine router demonstrates favorable performance in image and language tasks and exhibits better ability to mitigate the representation collapse issue, which often leads to parameter redundancy and limited representation potentials. Despite its empirical success, a comprehensive analysis of the cosine router in MoE has been lacking. Considering the least square estimation of the cosine routing MoE, we demonstrate that due to the intrinsic interaction of the model parameters in the cosine router via some partial differential equations, regardless of the structures of the experts, the estimation rates of experts and model parameters can be as slow as $\mathcal{O}(1/\log^{\tau}(n))$ where $\tau > 0$ is some constant and n is the sample size. Surprisingly, these pessimistic non-polynomial convergence rates can be circumvented by the widely used technique in practice to stabilize the cosine router --- simply adding noises to the ℓ^2-norms in the cosine router, which we refer to as *perturbed cosine router*. Under the strongly identifiable settings of the expert functions, we prove that the estimation rates for both the experts and model parameters under the perturbed cosine routing MoE are significantly improved to polynomial rates. Finally, we conduct extensive simulation studies in both synthetic and real data settings to empirically validate our theoretical results.

****

Jesse C. Cresswell,Bhargava Kumar,Yi Sui,Mouloud Belbahri

Conformal Prediction Sets Can Cause Disparate Impact

Conformal prediction is a statistically rigorous method for quantifying uncertainty in models by having them output sets of predictions, with larger sets indicating more uncertainty. However, prediction sets are not inherently actionable; many applications require a single output to act on, not several. To overcome this limitation, prediction sets can be provided to a human who then makes an informed decision. In any such system it is crucial to ensure the fairness of outcomes across protected groups, and researchers have proposed that Equalized Coverage be used as the standard for fairness. By conducting experiments with human participants, we demonstrate that providing prediction sets can lead to disparate impact in decisions. Disquietingly, we find that providing sets that satisfy Equalized Coverage actually increases disparate impact compared to marginal coverage. Instead of equalizing coverage, we propose to equalize set sizes across groups which empirically leads to lower disparate impact.

****

Zichen Miao,Zhengyuan Yang,Kevin Lin,Ze Wang,Zicheng Liu,Lijuan Wang,Qiang Qiu

Tuning Timestep-Distilled Diffusion Model Using Pairwise Sample Optimization

Recent advancements in timestep-distilled diffusion models have enabled high-quality image generation that rivals non-distilled multi-step models, but with significantly fewer inference steps. While such models are attractive for applications due to the low inference cost and latency, fine-tuning them with a naive diff

usion objective would result in degraded and blurry outputs. An intuitive altern
ative is to repeat the diffusion distillation process with a fine-tuned teacher
model, which produces good results but is cumbersome and computationally intensi
ve: the distillation training usually requires magnitude higher of training comp
ute compared to fine-tuning for specific image styles. In this paper, we present
 an algorithm named pairwise sample optimization (PSO), which enables the direct
 fine-tuning of an arbitrary timestep-distilled diffusion model. PSO introduces
additional reference images sampled from the current time-step distilled model,
and increases the relative likelihood margin between the training images and ref
erence images. This enables the model to retain its few-step generation ability,
 while allowing for fine-tuning of its output distribution. We also demonstrate
that PSO is a generalized formulation which be flexible extended to both offline
-sampled and online-sampled pairwise data, covering various popular objectives f
or diffusion model preference optimization. We evaluate PSO in both preference o
ptimization and other fine-tuning tasks, including style transfer and concept cu
stomization. We show that PSO can directly adapt distilled models to human-prefe
rred generation with both offline and online-generated pairwise preference image
 data. PSO also demonstrates effectiveness in style transfer and concept customi
zation by directly tuning timestep-distilled diffusion models.
****

Matthew Riemer,Gopeshh Subbaraj,Glen Berseth,Irina Rish
Enabling Realtime Reinforcement Learning at Scale with Staggered Asynchronous In
ference
Realtime environments change even as agents perform action inference and learnin
g, thus requiring high interaction frequencies to effectively minimize regret. H
owever, recent advances in machine learning involve larger neural networks with
longer inference times, raising questions about their applicability in realtime
systems where reaction time is crucial. We present an analysis of lower bounds o
n regret in realtime reinforcement learning (RL) environments to show that minim
izing long-term regret is generally impossible within the typical sequential int
eraction and learning paradigm, but often becomes possible when sufficient async
hronous compute is available. We propose novel algorithms for staggering asynchr
onous inference processes to ensure that actions are taken at consistent time in
tervals, and demonstrate that use of models with high action inference times is
only constrained by the environment's effective stochasticity over the inference
 horizon, and not by action frequency. Our analysis shows that the number of inf
erence processes needed scales linearly with increasing inference times while en
abling use of models that are multiple orders of magnitude larger than existing
approaches when learning from a realtime simulation of Game Boy games such as Po
kemon and Tetris.
****

Xiangyu Qi,Boyi Wei,Nicholas Carlini,Yangsibo Huang,Tinghao Xie,Luxi He,Matthew
Jagielski,Milad Nasr,Prateek Mittal,Peter Henderson
On Evaluating the Durability of Safeguards for Open-Weight LLMs
Many stakeholders---from model developers to policymakers---seek to minimize the
 risks of large language models (LLMs). Key to this goal is whether technical sa
feguards can impede the misuse of LLMs, even when models are customizable via fi
ne-tuning or when model weights are openly available. Several recent studies hav
e proposed methods to produce durable LLM safeguards for open-weight LLMs that c
an withstand adversarial modifications of the model's weights via fine-tuning. T
his holds the promise of raising adversaries' costs even under strong threat mod
els where adversaries can directly fine-tune parameters. However, we caution ag
ainst over-reliance on such methods in their current state. Through several case
 studies, we demonstrate that even the evaluation of these defenses is exceeding
ly difficult and can easily mislead audiences into thinking that safeguards are
more durable than they really are. We draw lessons from the failure modes that w
e identify and suggest that future research carefully cabin claims to more const
rained, well-defined, and rigorously examined threat models, which can provide u
seful and candid assessments to stakeholders.
****

Sarkar Snigdha Sarathi Das,Ryo Kamoi,Bo Pang,Yusen Zhang,Caiming Xiong,Rui Zhang
GReaTer: Gradients Over Reasoning Makes Smaller Language Models Strong Prompt Optimizers
The effectiveness of large language models (LLMs) is closely tied to the design of prompts, making prompt optimization essential for enhancing their performance across a wide range of tasks. Although recent advancements have focused on automating prompt engineering, many existing approaches rely exclusively on textual feedback, refining prompts based solely on inference errors identified by large, computationally expensive LLMs. Unfortunately, smaller models struggle to generate high-quality feedback, resulting in complete dependence on large LLM judgment. Moreover, these methods fail to leverage more direct and finer-grained information, such as gradients, due to operating purely in text space. To this end, we introduce, we introduce *GReaTer*, a novel prompt optimization technique that directly incorporates *gradient information over task-specific reasoning*. By utilizing task loss gradients, *GReaTer* enables self-optimization of prompts for smaller, lightweight language models (LM) without the need for costly closed-source LLMs, while maintaining reasonable prompt structures. This allows high-performance prompt optimization without dependence on massive LLMs, closing the gap between smaller models and the sophisticated reasoning often needed for prompt refinement. Extensive evaluations across diverse tasks demonstrate that \ours consistently outperforms previous methods, even those reliant on powerful LLMs. Additionally, *GReaTer*-optimized prompts frequently exhibit better transferability and, in some cases, boost task performance to levels comparable to or surpassing those achieved by larger language models, highlighting the effectiveness of *"gradient over reasoning"*-based prompt optimization. Code of *GReaTer* is available at: https://github.com/psunlpgroup/GreaTer
****
Zijing Ou,Mingtian Zhang,Andi Zhang,Tim Z. Xiao,Yingzhen Li,David Barber
Improving Probabilistic Diffusion Models With Optimal Diagonal Covariance Matching
The probabilistic diffusion model has become highly effective across various domains. Typically, sampling from a diffusion model involves using a denoising distribution characterized by a Gaussian with a learned mean and either fixed or learned covariances. In this paper, we leverage the recently proposed covariance moment matching technique and introduce a novel method for learning the diagonal covariances. Unlike traditional data-driven covariance approximation approaches, our method involves directly regressing the optimal analytic covariance using a new, unbiased objective named Optimal Covariance Matching (OCM). This approach can significantly reduce the approximation error in covariance prediction. We demonstrate how our method can substantially enhance the sampling efficiency, recall rate and likelihood of both diffusion models and latent diffusion models.
****
Bocheng Zeng,Qi Wang,Mengtao Yan,Yang Liu,Ruizhi Chengze,Yi Zhang,Hongsheng Liu,Zidong Wang,Hao Sun
PhyMPGN: Physics-encoded Message Passing Graph Network for spatiotemporal PDE systems
Solving partial differential equations (PDEs) serves as a cornerstone for modeling complex dynamical systems. Recent progresses have demonstrated grand benefits of data-driven neural-based models for predicting spatiotemporal dynamics (e.g., tremendous speedup gain compared with classical numerical methods). However, most existing neural models rely on rich training data, have limited extrapolation and generalization abilities, and suffer to produce precise or reliable physical prediction under intricate conditions (e.g., irregular mesh or geometry, complex boundary conditions, diverse PDE parameters, etc.). To this end, we propose a new graph learning approach, namely, Physics-encoded Message Passing Graph Network (PhyMPGN), to model spatiotemporal PDE systems on irregular meshes given small training datasets. Specifically, we incorporate a GNN into a numerical integrator to approximate the temporal marching of spatiotemporal dynamics for a given PDE system. Considering that many physical phenomena are governed by diffusion processes, we further design a learnable Laplace block, which encodes the discr

ete Laplace-Beltrami operator, to aid and guide the GNN learning in a physically feasible solution space. A boundary condition padding strategy is also designed to improve the model convergence and accuracy. Extensive experiments demonstrate that PhyMPGN is capable of accurately predicting various types of spatiotemporal dynamics on coarse unstructured meshes, consistently achieves the state-of-the-art results, and outperforms other baselines with considerable gains.
****
Haoran He,Can Chang,Huazhe Xu,Ling Pan
Looking Backward: Retrospective Backward Synthesis for Goal-Conditioned GFlowNets
Generative Flow Networks (GFlowNets), a new family of probabilistic samplers, have demonstrated remarkable capabilities to generate diverse sets of high-reward candidates, in contrast to standard return maximization approaches (e.g., reinforcement learning) which often converge to a single optimal solution. Recent works have focused on developing goal-conditioned GFlowNets, which aim to train a single GFlowNet capable of achieving different outcomes as the task specifies. However, training such models is challenging due to extremely sparse rewards, particularly in high-dimensional problems. Moreover, previous methods suffer from the limited coverage of explored trajectories during training, which presents more pronounced challenges when only offline data is available. In this work, we propose a novel method called \textbf{R}etrospective \textbf{B}ackward \textbf{S}ynthesis (\textbf{RBS}) to address these critical problems. Specifically, RBS synthesizes new backward trajectories in goal-conditioned GFlowNets to enrich training trajectories with enhanced quality and diversity, thereby introducing copious learnable signals for effectively tackling the sparse reward problem. Extensive empirical results show that our method improves sample efficiency by a large margin and outperforms strong baselines on various standard evaluation benchmarks. Our codes are available at \url{https://github.com/tinnerhrhe/Goal-Conditioned-GFN}.
****
Jérémy Perez,Grgur Kova■,Corentin Léger,Cédric Colas,Gaia Molinaro,Maxime Derex,Pierre-Yves Oudeyer,Clément Moulin-Frier
When LLMs Play the Telephone Game: Cultural Attractors as Conceptual Tools to Evaluate LLMs in Multi-turn Settings
As large language models (LLMs) start interacting with each other and generating an increasing amount of text online, it becomes crucial to better understand how information is transformed as it passes from one LLM to the next. While significant research has examined individual LLM behaviors, existing studies have largely overlooked the collective behaviors and information distortions arising from iterated LLM interactions. Small biases, negligible at the single output level, risk being amplified in iterated interactions, potentially leading the content to evolve towards attractor states. In a series of _telephone game experiments_, we apply a transmission chain design borrowed from the human cultural evolution literature: LLM agents iteratively receive, produce, and transmit texts from the previous to the next agent in the chain. By tracking the evolution of text _toxicity_, _positivity_, _difficulty_, and _length_ across transmission chains, we uncover the existence of biases and attractors, and study their dependence on the initial text, the instructions, language model, and model size. For instance, we find that more open-ended instructions lead to stronger attraction effects compared to more constrained tasks. We also find that different text properties display different sensitivity to attraction effects, with _toxicity_ leading to stronger attractors than _length_. These findings highlight the importance of accounting for multi-step transmission dynamics and represent a first step towards a more comprehensive understanding of LLM cultural dynamics.
****
Zhifan Ye,Kejing Xia,Yonggan Fu,Xin Dong,Jihoon Hong,Xiangchi Yuan,Shizhe Diao,Jan Kautz,Pavlo Molchanov,Yingyan Celine Lin
LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement
State space models (SSMs) have emerged as an efficient alternative to Transforme

r models for language modeling, offering linear computational complexity and con
stant memory usage as context length increases. However, despite their efficienc
y in handling long contexts, recent studies have shown that SSMs, such as Mamba
models, generally underperform compared to Transformers in long-context understa
nding tasks. To address this significant shortfall and achieve both efficient an
d accurate long-context understanding, we propose LongMamba, a training-free tec
hnique that significantly enhances the long-context capabilities of Mamba models
. LongMamba builds on our discovery that the hidden channels in Mamba can be cat
egorized into local and global channels based on their receptive field lengths,
with global channels primarily responsible for long-context capability. These gl
obal channels can become the key bottleneck as the input context lengthens. Spec
ifically, when input lengths largely exceed the training sequence length, global
 channels exhibit limitations in adaptively extend their receptive fields, leadi
ng to Mamba’s poor long-context performance. The key idea of LongMamba is to mit
igate the hidden state memory decay in these global channels by preventing the a
ccumulation of unimportant tokens in their memory. This is achieved by first ide
ntifying critical tokens in the global channels and then applying token filterin
g to accumulate only those critical tokens. Through extensive benchmarking acros
s synthetic and real-world long-context scenarios, LongMamba sets a new standard
 for Mamba’s long-context performance, significantly extending its operational r
ange without requiring additional training. Our code is available at https://git
hub.com/GATECH-EIC/LongMamba.
****

Site Bai,Brian Bullins

Tight Lower Bounds under Asymmetric High-Order Hölder Smoothness and Uniform Con
vexity

In this paper, we provide tight lower bounds for the oracle complexity of minimi
zing high-order Hölder smooth and uniformly convex functions. Specifically, for
a function whose p^{th}-order derivatives are Hölder continuous with degree $\
nu$ and parameter H, and that is uniformly convex with degree q and paramete
r σ, we focus on two asymmetric cases: (1) $q > p + \nu$, and (2) $q < p+
\nu$. Given up to p^{th}-order oracle access, we establish worst-case oracle c
omplexities of $\Omega\left(\left(\frac{H}{\sigma}\right)^\frac{2}{3(p+\nu)-2}
\left(\frac{\sigma}{\epsilon}\right)^\frac{2(q-p-\nu)}{q(3(p+\nu)-2)}\right)$ i
n the first case with an ℓ_∞-ball-truncated-Gaussian smoothed hard fun
ction and $\Omega\left(\left(\frac{H}{\sigma}\right)^\frac{2}{3(p+\nu)-2}+ \log\
log\left(\left(\frac{\sigma^{p+\nu}}{H^q}\right)^\frac{1}{p+\nu-q}\frac{1}{\epsi
lon}\right)\right)$ in the second case, for reaching an ϵ-approximate s
olution in terms of the optimality gap. Our analysis generalizes previous lower
bounds for functions under first- and second-order smoothness as well as those f
or uniformly convex functions, and furthermore our results match the correspondi
ng upper bounds in this general setting.
****

Shengyuan Hu,Yiwei Fu,Steven Wu,Virginia Smith

Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relea
rning

Machine unlearning is a promising approach to mitigate undesirable memorization
of training data in ML models. However, in this work we show that existing appro
aches for unlearning in LLMs are surprisingly susceptible to a simple set of ben
ign relearning attacks. With access to only a small and potentially loosely rela
ted set of data, we find that we can “jog” the memory of unlearned models to rev
erse the effects of unlearning. For example, we show that relearning on public m
edical articles can lead an unlearned LLM to output harmful knowledge about biow
eapons, and relearning general wiki information about the book series Harry Pott
er can force the model to output verbatim memorized text. We formalize this unle
arning-relearning pipeline, explore the attack across three popular unlearning b
enchmarks, and discuss future directions and guidelines that result from our stu
dy. Our work indicates that current approximate unlearning methods simply suppre
ss the model outputs and fail to robustly forget target knowledge in the LLMs.
****

Lizhe Fang,Yifei Wang,Zhaoyang Liu,Chenheng Zhang,Stefanie Jegelka,Jinyang Gao,Bolin Ding,Yisen Wang
What is Wrong with Perplexity for Long-context Language Modeling?
Handling long-context inputs is crucial for large language models (LLMs) in tasks such as extended conversations, document summarization, and many-shot in-context learning. While recent approaches have extended the context windows of LLMs and employed perplexity (PPL) as a standard evaluation metric, PPL has proven unreliable for assessing long-context capabilities. The underlying cause of this limitation has remained unclear. In this work, we provide a comprehensive explanation for this issue. We find that PPL overlooks key tokens, which are essential for long-context understanding, by averaging across all tokens and thereby obscuring the true performance of models in long-context scenarios. To address this, we propose \textbf{LongPPL}, a novel metric that focuses on key tokens by employing a long-short context contrastive method to identify them. Our experiments demonstrate that LongPPL strongly correlates with performance on various long-context benchmarks (e.g., Pearson correlation of -0.96), significantly outperforming traditional PPL in predictive accuracy. Additionally, we introduce \textbf{LongCE} (Long-context Cross-Entropy) loss, a re-weighting strategy for fine-tuning that prioritizes key tokens, leading to consistent improvements across diverse benchmarks. In summary, these contributions offer deeper insights into the limitations of PPL and present effective solutions for accurately evaluating and enhancing the long-context capabilities of LLMs. Code is available at https://github.com/PKU-ML/LongPPL.
****
Xiu-Chuan Li,Jun Wang,Tongliang Liu
Recovery of Causal Graph Involving Latent Variables via Homologous Surrogates
Causal discovery with latent variables is an important and challenging problem. To identify latent variables and infer their causal relations, most existing works rely on the assumption that latent variables have pure children. Considering that this assumption is potentially restrictive in practice and not strictly necessary in theory, in this paper, by introducing the concept of homologous surrogate, we eliminate the need for pure children in the context of causal discovery with latent variables. The homologous surrogate fundamentally differs from the pure child in the sense that the latter is characterized by having strictly restricted parents while the former allows for much more flexible parents. We formulate two assumptions involving homologous surrogates and develop theoretical results under each assumption. Under the weaker assumption, our theoretical results imply that we can determine each variable's ancestors, that is, partially recover the causal graph. The stronger assumption further enables us to determine each variable's parents exactly, that is, fully recover the causal graph. Building on these theoretical results, we derive an algorithm that fully leverages the properties of homologous surrogates for causal graph recovery. Also, we validate its efficacy through experiments. Our work broadens the applicability of causal discovery. Our code is available at: https://github.com/XiuchuanLi/ICLR2025-CDHS
****
Milad Sefidgaran,Abdellatif Zaidi,Piotr Krasnowski
Generalization Guarantees for Representation Learning via Data-Dependent Gaussian Mixture Priors
We establish in-expectation and tail bounds on the generalization error of representation learning type algorithms. The bounds are in terms of the relative entropy between the distribution of the representations extracted from the training and "test'' datasets and a data-dependent symmetric prior, i.e., the Minimum Description Length (MDL) of the latent variables for the training and test datasets. Our bounds are shown to reflect the "structure" and "simplicity'' of the encoder and significantly improve upon the few existing ones for the studied model. We then use our in-expectation bound to devise a suitable data-dependent regularizer; and we investigate thoroughly the important question of the selection of the prior. We propose a systematic approach to simultaneously learning a data-dependent Gaussian mixture prior and using it as a regularizer. Interestingly, we show that a weighted attention mechanism emerges naturally in this procedure. Our

experiments show that our approach outperforms the now popular Variational Information Bottleneck (VIB) method as well as the recent Category-Dependent VIB (CDVIB).
****
Kaiyan Zhang,Jiayuan Zhang,Haoxin Li,Xuekai Zhu,Ermo Hua,Xingtai Lv,Ning Ding,Biqing Qi,Bowen Zhou
OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees
Scaling inference-time computation is increasingly seen as the next frontier in scaling laws for large language models. Previous work in mathematics and coding has demonstrated the remarkable potential for inference-time scaling. During such scaling, fine-grained supervision through process-based reward models (PRMs) is essential for enhancement. However, exploration of inference-time scaling and PRMs in open-domain problems remains limited, where lacking exact answers and obtaining process supervision prove challenging. In this paper, we explore the construction of PRMs for open-domain tasks, specifically for instruction-following tasks. Utilizing existing outcome-based reward models (ORMs), we develop sentence-level preference trees based on the prefix similarity of parallel sampled candidates from datasets like UltraFeedback. This setup allows us to derive weak supervision for processes via back-propagation from outcome-level rewards. Subsequently, we integrate ORMs and PRMs under the same pairwise ranking objectives, resulting in our newly developed reward models, named OpenPRM. This approach significantly enhances the scalability of process-level supervision in open domains at minimal cost. We assess the performance of OpenPRM across various reward benchmarks, demonstrating its competitive edge over traditional ORMs in open domains and PRMs in specialized domains. Additionally, we investigate the scalability of inference-time computation for open-domain instructions. Our results highlight the limitations of ORMs’ scalability, while OpenPRM shows superior performance in scaled settings. Despite these advances, achieving automatic fine-grained supervision for open-domain inference-time scaling remains a substantial challenge. We hope these findings will spur further development of process supervision reward models in open-domain scenarios.
****
Ruohong Liu,Yuxin Pan,Linjie Xu,Lei Song,Pengcheng You,Yize Chen,Jiang Bian
Efficient Discovery of Pareto Front for Multi-Objective Reinforcement Learning
Multi-objective reinforcement learning (MORL) excels at handling rapidly changing preferences in tasks that involve multiple criteria, even for unseen preferences. However, previous dominating MORL methods typically generate a fixed policy set or preference-conditioned policy through multiple training iterations exclusively for sampled preference vectors, and cannot ensure the efficient discovery of the Pareto front. Furthermore, integrating preferences into the input of policy or value functions presents scalability challenges, in particular as the dimension of the state and preference space grow, which can complicate the learning process and hinder the algorithm's performance on more complex tasks. To address these issues, we propose a two-stage Pareto front discovery algorithm called Constrained MORL (C-MORL), which serves as a seamless bridge between constrained policy optimization and MORL. Concretely, a set of policies is trained in parallel in the initialization stage, with each optimized towards its individual preference over the multiple objectives. Then, to fill the remaining vacancies in the Pareto front, the constrained optimization steps are employed to maximize one objective while constraining the other objectives to exceed a predefined threshold. Empirically, compared to recent advancements in MORL methods, our algorithm achieves more consistent and superior performances in terms of hypervolume, expected utility, and sparsity on both discrete and continuous control tasks, especially with numerous objectives (up to nine objectives in our experiments).
****
Ziyao Shangguan,Chuhan Li,Yuxuan Ding,Yanan Zheng,Yilun Zhao,Tesca Fitzgerald,Arman Cohan
TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models
Existing benchmarks often highlight the remarkable performance achieved by state

-of-the-art Multimodal Foundation Models (MFMs) in leveraging temporal context f
or video understanding.
However, *how well do the models truly perform visual temporal reasoning?*
Our study of existing benchmarks shows that this capability of MFMs is likely ov
erestimated as many questions can be solved by using a single, few, or out-of-or
der frames.
To systematically examine current visual temporal reasoning tasks, we propose th
ree principles with corresponding metrics:
(1) *Multi-Frame Gain*,
(2) *Frame Order Sensitivity*,
and (3) *Frame Information Disparity*.
Following these principles, we introduce **TOMATO**, **T**emp**O**ral Reasoning
Multimod**A**l Evalua**T**i**O**n, a novel benchmark crafted to rigorously a
ssess MFMs' temporal reasoning capabilities in video understanding.
TOMATO comprises 1,484 carefully curated, *human-annotated* questions spanning *
six* tasks (i.e. *action count, direction, rotation, shape & trend, velocity & f
requency, and visual cues*), applied to 1,417 videos, including 805 self-recorde
d and -generated videos, that encompass human-centric, real-world, and simulated
 scenarios.
Our comprehensive evaluation reveals a human-model performance gap of 57.3% with
 the best-performing model.
Moreover, our in-depth analysis uncovers more fundamental limitations beyond thi
s gap in current MFMs. While they can accurately recognize events in isolated fr
ames, they fail to interpret these frames as a continuous sequence.
We believe TOMATO will serve as a crucial testbed for evaluating the next-genera
tion MFMs and as a call to the community to develop AI systems capable of compre
hending the human world dynamics through the video modality.
****
Yu Feng,Ben Zhou,Weidong Lin,Dan Roth
BIRD: A Trustworthy Bayesian Inference Framework for Large Language Models
Predictive models often need to work with incomplete information in real-world t
asks. Consequently, they must provide reliable probability or confidence estimat
ion, especially in large-scale decision-making and planning tasks. Current large
 language models (LLMs) are insufficient for accurate estimations, but they can
generate relevant factors that may affect the probabilities, produce coarse-grai
ned probabilities when the information is more complete, and help determine whic
h factors are relevant to specific downstream contexts. In this paper, we make u
se of these capabilities of LLMs to provide a significantly more accurate probab
ilistic estimation. We propose BIRD, a novel probabilistic inference framework t
hat aligns a Bayesian network with LLM abductions and then estimates more accura
te probabilities in a deduction step. We show BIRD provides reliable probability
 estimations that are 30% better than those provided directly by LLM baselines.
These estimates further contribute to better and more trustworthy decision makin
g.
****
Zhaowei Zhang,Fengshuo Bai,Qizhi Chen,Chengdong Ma,Mingzhi Wang,Haoran Sun,Zilon
g Zheng,Yaodong Yang
Amulet: ReAlignment During Test Time for Personalized Preference Adaptation of L
LMs
How to align large language models (LLMs) with user preferences from a static ge
neral dataset has been frequently studied. However, user preferences are usually
 personalized, changing, and diverse. This leads to the problem that the actual
user preferences often do not coincide with those trained by the model developer
s in the practical use of LLMs. Since we cannot collect enough data and retrain
for every demand, researching efficient real-time preference adaptation methods
based on the backbone LLMs during test time is important. To this end, we introd
uce **Amulet**, a novel, training-free framework that formulates the decoding pr
ocess of every token as a separate online learning problem with the guidance of
simple user-provided prompts, thus enabling real-time optimization to satisfy us
ers' personalized preferences. To reduce the computational cost brought by this

optimization process for each token, we additionally provide a closed-form solut
ion for each iteration step of the optimization process, thereby reducing the co
mputational time cost to a negligible level. The detailed experimental results d
emonstrate that Amulet can achieve significant performance improvements in rich
settings with combinations of different LLMs, datasets, and user preferences, wh
ile maintaining acceptable computational efficiency.
****
Xianghui Ze,Zhenbo Song,Qiwei Wang,Jianfeng Lu,Yujiao Shi
Controllable Satellite-to-Street-View Synthesis with Precise Pose Alignment and
Zero-Shot Environmental Control
Generating street-view images from satellite imagery is a challenging task, part
icularly in maintaining accurate pose alignment and incorporating diverse enviro
nmental conditions. While diffusion models have shown promise in generative task
s, their ability to maintain strict pose alignment throughout the diffusion proc
ess is limited. In this paper, we propose a novel Iterative Homography Adjustmen
t (IHA) scheme applied during the denoising process, which effectively addresses
 pose misalignment and ensures spatial consistency in the generated street-view
images. Additionally, currently, available datasets for satellite-to-street-view
 generation are limited in their diversity of illumination and weather condition
s, thereby restricting the generalizability of the generated outputs. To mitigat
e this, we introduce a text-guided illumination and weather-controlled sampling
strategy that enables fine-grained control over the environmental factors. Exten
sive quantitative and qualitative evaluations demonstrate that our approach sign
ificantly improves pose accuracy and enhances the diversity and realism of gener
ated street-view images, setting a new benchmark for satellite-to-street-view ge
neration tasks.
****
Ziyu Liu,Yuhang Zang,Xiaoyi Dong,Pan Zhang,Yuhang Cao,Haodong Duan,Conghui He,Yu
anjun Xiong,Dahua Lin,Jiaqi Wang
MIA-DPO: Multi-Image Augmented Direct Preference Optimization For Large Vision-L
anguage Models
Visual preference alignment involves training Large Vision-Language Models (LVLM
s) to predict human preferences between visual inputs. This is typically achieve
d by using labeled datasets of chosen/rejected pairs and employing optimization
algorithms like direct preference optimization (DPO).
Existing visual alignment methods, primarily designed for single-image scenarios
, struggle to effectively handle the complexity of multi-image tasks due to the
scarcity of diverse training data and the high cost of annotating chosen/rejecte
d pairs.
We present Multi-Image Augmented Direct Preference Optimization (MIA-DPO), a vis
ual preference alignment approach that effectively handles multi-image inputs.
MIA-DPO mitigates the scarcity of diverse multi-image training data by extending
 single-image data with unrelated images arranged in grid collages or pic-in-pic
 formats, significantly reducing the costs associated with multi-image data anno
tations.
Our observation reveals that attention values of LVLMs vary considerably across
different images. We use attention values to identify and filter out rejected re
sponses the model may have mistakenly focused on.
Our attention-aware selection for constructing the chosen/rejected pairs without
 relying on (i) human annotation, (ii) extra data, and (iii) external models or
APIs.
MIA-DPO is compatible with various architectures and outperforms existing method
s on five multi-image benchmarks, achieving an average performance boost of 3.0%
 on LLaVA-v1.5 and 4.3% on the recent InternLM-XC2.5.
Moreover, MIA-DPO has a minimal effect on the model's ability to understand sing
le images.
****
Zhen Guo,Adriana Meza Soria,Wei Sun,Yikang Shen,Rameswar Panda
API Pack: A Massive Multi-Programming Language Dataset for API Call Generation
We introduce API Pack, a massive multi-programming language dataset containing o

ver one million instruction-API calls for improving the API call generation capa
bilities of large language models. Our evaluation highlights three key findings:
 First, fine-tuning on API Pack enables open-source models to outperform GPT-3.5
 and GPT-4 in generating code for entirely new API calls. We show this by fine-t
uning CodeLlama-13B on 20,000 Python instances from API Pack. Second, fine-tunin
g on a large dataset in one language, combined with smaller datasets from others
, improves API generation accuracy across multiple languages. Third, we confirm
the benefits of larger datasets for API generalization, as increasing fine-tunin
g data to one million instances enhances generalization to new APIs. To support
further research, we open-source the API Pack dataset, trained model, and code a
t https://github.com/zguo0525/API-Pack.
****

Wenhong Zhu,Zhiwei He,Xiaofeng Wang,Pengfei Liu,Rui Wang

Weak-to-Strong Preference Optimization: Stealing Reward from Weak Aligned Model

Aligning language models (LMs) with human preferences has become a key area of r
esearch, enabling these models to meet diverse user needs better. Inspired by we
ak-to-strong generalization, where a strong LM fine-tuned on labels generated by
 a weaker model can consistently outperform its weak supervisor, we extend this
idea to model alignment. In this work, we observe that the alignment behavior in
 weaker models can be effectively transferred to stronger models and even exhibi
t an amplification effect. Based on this insight, we propose a method called Wea
k-to-Strong Preference Optimization (WSPO), which achieves strong model alignmen
t by learning the distribution differences before and after the alignment of the
 weak model. Experiments demonstrate that WSPO delivers outstanding performance,
 improving the win rate of Qwen2-7B-Instruct on Arena-Hard from 39.70 to 49.60,
achieving a remarkable 47.04 length-controlled win rate on AlpacaEval 2, and sco
ring 7.33 on MT-bench. Our results suggest that using the weak model to elicit a
 strong model with a high alignment ability is feasible. The code is available a
t https://github.com/zwhong714/weak-to-strong-preference-optimization.
****

Xiusheng Huang,Jiaxiang Liu,Yequan Wang,Jun Zhao,Kang Liu

Capability Localization: Capabilities Can be Localized rather than Individual Kn
owledge

Large scale language models have achieved superior performance in tasks related
to natural language processing, however, it is still unclear how model parameter
s affect performance improvement. Previous studies assumed that individual knowl
edge is stored in local parameters, and the storage form of individual knowledge
 is dispersed parameters, parameter layers, or parameter chains, which are not u
nified. We found through fidelity and reliability evaluation experiments that in
dividual knowledge cannot be localized. Afterwards, we constructed a dataset for
 decoupling experiments and discovered the potential for localizing data commona
lities. To further reveal this phenomenon, this paper proposes a **C**ommonality
 Neuron **L**ocalization (**CNL**) method, which successfully locates common
ality neurons and achieves a neuron overlap rate of 96.42% on the GSM8K dataset.
 Finally, we have demonstrated through cross data experiments that commonality n
eurons are a collection of capability neurons that possess the capability to enh
ance performance. Our code is available at https://github.com/nlpkeg/Capability-
Neuron-Localization.
****

Giuseppe Serra,Florian Buettner

Federated Continual Learning Goes Online: Uncertainty-Aware Memory Management fo
r Vision Tasks and Beyond

Given the ability to model more realistic and dynamic problems, Federated Contin
ual Learning (FCL) has been increasingly investigated recently. A well-known pro
blem encountered in this setting is the so-called catastrophic forgetting, for w
hich the learning model is inclined to focus on more recent tasks while forgetti
ng the previously learned knowledge. The majority of the current approaches in F
CL propose generative-based solutions to solve said problem. However, this setti
ng requires multiple training epochs over the data, implying an offline setting
where datasets are stored locally and remain unchanged over time. Furthermore, t

he proposed solutions are tailored for vision tasks solely. To overcome these li
mitations, we propose a new approach to deal with different modalities in the on
line scenario where new data arrive in streams of mini-batches that can only be
processed once. To solve catastrophic forgetting, we propose an uncertainty-awar
e memory-based approach. Specifically, we suggest using an estimator based on t
he Bregman Information (BI) to compute the model's variance at the sample level.
 Through measures of predictive uncertainty, we retrieve samples with specific c
haracteristics, and – by retraining the model on such samples – we demonstrate t
he potential of this approach to reduce the forgetting effect in realistic setti
ngs while maintaining data confidentiality and competitive communication efficie
ncy compared to state-of-the-art approaches.
****

Ziqing Fan,Siyuan Du,Shengchao Hu,Pingjie Wang,Li Shen,Ya Zhang,Dacheng Tao,Yanf
eng Wang
Combatting Dimensional Collapse in LLM Pre-Training Data via Submodular File Sel
ection
Selecting high-quality pre-training data for large language models (LLMs) is cru
cial for enhancing their overall performance under limited computation budget, i
mproving both training and sample efficiency. Recent advancements in file select
ion primarily rely on using an existing or trained proxy model to assess the sim
ilarity of samples to a target domain, such as high quality sources BookCorpus a
nd Wikipedia. However, upon revisiting these methods, the domain-similarity sele
ction criteria demonstrates a diversity dilemma, i.e. dimensional collapse in th
e feature space, improving performance on the domain-related tasks but causing s
evere degradation on generic performance.To prevent collapse and enhance diversi
ty, we propose a DiverSified File selection algorithm (DiSF), which selects the
most decorrelated text files in the feature space. We approach this with a class
ical greedy algorithm to achieve more uniform eigenvalues in the feature covaria
nce matrix of the selected texts, analyzing its approximation to the optimal sol
ution under a formulation of γ-weakly submodular optimization problem. Em
pirically, we establish a benchmark and conduct extensive experiments on the Tin
yLlama architecture with models from 120M to 1.1B parameters. Evaluating across
nine tasks from the Harness framework, DiSF demonstrates a significant improveme
nt on overall performance. Specifically, DiSF saves 98.5\% of 590M training file
s in SlimPajama, outperforming the full-data pre-training within a 50B training
budget, and achieving about 1.5x training efficiency and 5x data efficiency. Sou
rce code
is available at: https://github.com/MediaBrain-SJTU/DiSF.git.
****

Harin Lee,Taehyun Hwang,Min-hwan Oh
Lasso Bandit with Compatibility Condition on Optimal Arm
We consider a stochastic sparse linear bandit problem where only a sparse subset
 of context features affects the expected reward function, i.e., the unknown rew
ard parameter has a sparse structure.
In the existing Lasso bandit literature, the compatibility conditions, together
with additional diversity conditions on the context features are imposed to achi
eve regret bounds that only depend logarithmically on the ambient dimension d.
In this paper, we demonstrate that even without the additional diversity assumpt
ions, the \textit{compatibility condition on the optimal arm} is sufficient to d
erive a regret bound that depends logarithmically on d, and our assumption is
strictly weaker than those used in the lasso bandit literature under the single-
parameter setting.
We propose an algorithm that adapts the forced-sampling technique and prove that
 the proposed algorithm achieves $\mathcal{O}(\text{poly}\log dT)$ regret under
the margin condition.
To our knowledge, the proposed algorithm requires the weakest assumptions among
Lasso bandit algorithms under the single-parameter setting that achieve $\mathca
l{O}(\text{poly}\log dT)$ regret.
Through numerical experiments, we confirm the superior performance of our propos
ed algorithm.

****
Cevahir Koprulu,Thiago D. Simão,Nils Jansen,ufuk topcu
Safety-Prioritizing Curricula for Constrained Reinforcement Learning
Curriculum learning aims to accelerate reinforcement learning (RL) by generating curricula, i.e., sequences of tasks of increasing difficulty.
Although existing curriculum generation approaches provide benefits in sample efficiency, they overlook safety-critical settings where an RL agent must adhere to safety constraints.
Thus, these approaches may generate tasks that cause RL agents to violate safety constraints during training and behave suboptimally after.
We develop a safe curriculum generation approach (SCG) that aligns the objectives of constrained RL and curriculum learning: improving safety during training and boosting sample efficiency.
SCG generates sequences of tasks where the RL agent can be safe and performant by initially generating tasks with minimum safety violations over high-reward ones.
We empirically show that compared to the state-of-the-art curriculum learning approaches and their naively modified safe versions, SCG achieves optimal performance and the lowest amount of constraint violations during training.
****
Yihui Qiu,Deepu Rajan
Action Sequence Augmentation for Action Anticipation
Action anticipation models require an understanding of temporal action patterns and dependencies to predict future actions from previous events. The key challenges arise from the vast number of possible action sequences, given the flexibility in action ordering and the interleaving of multiple goals. Since only a subset of such action sequences are present in action anticipation datasets, there is an inherent ordering bias in them. Another challenge is the presence of noisy input to the models due to erroneous action recognition or other upstream tasks.
This paper addresses these challenges by introducing a novel data augmentation strategy that separately augments observed action sequences and next actions. To address biased action ordering, we introduce a grammar induction algorithm that derives a powerful context-free grammar from action sequence data. We also develop an efficient parser to generate plausible next-action candidates beyond the ground truth. For noisy input, we enhance model robustness by randomly deleting or replacing actions in observed sequences. Our experiments on the 50Salads, EGTEA Gaze+, and Epic-Kitchens-100 datasets demonstrate significant performance improvements over existing state-of-the-art methods.
****
Jiawei Chen,Wentao Chen,Jing Su,Jingjing Xu,Hongyu Lin,Mengjie Ren,Yaojie Lu,Xianpei Han,Le Sun
The Rise and Down of Babel Tower: Investigating the Evolution Process of Multilingual Code Large Language Model
Large language models (LLMs) have shown significant multilingual capabilities. However, the mechanisms underlying the development of these capabilities during pre-training are not well understood. In this paper, we use code LLMs as an experimental platform to explore the evolution of multilingual capabilities in LLMs during the pre-training process. Based on our observations, we propose the Babel Tower Hypothesis, which describes the entire process of LLMs acquiring new language capabilities. During the learning process, multiple languages initially share a single knowledge system dominated by the primary language and gradually develop language-specific knowledge systems. We then validate the above hypothesis by tracking the internal states of the LLM using specific methods. Experimental results show that the internal state changes of the LLM are consistent with our Babel Tower Hypothesis. Building on these insights, we propose a novel method to construct an optimized pre-training corpus for multilingual code LLMs, which significantly outperforms LLMs trained on the original corpus. The proposed Babel Tower Hypothesis provides new insights into designing pre-training data distributions to achieve optimal multilingual capabilities in LLMs.
****

Hyunsu Kim,Giung Nam,Chulhee Yun,Hongseok Yang,Juho Lee
Parameter Expanded Stochastic Gradient Markov Chain Monte Carlo
Bayesian Neural Networks (BNNs) provide a promising framework for modeling predictive uncertainty and enhancing out-of-distribution robustness (OOD) by estimating the posterior distribution of network parameters. Stochastic Gradient Markov Chain Monte Carlo (SGMCMC) is one of the most powerful methods for scalable posterior sampling in BNNs, achieving efficiency by combining stochastic gradient descent with second-order Langevin dynamics. However, SGMCMC often suffers from limited sample diversity in practice, which affects uncertainty estimation and model performance. We propose a simple yet effective approach to enhance sample diversity in SGMCMC without the need for tempering or running multiple chains. Our approach reparameterizes the neural network by decomposing each of its weight matrices into a product of matrices, resulting in a sampling trajectory that better explores the target parameter space. This approach produces a more diverse set of samples, allowing faster mixing within the same computational budget. Notably, our sampler achieves these improvements without increasing the inference cost compared to the standard SGMCMC. Extensive experiments on image classification tasks, including OOD robustness, diversity, loss surface analyses, and a comparative study with Hamiltonian Monte Carlo, demonstrate the superiority of the proposed approach.
****
Ahmed Imtiaz Humayun,Ibtihel Amara,Cristina Nader Vasconcelos,Deepak Ramachandran,Candice Schumann,Junfeng He,Katherine A Heller,Golnoosh Farnadi,Negar Rostamzadeh,Mohammad Havaei
What Secrets Do Your Manifolds Hold? Understanding the Local Geometry of Generative Models
Deep Generative Models are frequently used to learn continuous representations of complex data distributions using a finite number of samples. For any generative model, including pre-trained foundation models with GAN, Transformer or Diffusion architectures, generation performance can vary significantly based on which part of the learned data manifold is sampled. In this paper we study the post-training local geometry of the learned manifold and its relationship to generation outcomes for models ranging from toy settings to the latent decoder of the near state-of-the-art Stable Diffusion 1.4 Text-to-Image model. Building on the theory of continuous piecewise-linear (CPWL) generators, we characterize the local geometry in terms of three geometric descriptors - scaling (ψ), rank (ν), and complexity (δ). We provide quantitative and qualitative evidence showing that for a given latent, the local descriptors are indicative of generation aesthetics, artifacts, diversity, and memorization. Finally we demonstrate that training a reward model using the local geometry allows us to control the log-likelihood of a generated sample under the learned distribution, and improve the qualitative aspects of an image.
****
Elvis Dohmatob,Yunzhen Feng,Arjun Subramonian,Julia Kempe
Strong Model Collapse
Within the scaling laws paradigm, which underpins the training of large neural networks like ChatGPT and Llama, we consider a supervised regression setting and establish a strong form of the model collapse phenomenon, a critical performance degradation due to synthetic data in the training corpus. Our results show that even the smallest fraction of synthetic data (e.g., as little as 1 per 1000) can still lead to model collapse: larger and larger training sets do not enhance performance. We further investigate whether increasing model size, an approach aligned with current trends in training large language models, exacerbates or mitigates model collapse. In a simplified regime where neural networks are approximated via random projections of tunable size, we both theoretically and empirically show that larger models can amplify model collapse. Interestingly, our theory also indicates that, beyond the interpolation threshold (which can be extremely high for very large datasets), larger models may mitigate the collapse, although they do not entirely prevent it. Our theoretical findings are empirically verified through experiments on language models and neural networks for images.

****
Bruno Kacper Mlodozeniec,Runa Eschenhagen,Juhan Bae,Alexander Immer,David Kruege r,Richard E. Turner
Influence Functions for Scalable Data Attribution in Diffusion Models
Diffusion models have led to significant advancements in generative modelling. Y et their widespread adoption poses challenges regarding data attribution and int erpretability. In this paper, we aim to help address such challenges in diffusio n models by extending influence functions. Influence function-based data attribu tion methods approximate how a model's output would have changed if some trainin g data were removed. In supervised learning, this is usually used for predicting how the loss on a particular example would change. For diffusion models, we foc us on predicting the change in the probability of generating a particular exampl e via several proxy measurements. We show how to formulate influence functions f or such quantities and how previously proposed methods can be interpreted as par ticular design choices in our framework. To ensure scalability of the Hessian co mputations in influence functions, we use a K-FAC approximation based on general ised Gauss-Newton matrices specifically tailored to diffusion models. We show th at our recommended method outperforms previously proposed data attribution metho ds on common data attribution evaluations, such as the Linear Data-modelling Sco re (LDS) or retraining without top influences, without the need for method-speci fic hyperparameter tuning.
****
Tsiry Mayet,Pourya Shamsolmoali,Simon Bernard,Eric Granger,Romain HÉRAULT,Clemen t Chatelain
TD-Paint: Faster Diffusion Inpainting Through Time-Aware Pixel Conditioning
Diffusion models have emerged as highly effective techniques for inpainting, how ever, they remain constrained by slow sampling rates. While recent advances have enhanced generation quality, they have also increased sampling time, thereby li miting scalability in real-world applications. We investigate the generative sam pling process of diffusion-based inpainting models and observe that these models make minimal use of the input condition during the initial sampling steps. As a result, the sampling trajectory deviates from the data manifold, requiring comp lex synchronization mechanisms to realign the generation process. To address thi s, we propose Time-aware Diffusion Paint (TD-Paint), a novel approach that adapt s the diffusion process by modeling variable noise levels at the pixel level. Th is technique allows the model to efficiently use known pixel values from the sta rt, guiding the generation process toward the target manifold. By embedding this information early in the diffusion process, TD-Paint significantly accelerates sampling without compromising image quality. Unlike conventional diffusion-based inpainting models, which require a dedicated architecture or an expensive gener ation loop, TD-Paint achieves faster sampling times without architectural modifi cations. Experimental results across three datasets show that TD-Paint outperfor ms state-of-the-art diffusion models while maintaining lower complexity.
****
Chenxin An,Jun Zhang,Ming Zhong,Lei Li,Shansan Gong,Yao Luo,Jingjing Xu,Lingpeng Kong
Why Does the Effective Context Length of LLMs Fall Short?
Advancements in distributed training and efficient attention mechanisms have sig nificantly expanded the context window sizes of large language models (LLMs). Ho wever, recent work reveals that the effective context lengths of open-source LLM s often fall short, typically not exceeding half of their training lengths. In t his work, we attribute this limitation to the left-skewed frequency distribution of relative positions formed in LLMs pretraining and post-training stages, whic h impedes their ability to effectively gather distant information.
To address this challenge, we introduce Shifted Rotray Position Embedding (STRIN G). STRING shifts well-trained positions to overwrite the original ineffective positions during inference, enhancing performance within their existing trainin g lengths.
Experimental results show that without additional training, STRING dramatically improves the performance of the latest large-scale models, such as Llama3.1 70B

and Qwen2 72B, by over 10 points on popular long-context benchmarks RULER and In finiteBench, establishing new state-of-the-art results for open-source LLMs. Com pared to commercial models, Llama 3.1 70B with STRING even achieves better perf ormance than GPT-4-128K and clearly surpasses Claude 2 and Kimi-chat.
****

Yudi Xie,Weichen Huang,Esther Alter,Jeremy Schwartz,Joshua B. Tenenbaum,James J. DiCarlo

Vision CNNs trained to estimate spatial latents learned similar ventral-stream-a ligned representations

Studies of the functional role of the primate ventral visual stream have traditi onally focused on object categorization, often ignoring -- despite much prior ev idence -- its role in estimating "spatial" latents such as object position and p ose. Most leading ventral stream models are derived by optimizing networks for o bject categorization, which seems to imply that the ventral stream is also deriv ed under such an objective. Here, we explore an alternative hypothesis: Might th e ventral stream be optimized for estimating spatial latents? And a closely rela ted question: How different -- if at all -- are representations learned from spa tial latent estimation compared to categorization? To ask these questions, we le veraged synthetic image datasets generated by a 3D graphic engine and trained co nvolutional neural networks (CNNs) to estimate different combinations of spatial and category latents. We found that models trained to estimate just a few spati al latents achieve neural alignment scores comparable to those trained on hundre ds of categories, and the spatial latent performance of models strongly correlat es with their neural alignment. Spatial latent and category-trained models have very similar -- but not identical -- internal representations, especially in the ir early and middle layers. We provide evidence that this convergence is partly driven by non-target latent variability in the training data, which facilitates the implicit learning of representations of those non-target latents. Taken toge ther, these results suggest that many training objectives, such as spatial laten ts, can lead to similar models aligned neurally with the ventral stream. Thus, o ne should not assume that the ventral stream is optimized for object categorizat ion only. As a field, we need to continue to sharpen our measures of comparing m odels to brains to better understand the functional roles of the ventral stream.
****

Xingyu Jiang,Ning Gao,Xiuhui Zhang,Hongkun Dou,Yue Deng

Value-aligned Behavior Cloning for Offline Reinforcement Learning via Bi-level O ptimization

Offline reinforcement learning (RL) aims to optimize policies under pre-collecte d data, without requiring any further interactions with the environment. Derived from imitation learning, Behavior cloning (BC) is extensively utilized in offli ne RL for its simplicity and effectiveness. Although BC inherently avoids out-of -distribution deviations, it lacks the ability to discern between high and low-q uality data, potentially leading to sub-optimal performance when facing with poo r-quality data. Current offline RL algorithms attempt to enhance BC by incorpora ting value estimation, yet often struggle to effectively balance these two criti cal components, specifically the alignment between the behavior policy and the p re-trained value estimations under in-sample offline data. To address this chall enge, we propose the Value-aligned Behavior Cloning via Bi-level Optimization (V ACO), a novel bi-level framework that seamlessly integrates an inner loop for we ighted supervised behavior cloning (BC) with an outer loop dedicated to value al ignment. In this framework, the inner loop employs a meta-scoring network to eva luate and appropriately weight each training sample, while the outer loop maximi zes value estimation for alignment with controlled noise to facilitate limited e xploration. This bi-level structure allows VACO to identify the optimal weighted BC policy, ultimately maximizing the expected estimated return conditioned on t he learned value function. We conduct a comprehensive evaluation of VACO across a variety of continuous control benchmarks in offline RL, where it consistently achieves superior performance compared to existing state-of-the-art methods.
****

Yiming Zhang,Javier Rando,Ivan Evtimov,Jianfeng Chi,Eric Michael Smith,Nicholas

Carlini,Florian Tramèr,Daphne Ippolito
Persistent Pre-training Poisoning of LLMs
Large language models are pre-trained on uncurated text datasets consisting of t
rillions of tokens scraped from the Web.
Prior work has shown that: (1) web-scraped pre-training datasets can be practica
lly poisoned by malicious actors; and (2) adversaries can compromise language mo
dels after poisoning fine-tuning datasets.
Our work evaluates for the first time whether language models can also be \emph{
compromised during pre-training}, with a focus on the persistence of pre-trainin
g attacks after models are fine-tuned as helpful and harmless chatbots (i.e., af
ter SFT and DPO).
We pre-train a series of LLMs from scratch to measure the impact of a potential
poisoning adversary under four different attack objectives (denial-of-service, b
elief manipulation, jailbreaking, and prompt stealing), and across a wide range
of model sizes (from 600M to 7B).
Our main result is that poisoning only 0.1% of a model's pre-training dataset is
 sufficient for three out of four attacks to measurably persist through post-tra
ining. Moreover, simple attacks like denial-of-service persist through post-trai
ning with a poisoning rate of only 0.001%.
****
Hangtao Zhang,Chenyu Zhu,Xianlong Wang,Ziqi Zhou,Changgan Yin,Minghui Li,Lulu Xu
e,Yichen Wang,Shengshan Hu,Aishan Liu,Peijin Guo,Leo Yu Zhang
BadRobot: Jailbreaking Embodied LLM Agents in the Physical World
Embodied AI represents systems where AI is integrated into physical entities. Mu
ltimodal Large Language Model (LLM), which exhibits powerful language understand
ing abilities, has been extensively employed in embodied AI by facilitating soph
isticated task planning. However, a critical safety issue remains overlooked: co
uld these embodied LLMs perpetrate harmful behaviors? In response, we introduce
BadRobot, the first attack paradigm designed to jailbreak robotic manipulation,
making embodied LLMs violate safety and ethical constraints through typical voic
e-based user-system interactions. Specifically, three vulnerabilities are exploi
ted to achieve this type of attack: (i) manipulation of LLMs within robotic syst
ems, (ii) misalignment between linguistic outputs and physical actions, and
(iii) unintentional hazardous behaviors caused by world knowledge's flaws. Furth
ermore, we construct a benchmark of various malicious physical action queries to
 evaluate BadRobot's attack performance. Based on this benchmark, extensive expe
riments against existing prominent embodied LLM frameworks (e.g., Voxposer, Code
 as Policies, and ProgPrompt) demonstrate the effectiveness of our BadRobot. We
emphasize that addressing this emerging vulnerability is crucial for the secure
deployment of LLMs in robotics.

Warning: This paper contains harmful AI-generated language and aggressive action
s.
****
Théo Uscidda,Luca Eyring,Karsten Roth,Fabian J Theis,Zeynep Akata,marco cuturi
Disentangled Representation Learning with the Gromov-Monge Gap
Learning disentangled representations from unlabelled data is a fundamental chal
lenge in machine learning. Solving it may unlock other problems, such as general
ization, interpretability, or fairness. Although remarkably challenging to solve
 in theory, disentanglement is often achieved in practice through prior matching
. Furthermore, recent works have shown that prior matching approaches can be enh
anced by leveraging geometrical considerations, e.g., by learning representation
s that preserve geometric features of the data, such as distances or angles betw
een points. However, matching the prior while preserving geometric features is c
hallenging, as a mapping that *fully* preserves these features while aligning th
e data distribution with the prior does not exist in general. To address these c
hallenges, we introduce a novel approach to disentangled representation learning
 based on quadratic optimal transport. We formulate the problem using Gromov-Mon
ge maps that transport one distribution onto another with minimal distortion of
predefined geometric features, preserving them *as much as can be achieved*. To

compute such maps, we propose the Gromov-Monge-Gap (GMG), a regularizer quantifying whether a map moves a reference distribution with minimal geometry distortion. We demonstrate the effectiveness of our approach for disentanglement across four standard benchmarks, outperforming other methods leveraging geometric considerations.
****
Kaiwen Zheng,Guande He,Jianfei Chen,Fan Bao,Jun Zhu
Diffusion Bridge Implicit Models
Denoising diffusion bridge models (DDBMs) are a powerful variant of diffusion models for interpolating between two arbitrary paired distributions given as endpoints. Despite their promising performance in tasks like image translation, DDBMs require a computationally intensive sampling process that involves the simulation of a (stochastic) differential equation through hundreds of network evaluations. In this work, we take the first step in fast sampling of DDBMs without extra training, motivated by the well-established recipes in diffusion models. We generalize DDBMs via a class of non-Markovian diffusion bridges defined on the discretized timesteps concerning sampling, which share the same marginal distributions and training objectives, give rise to generative processes ranging from stochastic to deterministic, and result in diffusion bridge implicit models (DBIMs). DBIMs are not only up to 25$\times$ faster than the vanilla sampler of DDBMs but also induce a novel, simple, and insightful form of ordinary differential equation (ODE) which inspires high-order numerical solvers. Moreover, DBIMs maintain the generation diversity in a distinguished way, by using a booting noise in the initial sampling step, which enables faithful encoding, reconstruction, and semantic interpolation in image translation tasks. Code is available at \url{https://github.com/thu-ml/DiffusionBridge}.
****
Yuxiang Lu,Shengcao Cao,Yu-Xiong Wang
Swiss Army Knife: Synergizing Biases in Knowledge from Vision Foundation Models for Multi-Task Learning
Vision Foundation Models (VFMs) have demonstrated outstanding performance on numerous downstream tasks. However, due to their inherent representation biases originating from different training paradigms, VFMs exhibit advantages and disadvantages across distinct vision tasks. Although amalgamating the strengths of multiple VFMs for downstream tasks is an intuitive strategy, effectively exploiting these biases remains a significant challenge. In this paper, we propose a novel and versatile "Swiss Army Knife" (SAK) solution, which adaptively distills knowledge from a committee of VFMs to enhance multi-task learning. Unlike existing methods that use a single backbone for knowledge transfer, our approach preserves the unique representation bias of each teacher by collaborating the lightweight Teacher-Specific Adapter Path modules with the Teacher-Agnostic Stem. Through dynamic selection and combination of representations with Mixture-of-Representations Routers, our SAK is capable of synergizing the complementary strengths of multiple VFMs. Extensive experiments show that our SAK remarkably outperforms prior state of the arts in multi-task learning by 10% on the NYUD-v2 benchmark, while also providing a flexible and robust framework that can readily accommodate more advanced model designs.
****
Anish Dhir,Matthew Ashman,James Requeima,Mark van der Wilk
A Meta-Learning Approach to Bayesian Causal Discovery
Discovering a unique causal structure is difficult due to both inherent identifiability issues, and the consequences of finite data.
As such, uncertainty over causal structures, such as those obtained from a Bayesian posterior, are often necessary for downstream tasks.
Finding an accurate approximation to this posterior is challenging, due to the large number of possible causal graphs, as well as the difficulty in the subproblem of finding posteriors over the functional relationships of the causal edges.
Recent works have used Bayesian meta learning to view the problem of posterior estimation as a supervised learning task.
Yet, these methods are limited as they cannot reliably sample from the posterior

over causal structures and fail to encode key properties of the posterior, such as correlation between edges and permutation equivariance with respect to nodes.
To address these limitations, we propose a Bayesian meta learning model that allows for sampling causal structures from the posterior and encodes these key properties.
We compare our meta-Bayesian causal discovery against existing Bayesian causal discovery methods, demonstrating the advantages of directly learning a posterior over causal structure.
****
Gihyun Kwon,Jong Chul Ye
TweedieMix: Improving Multi-Concept Fusion for Diffusion-based Image/Video Generation
Despite significant advancements in customizing text-to-image and video generation models, generating images and videos that effectively integrate multiple personalized concepts remains challenging. To address this, we present TweedieMix, a novel method for composing customized diffusion models during the inference phase. By analyzing the properties of reverse diffusion sampling, our approach divides the sampling process into two stages. During the initial steps, we apply a multiple object-aware sampling technique to ensure the inclusion of the desired target objects. In the later steps, we blend the appearances of the custom concepts in the de-noised image space using Tweedie's formula. Our results demonstrate that TweedieMix can generate multiple personalized concepts with higher fidelity than existing methods. Moreover, our framework can be effortlessly extended to image-to-video diffusion models by extending the residual layer's features across frames}, enabling the generation of videos that feature multiple personalized concepts.
****
Gautam Chandrasekaran,Adam Klivans,Lin Lin Lee,Konstantinos Stavropoulos
Learning Neural Networks with Distribution Shift: Efficiently Certifiable Guarantees
We give the first provably efficient algorithms for learning neural networks with respect to distribution shift. We work in the Testable Learning with Distribution Shift framework (TDS learning) of Klivans et al. (2024), where the learner receives labeled examples from a training distribution and unlabeled examples from a test distribution and must either output a hypothesis with low test error or reject if distribution shift is detected. No assumptions are made on the test distribution.

All prior work in TDS learning focuses on classification, while here we must handle the setting of nonconvex regression. Our results apply to real-valued networks with arbitrary Lipschitz activations and work whenever the training distribution has strictly sub-exponential tails. For training distributions that are bounded and hypercontractive, we give a fully polynomial-time algorithm for TDS learning one hidden-layer networks with sigmoid activations. We achieve this by importing classical kernel methods into the TDS framework using data-dependent feature maps and a type of kernel matrix that couples samples from both train and test distributions.
****
Raeid Saqur,Anastasis Kratsios,Florian Krach,Yannick Limmer,Blanka Horvath,Frank Rudzicz
Filtered not Mixed: Filtering-Based Online Gating for Mixture of Large Language Models
We propose MoE-F — a formalized mechanism for combining N pre-trained expert Large Language Models (LLMs) in online time-series prediction tasks by adaptively forecasting the best weighting of LLM predictions at every time step. Our mechanism leverages the conditional information in each expert's running performance to forecast the best combination of LLMs for predicting the time series in its next step. Diverging from static (learned) Mixture of Experts (MoE) methods, our approach employs time-adaptive stochastic filtering techniques to combine experts.

By framing the expert selection problem as a finite state-space, continuous-time Hidden Markov model (HMM), we can leverage the Wohman-Shiryaev filter. Our approach first constructs N parallel filters corresponding to each of the N individual LLMs. Each filter proposes its best combination of LLMs, given the information that they have access to. Subsequently, the N filter outputs are optimally aggregated to maximize their robust predictive power, and this update is computed efficiently via a closed-form expression, thus generating our ensemble predictor.

Our contributions are:

- **(I)** the MoE-F algorithm — deployable as a plug-and-play filtering harness,

- **(II)** theoretical optimality guarantees of the proposed filtering-based gating algorithm (via optimality guarantees for its parallel Bayesian filtering and its robust aggregation steps), and

- **(III)** empirical evaluation and ablative results using state-of-the-art foundational and MoE LLMs on a real-world _Financial Market Movement_ task where MoE-F attains a remarkable 17% absolute and 48.5% relative F1 measure improvement over the next best performing individual LLM expert predicting short-horizon market movement based on streaming news. Further, we provide empirical evidence of substantial performance gains in applying MoE-F over specialized models in the _long-horizon time-series forecasting_ domain. Code available on github: https://github.com/raeidsaqur/moe-f
****
Felix Jedidja Binder,James Chua,Tomek Korbak,Henry Sleight,John Hughes,Robert Long,Ethan Perez,Miles Turpin,Owain Evans
Looking Inward: Language Models Can Learn About Themselves by Introspection
Humans acquire knowledge by observing the external world, but also by introspection. Introspection gives a person privileged access to their current state of mind (e.g. thoughts and feelings) that are not accessible to external observers. Do LLMs have this introspective capability of privileged access? If they do, this would show that LLMs can acquire knowledge not contained in or inferable from training data.
We investigate LLMs predicting properties of their own behavior in hypothetical situations. If a model M1 has this capability, it should outperform a different model M2 in predicting M1's behavior—even if M2 is trained on M1's ground-truth behavior.
The idea is that M1 has privileged access to its own behavioral tendencies, and this enables it to predict itself better than M2 (even if M2 is generally stronger).
In experiments with GPT-4, GPT-4o, and Llama-3 models, we find that the model M1 outperforms M2 in predicting itself, providing evidence for privileged access.
Further experiments and ablations provide additional evidence.
Our results show that LLMs can offer reliable self-information independent of external data in certain domains. By demonstrating this, we pave the way for further work on introspection in more practical domains, which would have significant implications for model transparency and explainability. However, while we successfully show introspective capabilities in simple tasks, we are unsuccessful on more complex tasks or those requiring out-of-distribution generalization.
****
Chenguo Lin,Panwang Pan,Bangbang Yang,Zeming Li,Yadong MU
DiffSplat: Repurposing Image Diffusion Models for Scalable Gaussian Splat Generation
Recent advancements in 3D content generation from text or a single image struggle with limited high-quality 3D datasets and inconsistency from 2D multi-view generation. We introduce DiffSplat, a novel 3D generative framework that natively generates 3D Gaussian splats by taming large-scale text-to-image diffusion models. It differs from previous 3D generative models by effectively utilizing web-sca

le 2D priors while maintaining 3D consistency in a unified model. To bootstrap t
he training, a lightweight reconstruction model is proposed to instantly produce
 multi-view Gaussian splat grids for scalable dataset curation. In conjunction w
ith the regular diffusion loss on these grids, a 3D rendering loss is introduced
 to facilitate 3D coherence across arbitrary views. The compatibility with image
 diffusion models enables seamless adaptions of numerous techniques for image ge
neration to the 3D realm. Extensive experiments reveal the superiority of DiffSp
lat in text- and image-conditioned generation tasks and downstream applications.
 Thorough ablation studies validate the efficacy of each critical design choice
and provide insights into the underlying mechanism.
****

Yeoreum Lee,Jinwook Jung,Sungyong Baik
Mitigating Parameter Interference in Model Merging via Sharpness-Aware Fine-Tuni
ng
Large-scale deep learning models with a pretraining-finetuning paradigm have led
 to a surge of numerous task-specific models fine-tuned from a common pre-traine
d model.
Recently, several research efforts have been made on merging these large models
into a single multi-task model, particularly with simple arithmetic on parameter
s.
Such merging methodology faces a central challenge: interference between model p
arameters fine-tuned on different tasks.
Few recent works have focused on designing a new fine-tuning scheme that can lea
d to small parameter interference, however at the cost of the performance of eac
h task-specific fine-tuned model and thereby limiting that of a merged model.
To improve the performance of a merged model, we note that a fine-tuning scheme
should aim for (1) smaller parameter interference and (2) better performance of
each fine-tuned model on the corresponding task.
In this work, we aim to design a new fine-tuning objective function to work towa
rds these two goals.
In the course of this process, we find such objective function to be strikingly
similar to sharpness-aware minimization (SAM) objective function, which aims to
achieve generalization by finding flat minima.
Drawing upon our observation, we propose to fine-tune pre-trained models via sha
rpness-aware minimization.
The experimental and theoretical results showcase the effectiveness and orthogon
ality of our proposed approach, improving performance upon various merging and f
ine-tuning methods.
Our code is available at https://github.com/baiklab/SAFT-Merge.
****

Qingmao Yao,Zhichao Lei,Tianyuan Chen,Ziyue Yuan,Xuefan Chen,Jianxiang Liu,Faguo
 Wu,Xiao Zhang
Offline RL with Smooth OOD Generalization in Convex Hull and its Neighborhood
Offline Reinforcement Learning (RL) struggles with distributional shifts, leadin
g to the Q-value overestimation for out-of-distribution (OOD) actions. Existin
g methods address this issue by imposing constraints; however, they often become
 overly conservative when evaluating OOD regions, which constrains the Q-funct
ion generalization. This over-constraint issue results in poor Q-value estimat
ion and hinders policy improvement. In this paper, we introduce a novel approach
 to achieve better Q-value estimation by enhancing Q-function generalization
 in OOD regions within Convex Hull and its Neighborhood (CHN). Under the safety
generalization guarantees of the CHN, we propose the Smooth Bellman Operator (SB
O), which updates OOD Q-values by smoothing them with neighboring in-sample $Q
$-values. We theoretically show that SBO approximates true Q-values for both i
n-sample and OOD actions within the CHN. Our practical algorithm, Smooth Q-funct
ion OOD Generalization (SQOG), empirically alleviates the over-constraint issue,
 achieving near-accurate Q-value estimation. On the D4RL benchmarks, SQOG outp
erforms existing state-of-the-art methods in both performance and computational
efficiency. Code is available at <https://github.com/yqpqry/SQOG>.
****

Seongyun Lee,Geewook Kim,Jiyeon Kim,Hyunji Lee,Hoyeon Chang,Sue Hyun Park,Minjoon Seo
How Does Vision-Language Adaptation Impact the Safety of Vision Language Models?
Vision-Language adaptation (VL adaptation) transforms Large Language Models (LLMs) into Large Vision-Language Models (LVLMs) for multimodal tasks, but this process often compromises the inherent safety capabilities embedded in the original LLMs. Despite potential harmfulness due to weakened safety measures, in-depth analysis on the effects of VL adaptation on safety remains under-explored. This study examines how VL adaptation influences safety and evaluates the impact of safety fine-tuning methods. Our analysis reveals that safety degradation occurs during VL adaptation, even when the training data is safe. While safety tuning techniques like supervised fine-tuning with safety datasets or reinforcement learning from human feedback mitigate some risks, they still lead to safety degradation and a reduction in helpfulness due to over-rejection issues. Further analysis of internal model weights suggests that VL adaptation may impact certain safety-related layers, potentially lowering overall safety levels. Additionally, our findings demonstrate that the objectives of VL adaptation and safety tuning are divergent, which often results in their simultaneous application being suboptimal. To address this, we suggest the weight merging approach as an optimal solution effectively reducing safety degradation while maintaining helpfulness. These insights help guide the development of more reliable and secure LVLMs for real-world applications.
****
Guojin Zhong,pan wang,Jin Yuan,Zhiyong Li,Long Chen
Multi-Resolution Decomposable Diffusion Model for Non-Stationary Time Series Anomaly Detection
Recently, generative models have shown considerable promise in unsupervised time series anomaly detection. Nonetheless, the task of effectively capturing complex temporal patterns and minimizing false alarms becomes increasingly challenging when dealing with non-stationary time series, characterized by continuously fluctuating statistical attributes and joint distributions. To confront these challenges, we underscore the benefits of multi-resolution modeling, which improves the ability to distinguish between anomalies and non-stationary behaviors by leveraging correlations across various resolution scales. In response, we introduce a **M**ulti-Res**o**lution **De**composable Diffusion **M**odel (MODEM), which integrates a coarse-to-fine diffusion paradigm with a frequency-enhanced decomposable network to adeptly navigate the intricacies of non-stationarity. Technically, the coarse-to-fine diffusion model embeds cross-resolution correlations into the forward process to optimize diffusion transitions mathematically. It then innovatively employs low-resolution recovery to guide the reverse trajectories of high-resolution series in a coarse-to-fine manner, enhancing the model's ability to learn and elucidate underlying temporal patterns. Furthermore, the frequency-enhanced decomposable network operates in the frequency domain to extract globally shared time-invariant information and time-variant temporal dynamics for accurate series reconstruction. Extensive experiments conducted across five real-world datasets demonstrate that our proposed MODEM achieves state-of-the-art performance and can be generalized to other time series tasks.
****
Zihan Qiu,Zeyu Huang,Shuang Cheng,Yizhi Zhou,Zili Wang,Ivan Titov,Jie Fu
Layerwise Recurrent Router for Mixture-of-Experts
The scaling of large language models (LLMs) has revolutionized their capabilities in various tasks, yet this growth must be matched with efficient computational strategies.
The Mixture-of-Experts (MoE) architecture stands out for its ability to scale model size without significantly increasing training costs.
Despite their advantages, current MoE models often display parameter inefficiency.
For instance, a pre-trained MoE-based LLM with 52 billion parameters might perform comparably to a standard model with 6.7 billion.
Being a crucial part of MoE,

current routers in different layers independently assign tokens without leveragi
ng historical routing information, potentially leading to suboptimal token-exper
t combinations and the parameter inefficiency problem.
To alleviate this issue, we introduce the Layerwise Recurrent Router for Mixture
-of-Experts (RMoE).
RMoE leverages a Gated Recurrent Unit (GRU) to establish dependencies between ro
uting decisions across consecutive layers.
Such layerwise recurrence can be efficiently parallelly computed for input token
s and introduces negotiable costs.
Our extensive empirical evaluations demonstrate that RMoE-based language models
consistently outperform a spectrum of baseline models.
Furthermore, RMoE integrates a novel computation stage orthogonal to existing me
thods, allowing seamless compatibility with other MoE architectures.
Our analyses attribute RMoE's gains to its effective cross-layer information sha
ring, which also improves expert selection and diversity.
****
Xin Ding,Xiaoyu Liu,Zhijun Tu,Yun Zhang,Wei Li,Jie Hu,Hanting Chen,Yehui Tang,Zh
iwei Xiong,Baoqun Yin,Yunhe Wang
CBQ: Cross-Block Quantization for Large Language Models
Post-training quantization (PTQ) has played a pivotal role in compressing large
language models (LLMs) at ultra-low costs. Although current PTQ methods have ach
ieved promising results by addressing outliers and employing layer- or block-wis
e loss optimization techniques, they still suffer from significant performance d
egradation at ultra-low bits precision. To dissect this issue, we conducted an i
n-depth analysis of quantization errors specific to LLMs and surprisingly discov
ered that, unlike traditional sources of quantization errors, the growing number
 of model parameters, combined with the reduction in quantization bits, intensif
ies inter-layer and intra-layer dependencies, which severely impact quantization
 accuracy. This finding highlights a critical challenge in quantizing LLMs. To a
ddress this, we propose CBQ, a cross-block reconstruction-based PTQ method for L
LMs. CBQ leverages a cross-block dependency to establish long-range dependencies
 across multiple blocks and integrates an adaptive LoRA-Rounding technique to ma
nage intra-layer dependencies. To further enhance performance, CBQ incorporates
a coarse-to-fine pre-processing mechanism for processing weights and activations
. Extensive experiments show that CBQ achieves superior low-bit quantization (W4
A4, W4A8, W2A16) and outperforms existing state-of-the-art methods across variou
s LLMs and datasets. Notably, CBQ only takes 4.3 hours to quantize a weight-only
 quantization of a 4-bit LLAMA1-65B model, achieving a commendable trade off bet
ween performance and efficiency.
****
Roman Belaire,Arunesh Sinha,Pradeep Varakantham
On Minimizing Adversarial Counterfactual Error in Adversarial Reinforcement Lear
ning
Deep Reinforcement Learning (DRL) policies are highly susceptible to adversarial
 noise in observations, which poses significant risks in safety-critical scenari
os. The challenge inherent to adversarial perturbations is that by altering the
information observed by the agent, the state becomes only partially observable.
Existing approaches address this by either enforcing consistent actions across n
earby states or maximizing the worst-case value within adversarially perturbed o
bservations. However, the former suffers from performance degradation when attac
ks succeed, while the latter tends to be overly conservative, leading to subopti
mal performance in benign settings. We hypothesize that these limitations stem f
rom their failing to account for partial observability directly. To this end, we
 introduce a novel objective called Adversarial Counterfactual Error (ACoE), def
ined on the beliefs about the true state and balancing value optimization with r
obustness. To make ACoE scalable in model-free settings, we propose the theoreti
cally-grounded surrogate objective Cumulative-ACoE (C-ACoE). Our empirical evalu
ations on standard benchmarks (MuJoCo, Atari, and Highway) demonstrate that our
method significantly outperforms current state-of-the-art approaches for address
ing adversarial RL challenges, offering a promising direction for improving robu

stness in DRL under adversarial conditions. Our code is available at https://gitHub.com/romanbelaire/acoe-robust-rl.
****
Tao Feng,Yanzhen Shen,Jiaxuan You
GraphRouter: A Graph-based Router for LLM Selections
The rapidly growing number and variety of Large Language Models (LLMs) present significant challenges in efficiently selecting the appropriate LLM for a given query, especially considering the trade-offs between performance and computational cost. Current LLM selection methods often struggle to generalize across new LLMs and different tasks because of their limited ability to leverage contextual interactions among tasks, queries, and LLMs, as well as their dependence on a transductive learning framework. To address these shortcomings, we introduce a novel inductive graph framework, named as GraphRouter, which fully utilizes the contextual information among tasks, queries, and LLMs to enhance the LLM selection process. GraphRouter constructs a heterogeneous graph comprising task, query, and LLM nodes, with interactions represented as edges, which efficiently captures the contextual information between the query's requirements and the LLM's capabilities. Through an innovative edge prediction mechanism, GraphRouter is able to predict attributes (the effect and cost of LLM response) of potential edges, allowing for optimized recommendations that adapt to both existing and newly introduced LLMs without requiring retraining. Comprehensive experiments across three distinct effect-cost weight scenarios have
shown that GraphRouter substantially surpasses existing routers, delivering a minimum performance improvement of 12.3%. In addition, it achieves enhanced generalization across new LLMs settings and supports diverse tasks with at least a
9.5% boost in effect and a significant reduction in computational demands. This work endeavors to apply a graph-based approach for the contextual and adaptive selection of LLMs, offering insights for real-world applications.
****
Zeliang Zhang,Susan Liang,Daiki Shimada,Chenliang Xu
Rethinking Audio-Visual Adversarial Vulnerability from Temporal and Modality Perspectives
While audio-visual learning equips models with a richer understanding of the real world by leveraging multiple sensory modalities, this integration also introduces new vulnerabilities to adversarial attacks.
In this paper, we present a comprehensive study of the adversarial robustness of audio-visual models, considering both temporal and modality-specific vulnerabilities. We propose two powerful adversarial attacks: 1) a temporal invariance attack that exploits the inherent temporal redundancy across consecutive time segments and 2) a modality misalignment attack that introduces incongruence between the audio and visual modalities. These attacks are designed to thoroughly assess the robustness of audio-visual models against diverse threats. Furthermore, to defend against such attacks, we introduce a novel audio-visual adversarial training framework. This framework addresses key challenges in vanilla adversarial training by incorporating efficient adversarial perturbation crafting tailored to multi-modal data and an adversarial curriculum strategy. Extensive experiments in the Kinetics-Sounds dataset demonstrate that our proposed temporal and modality-based attacks in degrading model performance can achieve state-of-the-art performance, while our adversarial training defense largely improves the adversarial robustness as well as the adversarial training efficiency.
****
Martin Andrae,Tomas Landelius,Joel Oskarsson,Fredrik Lindsten
Continuous Ensemble Weather Forecasting with Diffusion models
Weather forecasting has seen a shift in methods from numerical simulations to data-driven systems. While initial research in the area focused on deterministic forecasting, recent works have used diffusion models to produce skillful ensemble forecasts. These models are trained on a single forecasting step and rolled out autoregressively. However, they are computationally expensive and accumulate er

rors for high temporal resolution due to the many rollout steps. We address thes
e limitations with Continuous Ensemble Forecasting, a novel and flexible method
for sampling ensemble forecasts in diffusion models. The method can generate tem
porally consistent ensemble trajectories completely in parallel, with no autoreg
ressive steps. Continuous Ensemble Forecasting can also be combined with autoreg
ressive rollouts to yield forecasts at an arbitrary fine temporal resolution wit
hout sacrificing accuracy. We demonstrate that the method achieves competitive r
esults for global weather forecasting with good probabilistic properties.
****

Zihan Zhou,Xiaoxue Wang,Tianshu Yu
Generating Physical Dynamics under Priors
Generating physically feasible dynamics in a data-driven context is challenging,
 especially when adhering to physical priors expressed in specific equations or
formulas. Existing methodologies often overlook the integration of ''physical pr
iors'', resulting in violation of basic physical laws and suboptimal performance
. In this paper, we introduce a novel framework that seamlessly incorporates phy
sical priors into diffusion-based generative models to address this limitation.
Our approach leverages two categories of priors: 1) distributional priors, such
as roto-translational invariance, and 2) physical feasibility priors, including
energy and momentum conservation laws and PDE constraints. By embedding these pr
iors into the generative process, our method can efficiently generate physically
 realistic dynamics, encompassing trajectories and flows. Empirical evaluations
demonstrate that our method produces high-quality dynamics across a diverse arra
y of physical phenomena with remarkable robustness, underscoring its potential t
o advance data-driven studies in AI4Physics. Our contributions signify a substan
tial advancement in the field of generative modeling, offering a robust solution
 to generate accurate and physically consistent dynamics.
****

Ru Peng,Kexin Yang,Yawen Zeng,Junyang Lin,Dayiheng Liu,Junbo Zhao
DataMan: Data Manager for Pre-training Large Language Models
The performance emergence of large language models (LLMs) driven by data scaling
 laws makes the selection of pre-training data increasingly important.
However, existing methods rely on limited heuristics and human intuition, lackin
g comprehensive and clear guidelines.
To address this, we are inspired by *``reverse thinking''* -- prompting LLMs to
self-identify which criteria benefit its performance.
As its pre-training capabilities are related to perplexity (PPL), we derive 14 q
uality criteria from the causes of text perplexity anomalies and introduce 15 co
mmon application domains to support domain mixing.
In this paper, we train a **Data** **Man**ager (**DataMan**) to learn quality ra
tings and domain recognition from pointwise rating, and use it to annotate a 447
B token pre-training corpus with 14 quality ratings and domain type.
Our experiments validate our approach, using DataMan to select 30B tokens to tra
in a 1.3B-parameter language model, demonstrating significant improvements in in
-context learning (ICL), perplexity, and instruction-following ability over the
state-of-the-art baseline.
The best-performing model, based on the *Overall Score l=5* surpasses a model tr
ained with 50% more data using uniform sampling.
We continue pre-training with high-rated, domain-specific data annotated by Data
Man to enhance domain-specific ICL performance and thus verify DataMan's domain
mixing ability.
Our findings emphasize the importance of quality ranking, the complementary natu
re of quality criteria, and their low correlation with perplexity, analyzing mis
alignment between PPL and ICL performance.
We also thoroughly analyzed our pre-training dataset, examining its composition,
 the distribution of quality ratings, and the original document sources.
****

Ryan Lucas,Rahul Mazumder
Preserving Deep Representations in One-Shot Pruning: A Hessian-Free Second-Order
 Optimization Framework

We present SNOWS, a one-shot post-training pruning framework aimed at reducing t
he cost of vision network inference without retraining. Current leading one-shot
 pruning methods minimize layer-wise least squares reconstruction error which do
es not take into account deeper network representations. We propose to optimize
a more global reconstruction objective. This objective accounts for nonlinear ac
tivations deep in the network to obtain a better proxy for the network loss. Thi
s nonlinear objective leads to a more challenging optimization problem---we demo
nstrate it can be solved efficiently using a specialized second-order optimizati
on framework. A key innovation of our framework is the use of Hessian-free optim
ization to compute exact Newton descent steps without needing to compute or stor
e the full Hessian matrix. A distinct advantage of SNOWS is that it can be readi
ly applied on top of any sparse mask derived from prior methods, readjusting the
ir weights to preserve deep feature representations. SNOWS obtains state-of-the-
art results on various one-shot pruning benchmarks including residual networks a
nd Vision Transformers (ViT/B-16 and ViT/L-16, 86m and 304m parameters respectiv
ely). Our open-source implementation is available at https://github.com/mazumder
-lab/SNOWS.
****

Huimin LU,Masaru Isonuma,Junichiro Mori,Ichiro Sakata
UniDetox: Universal Detoxification of Large Language Models via Dataset Distilla
tion
We present UniDetox, a universally applicable method designed to mitigate toxici
ty across various large language models (LLMs).
Previous detoxification methods are typically model-specific, addressing only in
dividual models or model families, and require careful hyperparameter tuning due
 to the trade-off between detoxification efficacy and language modeling performa
nce.
In contrast, UniDetox provides a detoxification technique that can be universall
y applied to a wide range of LLMs without the need for separate model-specific t
uning.
Specifically, we propose a novel and efficient dataset distillation technique fo
r detoxification using contrastive decoding.
This approach distills detoxifying representations in the form of synthetic text
 data, enabling universal detoxification of any LLM through fine-tuning with the
 distilled text.
Our experiments demonstrate that the detoxifying text distilled from GPT-2 can e
ffectively detoxify larger models, including OPT, Falcon, and LLaMA-2.
Furthermore, UniDetox eliminates the need for separate hyperparameter tuning for
 each model, as a single hyperparameter configuration can be seamlessly applied
across different models.
Additionally, analysis of the detoxifying text reveals a reduction in politicall
y biased content, providing insights into the attributes necessary for effective
 detoxification of LLMs.
****

Hanseul Cho,Jaeyoung Cha,Srinadh Bhojanapalli,Chulhee Yun
Arithmetic Transformers Can Length-Generalize in Both Operand Length and Count
Transformers often struggle with *length generalization*, meaning they fail to g
eneralize to sequences longer than those encountered during training. While arit
hmetic tasks are commonly used to study length generalization, certain tasks are
 considered notoriously difficult, e.g., multi-operand addition (requiring gener
alization over both the number of operands and their lengths) and multiplication
 (requiring generalization over both operand lengths). In this work, we achieve
approximately 2–3× length generalization on both tasks, which is the first such
achievement in arithmetic Transformers. We design task-specific scratchpads enab
ling the model to focus on a fixed number of tokens per each next-token predicti
on step, and apply multi-level versions of *Position Coupling* (Cho et al., 2024
; McLeish et al., 2024) to let Transformers know the right position to attend to
. On the theory side, we prove that a 1-layer Transformer using our method can s
olve multi-operand addition, up to operand length and operand count that are exp
onential in embedding dimension.

****
Leheng Sheng,An Zhang,Yi Zhang,Yuxin Chen,Xiang Wang,Tat-Seng Chua
Language Representations Can be What Recommenders Need: Findings and Potentials
Recent studies empirically indicate that language models (LMs) encode rich world knowledge beyond mere semantics, attracting significant attention across various fields.
However, in the recommendation domain, it remains uncertain whether LMs implicitly encode user preference information. Contrary to prevailing understanding that LMs and traditional recommenders learn two distinct representation spaces due to the huge gap in language and behavior modeling objectives, this work re-examines such understanding and explores extracting a recommendation space directly from the language representation space.
Surprisingly, our findings demonstrate that item representations, when linearly mapped from advanced LM representations, yield superior recommendation performance.
This outcome suggests the possible homomorphism between the advanced language representation space and an effective item representation space for recommendation, implying that collaborative signals may be implicitly encoded within LMs.
Motivated by the finding of homomorphism, we explore the possibility of designing advanced collaborative filtering (CF) models purely based on language representations without ID-based embeddings.
To be specific, we incorporate several crucial components (i.e., a multilayer perceptron (MLP), graph convolution, and contrastive learning (CL) loss function) to build a simple yet effective model, with the language representations of item textual metadata (i.e., title) as the input.
Empirical results show that such a simple model can outperform leading ID-based CF models on multiple datasets, which sheds light on using language representations for better recommendation.
Moreover, we systematically analyze this simple model and find several key features for using advanced language representations:
a good initialization for item representations, superior zero-shot recommendation abilities in new datasets, and being aware of user intention.
Our findings highlight the connection between language modeling and behavior modeling, which can inspire both natural language processing and recommender system communities.
****
Alexandre Variengien,Eric Winsor
Look Before You Leap: Universal Emergent Mechanism for Retrieval in Language Models
When solving challenging problems, language models (LMs) are able to identify relevant information from long and complicated contexts. To study how LMs solve retrieval tasks in diverse situations, we introduce ORION, a collection of structured retrieval tasks spanning six domains, from text understanding to coding. Each task in ORION can be represented abstractly by a request (e.g. a question) that retrieves an attribute (e.g. the character name) from a context (e.g. a story). We apply causal analysis on 18 open-source language models with sizes ranging from 125 million to 70 billion parameters. We find that LMs internally decompose retrieval tasks in a modular way: middle layers at the last token position process the request, while late layers retrieve the correct entity from the context. After causally enforcing this decomposition, models are still able to solve the original task, preserving 70% of the original correct token probability in 98 of the 106 studied model-task pairs. We connect our macroscopic decomposition with a microscopic description by performing a fine-grained case study of a question-answering task on Pythia-2.8b. Building on our high-level understanding, we demonstrate a proof of concept application for scalable internal oversight of LMs to mitigate prompt-injection while requiring human supervision on only a single input. Our solution improves accuracy drastically (from 15.5% to 97.5% on Pythia-12b). This work presents evidence of a universal emergent modular processing of tasks across varied domains and models and is a pioneering effort in applying interpretability for scalable internal oversight of LMs.

****
Majid Daliri,Juliana Freire,Danrong Li,Christopher Musco
Matrix Product Sketching via Coordinated Sampling
We revisit the well-studied problem of approximating a matrix product, $\bv{A}^T \bv{B}$, based on small space sketches $\mathcal{S}(\bv{A})$ and $\mathcal{S}(\bv{B})$ of $\bv{A} \in \R^{n \times d}$ and $\bv{B}\in \R^{n \times m}$. We are interested in the setting where the sketches must be computed independently of each other, except for the use of a shared random seed. We prove that, when $\bv{A}$ and $\bv{B}$ are sparse, methods based on \emph{coordinated random sampling} can outperform classical linear sketching approaches, like Johnson-Lindenstrauss Projection or CountSketch. For example, to obtain Frobenius norm error $\epsilon\|\bv{A}\|_F\|\bv{B}\|_F$, coordinated sampling requires sketches of size $O(s/\epsilon^2)$ when $\bv{A}$ and $\bv{B}$ have at most $s \leq d,m$ non-zeros per row. In contrast, linear sketching leads to sketches of size $O(d/\epsilon^2)$ and $O(m/\epsilon^2)$ for $\bv{A}$ and $\bv{B}$. We empirically evaluate our approach on two applications: 1) distributed linear regression in databases, a problem motivated by tasks like dataset discovery and augmentation, and 2) approximating attention matrices in transformer-based language models. In both cases, our sampling algorithms yield an order of magnitude improvement over linear sketching.
****
Jiyeon Kim,Hyunji Lee,Hyowon Cho,Joel Jang,Hyeonbin Hwang,Seungpil Won,Youbin Ahn,Dohaeng Lee,Minjoon Seo
Knowledge Entropy Decay during Language Model Pretraining Hinders New Knowledge Acquisition
In this work, we investigate how a model's tendency to broadly integrate its parametric knowledge evolves throughout pretraining, and how this behavior affects overall performance, particularly in terms of knowledge acquisition and forgetting. We introduce the concept of knowledge entropy, which quantifies the range of memory sources the model engages with; high knowledge entropy indicates that the model utilizes a wide range of memory sources, while low knowledge entropy suggests reliance on specific sources with greater certainty. Our analysis reveals a consistent decline in knowledge entropy as pretraining advances. We also find that the decline is closely associated with a reduction in the model's ability to acquire and retain knowledge, leading us to conclude that diminishing knowledge entropy (smaller number of active memory sources) impairs the model's knowledge acquisition and retention capabilities. We find further support for this by demonstrating that increasing the activity of inactive memory sources enhances the model's capacity for knowledge acquisition and retention.
****
Qizhi Pei,Rui Yan,Kaiyuan Gao,Jinhua Zhu,Lijun Wu
3D-MolT5: Leveraging Discrete Structural Information for Molecule-Text Modeling
The integration of molecular and natural language representations has emerged as a focal point in molecular science, with recent advancements in Language Models (LMs) demonstrating significant potential for comprehensive modeling of both domains. However, existing approaches face notable limitations, particularly in their neglect of three-dimensional (3D) information, which is crucial for understanding molecular structures and functions. While some efforts have been made to incorporate 3D molecular information into LMs using external structure encoding modules, significant difficulties remain, such as insufficient interaction across modalities in pre-training and challenges in modality alignment. To address the limitations, we propose \textbf{3D-MolT5}, a unified framework designed to model molecule in both sequence and 3D structure spaces. The key innovation of our approach lies in mapping fine-grained 3D substructure representations into a specialized 3D token vocabulary. This methodology facilitates the seamless integration of sequence and structure representations in a tokenized format, enabling 3D-MolT5 to encode molecular sequences, molecular structures, and text sequences within a unified architecture. Leveraging this tokenized input strategy, we build a foundation model that unifies the sequence and structure data formats. We then conduct joint pre-training with multi-task objectives to enhance the model's co

mprehension of these diverse modalities within a shared representation space. Th
us, our approach significantly improves cross-modal interaction and alignment, a
ddressing key challenges in previous work. Further instruction tuning demonstrat
ed that our 3D-MolT5 has strong generalization ability and surpasses existing me
thods with superior performance in multiple downstream tasks, such as nearly 70\
% improvement on the molecular property prediction task compared to state-of-the
-art methods. Our code is available at \url{https://github.com/QizhiPei/3D-MolT5
}.
****
Ziyue Li,Tianyi Zhou
Your Mixture-of-Experts LLM Is Secretly an Embedding Model for Free
While large language models (LLMs) excel on generation tasks, their decoder-only
 architecture often limits their potential as embedding models if no further rep
resentation finetuning is applied. Does this contradict their claim of generalis
ts? To answer the question, we take a closer look at Mixture-of-Experts (MoE) LL
Ms. Our study shows that the expert routers in MoE LLMs can serve as an off-the-
shelf embedding model with promising performance on a diverse class of embedding
-focused tasks, without requiring any finetuning. Moreover, our extensive analys
is shows that the MoE routing weights (RW) is complementary to the hidden state
(HS) of LLMs, a widely-used embedding. Compared to HS, we find that RW is more r
obust to the choice of prompts and focuses on high-level semantics. Motivated by
 the analysis, we propose MoEE combining RW and HS, which achieves better perfor
mance than using either separately. Our exploration of their combination and pro
mpting strategy shed several novel insights, e.g., a weighted sum of RW and HS s
imilarities outperforms the similarity on their concatenation. Our experiments a
re conducted on 6 embedding tasks with 20 datasets from the Massive Text Embeddi
ng Benchmark (MTEB). The results demonstrate the significant improvement brought
 by MoEE to LLM-based embedding without further finetuning.
****
Aldo Pareja,Nikhil Shivakumar Nayak,Hao Wang,Krishnateja Killamsetty,Shivchander
 Sudalairaj,Wenlong Zhao,Seungwook Han,Abhishek Bhandwaldar,Guangxuan Xu,Kai Xu,
Ligong Han,Luke Inglis,Akash Srivastava
Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs
The rise of large language models (LLMs) has created a significant disparity: in
dustrial research labs with their computational resources, expert teams, and adv
anced infrastructures, can effectively fine-tune LLMs, while individual develope
rs and small organizations face barriers due to limited resources to effectively
 explore the experiment space. In this paper, we aim to bridge this gap by prese
nting a comprehensive study on supervised fine-tuning of LLMs using instruction-
tuning datasets spanning diverse knowledge domains and skills. We focus on small
-sized LLMs (3B to 7B parameters) for their cost-efficiency and accessibility. W
e explore various training configurations and strategies across four open-source
 pre-trained models. We provide detailed documentation of these configurations,
revealing findings that challenge several common training practices, including h
yperparameter recommendations from TULU and phased training recommended by Orca.
 The code used for the experiments can be found here: https://github.com/instruc
tlab/training.

Key insights from our work include: (i) larger batch sizes paired with lower lea
rning rates lead to improved model performance on benchmarks such as MMLU, MTBen
ch, and Open LLM Leaderboard; (ii) early-stage training dynamics, such as lower
gradient norms and higher loss values, are strong indicators of better final mod
el performance, allowing for early termination of sub-optimal runs and significa
nt computational savings; (iii) through a thorough exploration of hyperparameter
s like warmup steps and learning rate schedules, we provide guidance for practit
ioners and find that certain simplifications do not compromise performance; and
(iv) we observe no significant difference in performance between phased (sequent
ially training on data divided into phases) and stacked (training on the entire
dataset at once) strategies, but stacked training is simpler and more sample eff
icient. With these findings holding robustly across datasets as well as model fa

milies and sizes, we hope this study serves as a guide for practitioners fine-tu
ning small LLMs and promotes a more inclusive research environment for LLM devel
opment.
****

Hongyu Qu,Jianan Wei,Xiangbo Shu,Wenguan Wang
Learning Clustering-based Prototypes for Compositional Zero-Shot Learning
Learning primitive (i.e., attribute and object) concepts from seen compositions
is the primary challenge of Compositional Zero-Shot Learning (CZSL). Existing CZ
SL solutions typically rely on oversimplified data assumptions, e.g., modeling e
ach primitive with a single centroid primitive presentation, ignoring the natura
l diversities of the attribute (resp. object) when coupled with different object
s (resp. attribute). In this work, we develop ClusPro, a robust clustering-based
 prototype mining framework for CZSL that defines the conceptual boundaries of p
rimitives through a set of diversified prototypes. Specifically, ClusPro conduct
s within-primitive clustering on the embedding space for automatically discoveri
ng and dynamically updating prototypes. To learn high-quality embeddings for dis
criminative prototype construction, ClusPro repaints a well-structured and indep
endent primitive embedding space, ensuring intra-primitive separation and inter-
primitive decorrelation through prototype-based contrastive learning and decorre
lation learning. Moreover, ClusPro effectively performs prototype clustering in
a non-parametric fashion without the introduction of additional learnable parame
ters or computational budget during testing. Experiments on three benchmarks dem
onstrate ClusPro outperforms various top-leading CZSL solutions under both close
d-world and open-world settings. Our code is available at CLUSPRO.
****

Terry Tong,Fei Wang,Zhe Zhao,Muhao Chen
BadJudge: Backdoor Vulnerabilities of LLM-As-A-Judge
This paper proposes a novel backdoor threat attacking the LLM-as-a-Judge evaluat
ion regime, where the adversary controls both the candidate and evaluator model.
 The backdoored evaluator victimizes benign users by unfairly assigning inflated
 scores to adversary. A trivial single token backdoor poisoning 1% of the evalua
tor training data triples the adversary's score with respect to their legitimate
 score. We systematically categorize levels of data access corresponding to thre
e real-world settings, (1) web poisoning, (2) malicious annotator, and (3) weigh
t poisoning. These regimes reflect a weak to strong escalation of data access th
at highly correlates with attack severity. Under the weakest assumptions - web p
oisoning (1), the adversary still induces a 20% score inflation. Likewise, in t
he (3) weight poisoning regime, the stronger assumptions enable the adversary to
 inflate their scores from 1.5/5 to 4.9/5. The backdoor threat generalizes acros
s different evaluator architectures, trigger designs, evaluation tasks, and pois
oning rates. By poisoning 10% of the evaluator training data, we control toxicit
y judges (Guardrails) to misclassify toxic prompts as non-toxic 89% of the time,
  and document reranker judges in RAG to rank the poisoned document first 97% of
 the time. LLM-as-a-Judge is uniquely positioned at the intersection of ethics a
nd technology, where social implications of mislead model selection and evaluati
on constrain the available defensive tools. Amidst these challenges, model mergi
ng emerges as a principled tool to offset the backdoor, reducing ASR to near 0%
whilst maintaining SOTA performance. Model merging's low computational cost and
convenient integration into the current LLM Judge training pipeline position it
as a promising avenue for backdoor mitigation in the LLM-as-a-Judge setting.
****

Guiyu Zhang,Huan-ang Gao,Zijian Jiang,Hao Zhao,Zhedong Zheng
Ctrl-U: Robust Conditional Image Generation via Uncertainty-aware Reward Modelin
g
In this paper, we focus on the task of conditional image generation, where an im
age is synthesized according to user instructions. The critical challenge underp
inning this task is ensuring both the fidelity of the generated images and their
 semantic alignment with the provided conditions. To tackle this issue, previous
 studies have employed supervised perceptual losses derived from pre-trained mod
els, i.e., reward models, to enforce alignment between the condition and the gen

erated result. However, we observe one inherent shortcoming: considering the div
ersity of synthesized images, the reward model usually provides inaccurate feedb
ack when encountering newly generated data, which can undermine the training pro
cess. To address this limitation, we propose an uncertainty-aware reward modelin
g, called Ctrl-U, including uncertainty estimation and uncertainty-aware regular
ization, designed to reduce the adverse effects of imprecise feedback from the r
eward model. Given the inherent cognitive uncertainty within reward models, even
 images generated under identical conditions often result in a relatively large
discrepancy in reward loss. Inspired by the observation, we explicitly leverage
such prediction variance as an uncertainty indicator. Based on the uncertainty e
stimation, we regularize the model training by adaptively rectifying the reward.
 In particular, rewards with lower uncertainty receive higher loss weights, whil
e those with higher uncertainty are given reduced weights to allow for larger va
riability. The proposed uncertainty regularization facilitates reward fine-tunin
g through consistency construction. Extensive experiments validate the effective
ness of our methodology in improving the controllability and generation quality,
 as well as its scalability across diverse conditional scenarios, including segm
entation mask, edge, and depth conditions.
****

Giuseppe Bruno,Federico Pasqualotto,Andrea Agazzi
Emergence of meta-stable clustering in mean-field transformer models
We model the evolution of tokens within a deep stack of Transformer layers as a
continuous-time flow on the unit sphere, governed by a mean-field interacting pa
rticle system, building on the framework introduced in Geshkovski et al. (2023).
 Studying the corresponding mean-field Partial Differential Equation (PDE), whic
h can be interpreted as a Wasserstein gradient flow, in this paper we provide a
mathematical investigation of the long-term behavior of this system, with a part
icular focus on the emergence and persistence of meta-stable phases and clusteri
ng phenomena, key elements in applications like next-token prediction. More spec
ifically, we perform a perturbative analysis of the mean-field PDE around the ii
d uniform initialization and prove that, in the limit of large number of tokens,
 the model remains close to a meta-stable manifold of solutions with a given str
ucture (e.g., periodicity). Further, the structure characterizing the meta-stabl
e manifold is explicitly identified, as a function of the inverse temperature pa
rameter of the model, by the index maximizing a certain rescaling of Gegenbauer
polynomials.
****

Ishank Juneja,Carlee Joe-Wong,Osman Yagan
Pairwise Elimination with Instance-Dependent Guarantees for Bandits with Cost Su
bsidy
Multi-armed bandits (MAB) are commonly used in sequential online decision-making
 when the reward of each decision is an unknown random variable. In practice, ho
wever, the typical goal of maximizing total reward may be less important than mi
nimizing the total cost of the decisions taken, subject to a reward constraint.
For example, we may seek to make decisions that have at least the reward of a re
ference ``default'' decision. This problem was recently introduced in the Multi-
Armed Bandits with Cost Subsidy (MAB-CS) framework. MAB-CS is broadly applicable
 to problem domains where a primary metric (cost) is constrained by a secondary
metric (reward), and there is an inability to explicitly determine the trade-off
 between these metrics. In our work, we first introduce the Pairwise-Elimination
 algorithm for a simplified variant of the cost subsidy problem with a known ref
erence arm. We then generalize PE to PE-CS to solve the MAB-CS problem in the se
tting where the reference arm is the un-identified optimal arm. Next, we analyze
 the performance of both PE and PE-CS on the dual metrics of Cost and Quality Re
gret. Our instance-dependent analysis of PE and PE-CS reveals that both algorith
ms have an order-wise logarithmic upper bound on Cost and Quality Regret, making
 our policy the first with such a guarantee. Finally, experiments are conducted
using the MovieLens 25M dataset for both PE and PE-CS and using a synthetic toy
experiment for PE-CS revealing that our method invariably outperforms the ETC-CS
 baseline from the literature.

****
Xiaojun Jia,Tianyu Pang,Chao Du,Yihao Huang,Jindong Gu,Yang Liu,Xiaochun Cao,Min Lin
Improved Techniques for Optimization-Based Jailbreaking on Large Language Models
Large language models (LLMs) are being rapidly developed, and a key component of their widespread deployment is their safety-related alignment. Many red-teaming efforts aim to jailbreak LLMs, where among these efforts, the Greedy Coordinate Gradient (GCG) attack's success has led to a growing interest in the study of optimization-based jailbreaking techniques. Although GCG is a significant milestone, its attacking efficiency remains unsatisfactory. In this paper, we present several improved (empirical) techniques for optimization-based jailbreaks like GCG. We first observe that the single target template of ”Sure'' largely limits the attacking performance of GCG; given this, we propose to apply diverse target templates containing harmful self-suggestion and/or guidance to mislead LLMs. Besides, from the optimization aspects, we propose an automatic multi-coordinate updating strategy in GCG (i.e., adaptively deciding how many tokens to replace in each step) to accelerate convergence, as well as tricks like easy-to-hard initialization. Then, we combine these improved technologies to develop an efficient jailbreak method, dubbed $\mathcal{I}$-GCG. In our experiments, we evaluate our $\mathcal{I}$-GCG on a series of benchmarks (such as NeurIPS 2023 Red Teaming Track). The results demonstrate that our improved techniques can help GCG outperform state-of-the-art jailbreaking attacks and achieve a nearly 100\% attack success rate.
The code is released at https://github.com/jiaxiaojunQAQ/I-GCG.
****
Sandeep Silwal,David Woodruff,Qiuyi Zhang
Beyond Worst-Case Dimensionality Reduction for Sparse Vectors
We study beyond worst-case dimensionality reduction for s-sparse vectors (vectors with at most s non-zero coordinates). Our work is divided into two parts, each focusing on a different facet of beyond worst-case analysis:

\noindent (a) We first consider average-case guarantees for embedding s-sparse vectors. Here, a well-known folklore upper bound based on the birthday-paradox states: For any collection X of s-sparse vectors in $\mathbb{R}^d$, there exists a linear map $A: \mathbb{R}^d \rightarrow \mathbb{R}^{O(s^2)}$ which \emph{exactly} preserves the norm of 99% of the vectors in X in any ℓ_p norm (as opposed to the usual setting where guarantees hold for all vectors). We provide novel lower bounds showing that this is indeed optimal in many settings. Specifically, any oblivious linear map satisfying similar average-case guarantees must map to $\Omega(s^2)$ dimensions. The same lower bound also holds for a wider class of sufficiently smooth maps, including `encoder-decoder schemes', where we compare the norm of the original vector to that of a smooth function of the embedding. These lower bounds reveal a surprising separation result for smooth embeddings of sparse vectors, as an upper bound of $O(s \log(d))$ is possible if we instead use arbitrary functions, e.g., via compressed sensing algorithms.

(b) Given these lower bounds, we specialize to sparse \emph{non-negative} vectors to hopes of improved upper bounds. For a dataset X of non-negative s-sparse vectors and any $p \ge 1$, we can non-linearly embed X to $O(s\log(|X|s)/\varepsilon^2)$ dimensions while preserving all pairwise distances in ℓ_p norm up to $1\pm \varepsilon$, with no dependence on p. Surprisingly, the non-negativity assumption enables much smaller embeddings than arbitrary sparse vectors, where the best known bound suffers an exponential $(\log |X|)^{O(p)}$ dependence. Our map also guarantees \emph{exact} dimensionality reduction for the ℓ_{∞} norm by embedding X into $O(s\log |X|)$ dimensions, which is tight. We further give separation results showing that both the non-linearity of f and the non-negativity of X are necessary, and provide downstream algorithmic improvements using our embedding.
****

Ruizhe Chen,Xiaotian Zhang,Meng Luo,Wenhao Chai,Zuozhu Liu
PAD: Personalized Alignment at Decoding-time
Aligning with personalized preferences, which vary significantly across cultural, educational, and political differences, poses a significant challenge due to the computational costs and data demands of traditional alignment methods. In response, this paper presents Personalized Alignment at Decoding-time (PAD), a novel framework designed to align LLM outputs with diverse personalized preferences during the inference phase, eliminating the need for additional training. By introducing a unique personalized reward modeling strategy, this framework decouples the text generation process from personalized preferences, facilitating the generation of generalizable token-level personalized rewards. The PAD algorithm leverages these rewards to guide the decoding process, dynamically tailoring the base model’s predictions to personalized preferences. Extensive experimental results demonstrate that PAD not only outperforms existing training-based alignment methods in terms of aligning with diverse preferences but also shows significant generalizability to preferences unseen during training and scalability across different base models. This work advances the capability of LLMs to meet user needs in real-time applications, presenting a substantial step forward in personalized LLM alignment.
****
Junyeob Baek,Yi-Fu Wu,Gautam Singh,Sungjin Ahn
Dreamweaver: Learning Compositional World Models from Pixels
Humans have an innate ability to decompose their perceptions of the world into objects and their attributes, such as colors, shapes, and movement patterns. This cognitive process enables us to imagine novel futures by recombining familiar concepts. However, replicating this ability in artificial intelligence systems has proven challenging, particularly when it comes to modeling videos into compositional concepts and generating unseen, recomposed futures without relying on auxiliary data, such as text, masks, or bounding boxes. In this paper, we propose __Dreamweaver__, a neural architecture designed to discover hierarchical and compositional representations from raw videos and generate compositional future simulations. Our approach leverages a novel Recurrent Block-Slot Unit (RBSU) to decompose videos into their constituent objects and attributes. In addition, Dreamweaver uses a multi-future-frame prediction objective to capture disentangled representations for dynamic concepts more effectively as well as static concepts. In experiments, we demonstrate our model outperforms current state-of-the-art baselines for world modeling when evaluated under the DCI framework across multiple datasets. Furthermore, we show how the modularized concept representations of our model enable compositional imagination, allowing the generation of novel videos by recombining attributes from previously seen objects. cun-bjy.github.io/dreamweaver-website
****
Xiu Yuan,Tongzhou Mu,Stone Tao,Yunhao Fang,Mengke Zhang,Hao Su
Policy Decorator: Model-Agnostic Online Refinement for Large Policy Model
Recent advancements in robot learning have used imitation learning with large models and extensive demonstrations to develop effective policies. However, these models are often limited by the quantity quality, and diversity of demonstrations. This paper explores improving offline-trained imitation learning models through online interactions with the environment. We introduce Policy Decorator, which uses a model-agnostic residual policy to refine large imitation learning models during online interactions. By implementing controlled exploration strategies, Policy Decorator enables stable, sample-efficient online learning. Our evaluation spans eight tasks across two benchmarks—ManiSkill and Adroit—and involves two state-of-the-art imitation learning models (Behavior Transformer and Diffusion Policy). The results show Policy Decorator effectively improves the offline-trained policies and preserves the smooth motion of imitation learning models, avoiding the erratic behaviors of pure RL policies. See our [project page](https://policydecorator.github.io/) for videos.
****
Cong Wang,Kuan Tian,Yonghang Guan,Fei Shen,Zhiwei Jiang,Qing Gu,Jun Zhang

Ensembling Diffusion Models via Adaptive Feature Aggregation
The success of the text-guided diffusion model has inspired the development and release of numerous powerful diffusion models within the open-source community. These models are typically fine-tuned on various expert datasets, showcasing diverse denoising capabilities. Leveraging multiple high-quality models to produce stronger generation ability is valuable, but has not been extensively studied. Existing methods primarily adopt parameter merging strategies to produce a new static model. However, they overlook the fact that the divergent denoising capabilities of the models may dynamically change across different states, such as when experiencing different prompts, initial noises, denoising steps, and spatial locations. In this paper, we propose a novel ensembling method, Adaptive Feature Aggregation (AFA), which dynamically adjusts the contributions of multiple models at the feature level according to various states (i.e., prompts, initial noises, denoising steps, and spatial locations), thereby keeping the advantages of multiple diffusion models, while suppressing their disadvantages. Specifically, we design a lightweight Spatial-Aware Block-Wise (SABW) feature aggregator that adaptive aggregates the block-wise intermediate features from multiple U-Net denoisers into a unified one. The core idea lies in dynamically producing an individual attention map for each model's features by comprehensively considering various states. It is worth noting that only SABW is trainable with about 50 million parameters, while other models are frozen. Both the quantitative and qualitative experiments demonstrate the effectiveness of our proposed method.
****
Seyedmorteza Sadat,Otmar Hilliges,Romann M. Weber
Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models
Classifier-free guidance (CFG) is crucial for improving both generation quality and alignment between the input condition and final output in diffusion models. While a high guidance scale is generally required to enhance these aspects, it also causes oversaturation and unrealistic artifacts. In this paper, we revisit the CFG update rule and introduce modifications to address this issue. We first decompose the update term in CFG into parallel and orthogonal components with respect to the conditional model prediction and observe that the parallel component primarily causes oversaturation, while the orthogonal component enhances image quality. Accordingly, we propose down-weighting the parallel component to achieve high-quality generations without oversaturation. Additionally, we draw a connection between CFG and gradient ascent and introduce a new rescaling and momentum method for the CFG update rule based on this insight. Our approach, termed adaptive projected guidance (APG), retains the quality-boosting advantages of CFG while enabling the use of higher guidance scales without oversaturation. APG is easy to implement and introduces practically no additional computational overhead to the sampling process. Through extensive experiments, we demonstrate that APG is compatible with various conditional diffusion models and samplers, leading to improved FID, recall, and saturation scores while maintaining precision comparable to CFG, making our method a superior plug-and-play alternative to standard classifier-free guidance.
****
Xue Yan,Yan Song,Xidong Feng,Mengyue Yang,Haifeng Zhang,Haitham Bou Ammar,Jun Wang
Efficient Reinforcement Learning with Large Language Model Priors
In sequential decision-making (SDM) tasks, methods like reinforcement learning (RL) and heuristic search have made notable advances in specific cases. However, they often require extensive exploration and face challenges in generalizing across diverse environments due to their limited grasp of the underlying decision dynamics. In contrast, large language models (LLMs) have recently emerged as powerful general-purpose tools, due to their capacity to maintain vast amounts of domain-specific knowledge. To harness this rich prior knowledge for efficiently solving complex SDM tasks, we propose treating LLMs as prior action distributions and integrating them into RL frameworks through Bayesian inference methods, making use of variational inference and direct posterior sampling. The proposed appr

oaches facilitate the seamless incorporation of fixed LLM priors into both polic
y-based and value-based RL frameworks. Our experiments show that incorporating L
LM-based action priors significantly reduces exploration and optimization comple
xity, substantially improving sample efficiency compared to traditional RL techn
iques, e.g., using LLM priors decreases the number of required samples by over 9
0\% in offline learning scenarios.
****
Xiaoming Shi,Shiyu Wang,Yuqi Nie,Dianqi Li,Zhou Ye,Qingsong Wen,Ming Jin
Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts
Deep learning for time series forecasting has seen significant advancements over
 the past decades. However, despite the success of large-scale pre-training in l
anguage and vision domains, pre-trained time series models remain limited in sca
le and operate at a high cost, hindering the development of larger capable forec
asting models in real-world applications. In response, we introduce Time-MoE, a
scalable and unified architecture designed to pre-train larger, more capable for
ecasting foundation models while reducing inference costs. By leveraging a spars
e mixture-of-experts (MoE) design, Time-MoE enhances computational efficiency by
 activating only a subset of networks for each prediction, reducing computationa
l load while maintaining high model capacity. This allows Time-MoE to scale effe
ctively without a corresponding increase in inference costs. Time-MoE comprises
a family of decoder-only transformer models that operate in an auto-regressive m
anner and support flexible forecasting horizons with varying input context lengt
hs. We pre-trained these models on our newly introduced large-scale data Time-30
0B, which spans over 9 domains and encompassing over 300 billion time points. Fo
r the first time, we scaled a time series foundation model up to 2.4 billion par
ameters, achieving significantly improved forecasting precision. Our results val
idate the applicability of scaling laws for training tokens and model size in th
e context of time series forecasting. Compared to dense models with the same num
ber of activated parameters or equivalent computation budgets, our models consis
tently outperform them by large margin. These advancements position Time-MoE as
a state-of-the-art solution for tackling real-world time series forecasting chal
lenges with superior capability, efficiency, and flexibility. Code is available
at https://github.com/Time-MoE/Time-MoE
****
Yunlu Yan,Chun-Mei Feng,Wangmeng Zuo,Lei Zhu,Rick Siow Mong Goh,Yong Liu
Federated Residual Low-Rank Adaption of Large Language Models
Low-Rank Adaptation (LoRA) presents an effective solution for federated fine-tun
ing of Large Language Models (LLMs), as it substantially reduces communication o
verhead. However, a straightforward combination of FedAvg and LoRA results in su
boptimal performance, especially under data heterogeneity. We noted this stems f
rom both intrinsic (i.e., constrained parameter space) and extrinsic (i.e., clie
nt drift) limitations, which hinder it effectively learn global knowledge. In th
is work, we proposed a novel Federated Residual Low-Rank Adaption method, namely
 FRLoRA, to tackle above two limitations. It directly sums the weight of the glo
bal model parameters with a residual low-rank matrix product (\ie, weight change
) during the global update step, and synchronizes this update for all local mode
ls. By this, FRLoRA performs global updates in a higher-rank parameter space, en
abling a better representation of complex knowledge structure. Furthermore, FRLo
RA reinitializes the local low-rank matrices with the principal singular values
 and vectors of the pre-trained weights in each round, to calibrate their incons
istent convergence, thereby mitigating client drift. Our extensive experiments d
emonstrate that FRLoRA consistently outperforms various state-of-the-art FL meth
ods across nine different benchmarks in natural language understanding and gener
ation under different FL scenarios.
****
Karthik Prakhya,Tolga Birdal,Alp Yurtsever
Convex Formulations for Training Two-Layer ReLU Neural Networks
Solving non-convex, NP-hard optimization problems is crucial for training machin
e learning models, including neural networks. However, non-convexity often leads
 to black-box machine learning models with unclear inner workings. While convex

formulations have been used for verifying neural network robustness, their appli
cation to training neural networks remains less explored. In response to this ch
allenge, we reformulate the problem of training infinite-width two-layer ReLU ne
tworks as a convex completely positive program in a finite-dimensional (lifted)
space. Despite the convexity, solving this problem remains NP-hard due to the co
mplete positivity constraint. To overcome this challenge, we introduce a semidef
inite relaxation that can be solved in polynomial time. We then experimentally e
valuate the tightness of this relaxation, demonstrating its competitive performa
nce in test accuracy across a range of classification tasks.
****

Siru Ouyang,Wenhao Yu,Kaixin Ma,Zilin Xiao,Zhihan Zhang,Mengzhao Jia,Jiawei Han,
Hongming Zhang,Dong Yu
RepoGraph: Enhancing AI Software Engineering with Repository-level Code Graph
Large Language Models (LLMs) excel in code generation yet struggle with modern A
I software engineering tasks.
Unlike traditional function-level or file-level coding tasks, AI software engine
ering requires not only basic coding proficiency but also advanced skills in man
aging and interacting with code repositories. However, existing methods often ov
erlook the need for repository-level code understanding, which is crucial for ac
curately grasping the broader context and developing effective solutions. On thi
s basis, we present RepoGraph, a plug-in module that manages a repository-level
structure for modern AI software engineering solutions. RepoGraph offers the des
ired guidance and serves as a repository-wide navigation for AI software enginee
rs. We evaluate RepoGraph on the SWE-bench by plugging it into four different me
thods of two lines of approaches, where RepoGraph substantially boosts the perfo
rmance of all systems, leading to a new state-of-the-art among open-source frame
works. Our analyses also demonstrate the extensibility and flexibility of RepoGr
aph by testing on another repo-level coding benchmark, CrossCodeEval. Our code i
s available at https://github.com/ozyyshr/RepoGraph.
****

Dai Shi,Lequan Lin,Andi Han,Zhiyong Wang,Yi Guo,Junbin Gao
When Graph Neural Networks Meet Dynamic Mode Decomposition
Graph Neural Networks (GNNs) have emerged as fundamental tools for a wide range
of prediction tasks on graph-structured data. Recent studies have drawn analogie
s between GNN feature propagation and diffusion processes, which can be interpre
ted as dynamical systems. In this paper, we delve deeper into this perspective b
y connecting the dynamics in GNNs to modern Koopman theory and its numerical met
hod, Dynamic Mode Decomposition (DMD). We illustrate how DMD can estimate a low-
rank, finite-dimensional linear operator based on multiple states of the system,
 effectively approximating potential nonlinear interactions between nodes in the
 graph. This approach allows us to capture complex dynamics within the graph acc
urately and efficiently. We theoretically establish a connection between the DMD
-estimated operator and the original dynamic operator between system states. Bui
lding upon this foundation, we introduce a family of DMD-GNN models that effecti
vely leverage the low-rank eigenfunctions provided by the DMD algorithm. We furt
her discuss the potential of enhancing our approach by incorporating domain-spec
ific constraints such as symmetry into the DMD computation, allowing the corresp
onding GNN models to respect known physical properties of the underlying system.
 Our work paves the path for applying advanced dynamical system analysis tools v
ia GNNs. We validate our approach through extensive experiments on various learn
ing tasks, including directed graphs, large-scale graphs, long-range interaction
s, and spatial-temporal graphs. We also empirically verify that our proposed mod
els can serve as powerful encoders for link prediction tasks. The results demons
trate that our DMD-enhanced GNNs achieve state-of-the-art performance, highlight
ing the effectiveness of integrating DMD into GNN frameworks.
****

Shen Yan,Xingyan Bin,Sijun Zhang,Yisen Wang,Zhouchen Lin
TC-MoE: Augmenting Mixture of Experts with Ternary Expert Choice
The Mixture of Experts (MoE) architecture has emerged as a promising solution to
 reduce computational overhead by selectively activating subsets of model parame

ters.
The effectiveness of MoE models depends primarily on their routing mechanisms, w
ith the widely adopted Top-K routing scheme used for activating experts.
However, the Top-K scheme has notable limitations,
including unnecessary activations and underutilization of experts.
In this work,
rather than modifying the routing mechanism as done in previous studies,
we propose the Ternary Choice MoE (TC-MoE),
a novel approach that expands the expert space by applying the ternary set {-1,
0, 1} to each expert.
This expansion allows more efficient and effective expert activations without in
curring significant computational costs.
Additionally,
given the unique characteristics of the expanded expert space,
we introduce a new load balance loss and reward loss to ensure workload balance
and achieve a flexible trade-off between effectiveness and efficiency.
Extensive experiments demonstrate that TC-MoE achieves an average improvement of
 over 1.1% compared with traditional approaches,
while reducing the average number of activated experts by up to 9%.
These results confirm that TC-MoE effectively addresses the inefficiencies of co
nventional routing schemes,
offering a more efficient and scalable solution for MoE-based large language mod
els.
Code and models are available at https://github.com/stiger1000/TC-MoE.
****
Josiah C Kratz,Jacob Adamczyk
Reinforcement Learning for Control of Non-Markovian Cellular Population Dynamics
Many organisms and cell types, from bacteria to cancer cells, exhibit a remarkab
le ability to adapt to fluctuating environments. Additionally, cells can leverag
e memory of past environments to better survive previously-encountered stressors
. From a control perspective, this adaptability poses significant challenges in
driving cell populations toward extinction, and is thus an open question with gr
eat clinical significance. In this work, we focus on drug dosing in cell populat
ions exhibiting phenotypic plasticity. For specific dynamical models switching b
etween resistant and susceptible states, exact solutions are known. However, whe
n the underlying system parameters are unknown, and for complex memory-based sys
tems, obtaining the optimal solution is currently intractable. To address this c
hallenge, we apply reinforcement learning (RL) to identify informed dosing strat
egies to control cell populations evolving under novel non-Markovian dynamics. W
e find that model-free deep RL is able to recover exact solutions and control ce
ll populations even in the presence of long-range temporal dynamics. To further
 test our approach in more realistic settings, we demonstrate performant RL-base
d control strategies in environments with dynamic memory strength.
****
Wei Liu,Li Yang,Mingxuan Zhao,Shuxun Wang,Jin Gao,Wenjuan Li,Bing Li,Weiming Hu
DeepTAGE: Deep Temporal-Aligned Gradient Enhancement for Optimizing Spiking Neur
al Networks
Spiking Neural Networks (SNNs), with their biologically inspired spatio-temporal
 dynamics and spike-driven processing, are emerging as a promising low-power alt
ernative to traditional Artificial Neural Networks (ANNs). However, the complex
neuronal dynamics and non-differentiable spike communication mechanisms in SNNs
present substantial challenges for efficient training. By analyzing the membrane
 potentials in spiking neurons, we found that their distributions can increasing
ly deviate from the firing threshold as time progresses, which tends to cause di
minished backpropagation gradients and unbalanced optimization. To address these
 challenges, we propose Deep Temporal-Aligned Gradient Enhancement (DeepTAGE), a
 novel approach that improves optimization gradients in SNNs from both internal
surrogate gradient functions and external supervision methods. Our DeepTAGE dyna
mically adjusts surrogate gradients in accordance with the membrane potential di
stribution across different time steps, enhancing their respective gradients in

a temporal-aligned manner that promotes balanced training. Moreover, to mitigate issues of gradient vanishing or deviating during backpropagation, DeepTAGE incorporates deep supervision at both spatial (network stages) and temporal (time steps) levels to ensure more effective and robust network optimization. Importantly, our method can be seamlessly integrated into existing SNN architectures without imposing additional inference costs or requiring extra control modules. We validate the efficacy of DeepTAGE through extensive experiments on static benchmarks (CIFAR10, CIFAR100, and ImageNet-1k) and a neuromorphic dataset (DVS-CIFAR10), demonstrating significant performance improvements.
****

Mason Lary,Richard Samuelson,Alexander Wilentz,Alina Zare,Matthew Klawonn,James Fairbanks
Learning Diagrams: A Graphical Language for Compositional Training Regimes
Motivated by deep learning regimes with multiple interacting yet distinct model components, we introduce learning diagrams, graphical depictions of training setups that capture parameterized learning as data rather than code. A learning diagram compiles to a unique loss function on which component models are trained. The result of training on this loss is a collection of models whose predictions ``agree" with one another. We show that a number of popular learning setups such as few-shot multi-task learning, knowledge distillation, and multi-modal learning can be depicted as learning diagrams. We further implement learning diagrams in a library that allows users to build diagrams of PyTorch and Flux.jl models. By implementing some classic machine learning use cases, we demonstrate how learning diagrams allow practitioners to build complicated models as compositions of smaller components, identify relationships between workflows, and manipulate models during or after training. Leveraging a category theoretic framework, we introduce a rigorous semantics for learning diagrams that puts such operations on a firm mathematical foundation.
****

Jinluan Yang,Anke Tang,Didi Zhu,Zhengyu Chen,Li Shen,Fei Wu
Mitigating the Backdoor Effect for Multi-Task Model Merging via Safety-Aware Subspace
Model merging has gained significant attention as a cost-effective approach to integrate multiple single-task fine-tuned models into a unified one that can perform well on multiple tasks. However, existing model merging techniques primarily focus on resolving conflicts between task-specific models, they often overlook potential security threats, particularly the risk of backdoor attacks in the open-source model ecosystem. In this paper, we first investigate the vulnerabilities of existing model merging methods to backdoor attacks, identifying two critical challenges: backdoor succession and backdoor transfer. To address these issues, we propose a novel Defense-Aware Merging (DAM) approach that simultaneously mitigates task interference and backdoor vulnerabilities. Specifically, DAM employs a meta-learning-based optimization method with dual masks to identify a shared and safety-aware subspace for model merging. These masks are alternately optimized: the Task-Shared mask identifies common beneficial parameters across tasks, aiming to preserve task-specific knowledge while reducing interference, while the Backdoor-Detection mask isolates potentially harmful parameters to neutralize security threats. This dual-mask design allows us to carefully balance the preservation of useful knowledge and the removal of potential vulnerabilities. Compared to existing merging methods, DAM achieves a more favorable balance between performance and security, reducing the attack success rate by 2-10 percentage points while sacrificing only about 1\% in accuracy. Furthermore, DAM exhibits robust performance and broad applicability across various types of backdoor attacks and the number of compromised models involved in the merging process. Our codes and models can be accessed through https://github.com/Yangjinluan/DAM.
****

Anirudh Sundara Rajan,Utkarsh Ojha,Jedidiah Schloesser,Yong Jae Lee
Aligned Datasets Improve Detection of Latent Diffusion-Generated Images
As latent diffusion models (LDMs) democratize image generation capabilities, there is a growing need to detect fake images. A good detector should focus on the

generative model’s fingerprints while ignoring image properties such as semantic content, resolution, file format, etc. Fake image detectors are usually built in a data-driven way, where a model is trained to separate real from fake images. Existing works primarily investigate network architecture choices and training recipes. In this work, we argue that in addition to these algorithmic choices, we also require a well-aligned dataset of real/fake images to train a robust detector. For the family of LDMs, we propose a very simple way to achieve this: we reconstruct all the real images using the LDM's autoencoder, without any denoising operation. We then train a model to separate these real images from their reconstructions. The fakes created this way are extremely similar to the real ones in almost every aspect (e.g., size, aspect ratio, semantic content), which forces the model to look for the LDM decoder's artifacts. We empirically show that this way of creating aligned real/fake datasets, which also sidesteps the computationally expensive denoising process, helps in building a detector that focuses less on spurious correlations, something that a very popular existing method is susceptible to. Finally, to demonstrate the effectivenss of dataset alignment, we build a detector using images that are not natural objects, and present promising results. Overall, our work identifies the subtle but significant issues that arise when training a fake image detector and proposes a simple and inexpensive solution to address these problems.
****

Abhishek Aich,Yumin Suh,Samuel Schulter,Manmohan Chandraker
Progressive Token Length Scaling in Transformer Encoders for Efficient Universal Segmentation
A powerful architecture for universal segmentation relies on transformers that encode multi-scale image features and decode object queries into mask predictions. With efficiency being a high priority for scaling such models, we observed that the state-of-the-art method Mask2Former uses \~50% of its compute only on the transformer encoder. This is due to the retention of a full-length token-level representation of all backbone feature scales at each encoder layer. With this observation, we propose a strategy termed PROgressive Token Length SCALing for Efficient transformer encoders (PRO-SCALE) that can be plugged-in to the Mask2Former segmentation architecture to significantly reduce the computational cost. The underlying principle of PRO-SCALE is: progressively scale the length of the tokens with the layers of the encoder. This allows PRO-SCALE to reduce computations by a large margin with minimal sacrifice in performance (\~52% GFLOPs reduction with no drop in performance on COCO dataset). We validate our framework on multiple public benchmarks. Our code will be publicly released.
****

Zetian Jiang,Jiaxin Lu,Haizhao Fan,Tianzhe Wang,Junchi Yan
Learning Structured Universe Graph with Outlier OOD Detection for Partial Matching
Partial matching is a kind of graph matching where only part of two graphs can be aligned. This problem is particularly important in computer vision applications, where challenges like point occlusion or annotation errors often occur when labeling key points. Previous work has often conflated point occlusion and annotation errors, despite their distinct underlying causes. We propose two components to address these challenges: (1) a structured universe graph is learned to connect two input graphs $X_{ij} = X_{iu} X_{ju}^\top$, effectively resolving the issue of point occlusion; (2) an energy-based out-of-distribution detection is designed to remove annotation errors from the input graphs before matching. We evaluated our method on the Pascal VOC and Willow Object datasets, focusing on scenarios involving point occlusion and random outliers. The experimental results demonstrate that our approach consistently outperforms state-of-the-art methods across all tested scenarios, highlighting the accuracy and robustness of our method.
****

Li Nanbo,Firas Laakom,Yucheng XU,Wenyi Wang,Jürgen Schmidhuber
FACTS: A Factored State-Space Framework for World Modelling
World modelling is essential for understanding and predicting the dynamics of co

mplex systems by learning both spatial and temporal dependencies. However, curre
nt frameworks, such as Transformers and selective state-space models like Mambas
, exhibit limitations in efficiently encoding spatial and temporal structures, p
articularly in scenarios requiring long-term high-dimensional sequence modelling
. To address these issues, we propose a novel recurrent framework, the FACTored
State-space (FACTS) model, for spatial-temporal world modelling. The FACTS frame
work constructs a graph-structured memory with a routing mechanism that learns p
ermutable memory representations, ensuring invariance to input permutations whil
e adapting through selective state-space propagation. Furthermore, FACTS support
s parallel computation of high-dimensional sequences. We empirically evaluate FA
CTS across diverse tasks, including multivariate time series forecasting, object
-centric world modelling, and spatial-temporal graph prediction, demonstrating t
hat it consistently outperforms or matches specialised state-of-the-art models,
despite its general-purpose world modelling design.
****
Changyu Chen,Zichen Liu,Chao Du,Tianyu Pang,Qian Liu,Arunesh Sinha,Pradeep Varak
antham,Min Lin
Bootstrapping Language Models with DPO Implicit Rewards
Human alignment in large language models (LLMs) is an active area of research. A
 recent groundbreaking work, direct preference optimization (DPO), has greatly s
implified the process from past work in reinforcement learning from human feedba
ck (RLHF) by bypassing the reward learning stage in RLHF. DPO, after training, p
rovides an implicit reward model. In this work, we make a novel observation that
 this implicit reward model can by itself be used in a bootstrapping fashion to
further align the LLM. Our approach is to use the rewards from a current LLM to
construct a preference dataset, which is then used in subsequent DPO rounds. We
incorporate two refinements to further improve our approach: 1) length-regulariz
ed reward shaping to make the preference dataset length-unbiased; 2) experience
replay to enhance the quality of the preference dataset. Our approach, named sel
f-alignment with DPO ImpliCit rEwards (DICE), shows great improvements in alignm
ent. It achieves an increase of more than 8$\\%$ in lengthcontrolled win rate on
 AlpacaEval 2 for all the different base models that we tried, without relying o
n external feedback. Our code is available at https://github.com/sail-sg/dice.
****
Liuyue Xie,Joel Julin,Koichiro Niinuma,Laszlo Attila Jeni
Gaussian Splatting Lucas-Kanade
Gaussian Splatting and its dynamic extensions are effective for reconstructing 3
D scenes from 2D images when there is significant camera movement to facilitate
motion parallax and when scene objects remain relatively static. However, in man
y real-world scenarios, these conditions are not met. As a consequence, data-dri
ven semantic and geometric priors have been favored as regularizers, despite the
ir bias toward training data and their neglect of broader movement dynamics.

Departing from this practice, we propose a novel analytical approach that adapts
 the classical Lucas-Kanade method to dynamic Gaussian splatting. By leveraging
the intrinsic properties of the forward warp field network, we derive an analyti
cal velocity field that, through time integration, facilitates accurate scene fl
ow computation. This enables the precise enforcement of motion constraints on wa
rp fields, thus constraining both 2D motion and 3D positions of the Gaussians. O
ur method excels in reconstructing highly dynamic scenes with minimal camera mov
ement, as demonstrated through experiments on both synthetic and real-world scen
es.
****
Shihao Zhang,Yuguang Yan,Angela Yao
Improving Deep Regression with Tightness
For deep regression, preserving the ordinality of the targets with respect to th
e feature representation improves performance across various tasks. However, a t
heoretical explanation for the benefits of ordinality is still lacking. This wor
k reveals that preserving ordinality reduces the conditional entropy $H(Z|Y)$ of
 representation Z conditional on the target Y. However, our findings reveal

that typical regression losses do little to reduce $H(Z|Y)$, even though it is v
ital for generalization performance. With this motivation, we introduce an opti
mal transport-based regularizer to preserve the similarity relationships of targ
ets in the feature space to reduce $H(Z|Y)$. Additionally, we introduce a simple
 yet efficient strategy of duplicating the regressor targets, also with the aim
of reducing $H(Z|Y)$. Experiments on three real-world regression tasks verify t
he effectiveness of our strategies to improve deep regression. Code: https://gi
thub.com/needylove/Regression_tightness
****

Ruochen Jin,Bojian Hou,Jiancong Xiao,Weijie J Su,Li Shen
Fine-Tuning Attention Modules Only: Enhancing Weight Disentanglement in Task Ari
thmetic
In recent years, *task arithmetic* has garnered increasing attention. This appro
ach edits pre-trained models directly in weight space by combining the fine-tune
d weights of various tasks into a *unified model*. Its efficiency and cost-effec
tiveness stem from its training-free combination, contrasting with traditional m
ethods that require model training on large datasets for multiple tasks. However
, applying such a unified model to individual tasks can lead to interference fro
m other tasks (lack of *weight disentanglement*). To address this issue, Neural
Tangent Kernel (NTK) linearization has been employed to leverage a ''kernel beha
vior'', facilitating weight disentanglement and mitigating adverse effects from
unrelated tasks. Despite its benefits, NTK linearization presents drawbacks, inc
luding doubled training costs, as well as reduced performance of individual mode
ls. To tackle this problem, we propose a simple yet effective and efficient meth
od that is to finetune the attention modules only in the Transformer. Our study
reveals that the attention modules exhibit kernel behavior, and fine-tuning the
attention modules only significantly improves weight disentanglement. To further
 understand how our method improves the weight disentanglement of task arithmeti
c, we present a comprehensive study of task arithmetic by differentiating the ro
le of the representation module and task-specific module. In particular, we find
 that the representation module plays an important role in improving weight dise
ntanglement whereas the task-specific modules such as the classification heads c
an degenerate the weight disentanglement performance.
****

Nikunj Saunshi,Nishanth Dikkala,Zhiyuan Li,Sanjiv Kumar,Sashank J. Reddi
Reasoning with Latent Thoughts: On the Power of Looped Transformers
Large language models have shown remarkable reasoning abilities and scaling laws
 suggest that large parameter count, especially along the depth axis, is the pri
mary driver. In this work, we make a stronger claim --- many reasoning problems
require a large depth but not necessarily many parameters. This unlocks a novel
application of looped models for reasoning. Firstly, we show that for many synth
etic reasoning problems like addition, p-hop induction, and math problems, a $
k$-layer transformer looped L times nearly matches the performance of a kL-l
ayer non-looped model, and is significantly better than a k-layer model. This
is further corroborated by theoretical results showing that many such reasoning
problems can be solved via iterative algorithms, and thus, can be solved effecti
vely using looped models with nearly optimal depth. Perhaps surprisingly, these
benefits also translate to practical settings of language modeling --- on many d
ownstream reasoning tasks, a language model with k-layers looped L times can
 be competitive to, if not better than, a kL-layer language model. In fact, ou
r empirical analysis reveals an intriguing phenomenon: looped and non-looped mod
els exhibit scaling behavior that depends on their effective depth, akin to the
inference-time scaling of chain-of-thought (CoT) reasoning. We further elucidate
 the connection to CoT reasoning by proving that looped models implicitly genera
te latent thoughts and can simulate T steps of CoT with T loops. Inspired by
 these findings, we also present an interesting dichotomy between reasoning and
memorization, and design a looping-based regularization that is effective on bot
h fronts.
****

Yuxian Gu,Li Dong,Hongning Wang,Yaru Hao,Qingxiu Dong,Furu Wei,Minlie Huang

Data Selection via Optimal Control for Language Models
This work investigates the selection of high-quality pre-training data from mass ive corpora to enhance LMs' capabilities for downstream usage.
We formulate data selection as a generalized Optimal Control problem, which can be solved theoretically by Pontryagin's Maximum Principle (PMP), yielding a set of necessary conditions that characterize the relationship between optimal data selection and LM training dynamics.
Based on these theoretical results, we introduce **P**MP-based **D**ata **S**ele ction (**PDS**), a framework that approximates optimal data selection by solving the PMP conditions.
In our experiments, we adopt PDS to select data from CommmonCrawl and show that the PDS-selected corpus accelerates the learning of LMs and constantly boosts th eir performance on a wide range of downstream tasks across various model sizes. Moreover, the benefits of PDS extend to ~400B models trained on ~10T tokens, as evidenced by the extrapolation of the test loss curves according to the Scaling Laws.
PDS also improves data utilization when the pre-training data is limited, by red ucing the data demand by 1.8 times, which helps mitigate the quick exhaustion of available web-crawled corpora. Our code, model, and data can be found at https: //github.com/microsoft/LMOps/tree/main/data_selection.
****
Stanley Wei,Sadhika Malladi,Sanjeev Arora,Amartya Sanyal
Provable unlearning in topic modeling and downstream tasks
Machine unlearning algorithms are increasingly important as legal concerns arise around the provenance of training data, but verifying the success of unlearning is often difficult. Provable guarantees for unlearning are often limited to sup ervised learning settings. In this paper, we provide the first theoretical guara ntees for unlearning in the pre-training and fine-tuning paradigm by studying to pic models, simple bag-of-words language models that can be adapted to solve dow nstream tasks like retrieval and classification. First, we design a provably eff ective unlearning algorithm for topic models that incurs a computational overhea d independent of the size of the original dataset. Our analysis additionally qua ntifies the deletion capacity of the model -- i.e., the number of examples that can be unlearned without incurring a significant cost in model performance. Fina lly, we formally extend our analyses to account for adaptation to a given downst ream task. In particular, we design an efficient algorithm to perform unlearning after fine-tuning the topic model via a linear head. Notably, we show that it i s easier to unlearn pre-training data from models that have been fine-tuned to a particular task, and one can unlearn this data without modifying the base model .
****
Qi Sun,Edoardo Cetin,Yujin Tang
Transformer-Squared: Self-adaptive LLMs
Self-adaptive large language models (LLMs) aim to solve the challenges posed by traditional fine-tuning methods, which are often computationally intensive and s tatic in their ability to handle diverse tasks. We introduce Transformer-Squared , a novel self-adaptation framework that adapts LLMs for unseen tasks in real-ti me by selectively adjusting only the singular components of their weight matrice s. During inference, Transformer-Squared employs a two-pass mechanism: first, a dispatch system identifies the task properties, and then task-specific 'expert' vectors, trained using reinforcement learning, are dynamically mixed to obtain t argeted behavior for the incoming prompt. Our method consistently outperforms ub iquitous approaches such as LoRA, with fewer parameters and greater efficiency. Furthermore, Transformer-Squared demonstrates versatility across different LLM a rchitectures and modalities, including vision-language tasks. Transformer-Square d represents a significant leap forward, offering a scalable, efficient solution for enhancing the adaptability and task-specific performance of LLMs, paving th e way for truly dynamic, self-organizing AI systems.
****
Toshiaki Koike-Akino,Francesco Tonin,Yongtao Wu,Zhengqing Wu,Leyla Naz Candogan,

Volkan Cevher
Quantum-PEFT: Ultra parameter-efficient fine-tuning
This paper introduces Quantum-PEFT that leverages quantum computations for parameter-efficient fine-tuning (PEFT). Unlike other additive PEFT methods, such as low-rank adaptation (LoRA), Quantum-PEFT exploits an underlying full-rank yet surprisingly parameter efficient _quantum unitary parameterization_. With the use of Pauli parameterization, the number of trainable parameters grows only logarithmically with the ambient dimension, as opposed to linearly as in LoRA-based PEFT methods. Quantum-PEFT achieves vanishingly smaller number of trainable parameters than the lowest-rank LoRA as dimensions grow, enhancing parameter efficiency while maintaining a competitive performance. We apply Quantum-PEFT to several transfer learning benchmarks in language and vision, demonstrating significant advantages in parameter efficiency.
****
Chenyang Cao,Yucheng Xin,Silang Wu,Longxiang He,Zichen Yan,Junbo Tan,Xueqian Wang
FOSP: Fine-tuning Offline Safe Policy through World Models
Offline Safe Reinforcement Learning (RL) seeks to address safety constraints by learning from static datasets and restricting exploration. However, these approaches heavily rely on the dataset and struggle to generalize to unseen scenarios safely. In this paper, we aim to improve safety during the deployment of vision-based robotic tasks through online fine-tuning an offline pretrained policy. To facilitate effective fine-tuning, we introduce model-based RL, which is known for its data efficiency. Specifically, our method employs in-sample optimization to improve offline training efficiency while incorporating reachability guidance to ensure safety. After obtaining an offline safe policy, a safe policy expansion approach is leveraged for online fine-tuning. The performance of our method is validated on simulation benchmarks with five vision-only tasks and through real-world robot deployment using limited data. It demonstrates that our approach significantly improves the generalization of offline policies to unseen safety-constrained scenarios. To the best of our knowledge, this is the first work to explore offline-to-online RL for safe generalization tasks. The videos are available at https://sunlighted.github.io/fosp_web/.
****
Boqian Wu,Qiao Xiao,Shunxin Wang,Nicola Strisciuglio,Mykola Pechenizkiy,Maurice van Keulen,Decebal Constantin Mocanu,Elena Mocanu
Dynamic Sparse Training versus Dense Training: The Unexpected Winner in Image Corruption Robustness
It is generally perceived that Dynamic Sparse Training opens the door to a new era of scalability and efficiency for artificial neural networks at, perhaps, some costs in accuracy performance for the classification task. At the same time, Dense Training is widely accepted as being the "de facto" approach to train artificial neural networks if one would like to maximize their robustness against image corruption. In this paper, we question this general practice. Consequently, \textit{we claim that}, contrary to what is commonly thought, the Dynamic Sparse Training methods can consistently outperform Dense Training in terms of robustness accuracy, particularly if the efficiency aspect is not considered as a main objective (i.e., sparsity levels between 10\% and up to 50\%), without adding (or even reducing) resource cost. We validate our claim on two types of data, images and videos, using several traditional and modern deep learning architectures for computer vision and three widely studied Dynamic Sparse Training algorithms. Our findings reveal a new yet-unknown benefit of Dynamic Sparse Training and open new possibilities in improving deep learning robustness beyond the current state of the art.
****
Yicheng Deng,Hideaki Hayashi,Hajime Nagahara
Gaussian-Based Instance-Adaptive Intensity Modeling for Point-Supervised Facial Expression Spotting
Point-supervised facial expression spotting (P-FES) aims to localize facial expression instances in untrimmed videos, requiring only a single timestamp label fo

r each instance during training. To address label sparsity, hard pseudo-labeling is often employed to propagate point labels to unlabeled frames; however, this approach can lead to confusion when distinguishing between neutral and expressio n frames with various intensities, which can negatively impact model performance . In this paper, we propose a two-branch framework for P-FES that incorporates a Gaussian-based instance-adaptive Intensity Modeling (GIM) module for soft pseud o-labeling. GIM models the expression intensity distribution for each instance. Specifically, we detect the pseudo-apex frame around each point label, estimate the duration, and construct a Gaussian distribution for each expression instance . We then assign soft pseudo-labels to pseudo-expression frames as intensity val ues based on the Gaussian distribution. Additionally, we introduce an Intensity-Aware Contrastive (IAC) loss to enhance discriminative feature learning and supp ress neutral noise by contrasting neutral frames with expression frames of vario us intensities. Extensive experiments on the SAMM-LV and CAS(ME)^2 datasets de monstrate the effectiveness of our proposed framework. Code is available at http s://github.com/KinopioIsAllIn/GIM.
****
Yu Wang,Ruihan Wu,Zexue He,Xiusi Chen,Julian McAuley
Large Scale Knowledge Washing
Large language models show impressive abilities in memorizing world knowledge, w hich leads to concerns regarding memorization of private information, toxic or s ensitive knowledge, and copyrighted content. We introduce the problem of Large S cale Knowledge Washing, focusing on unlearning an extensive amount of factual kn owledge. Previous unlearning methods usually define the reverse loss and update the model via backpropagation, which may affect the model's fluency and reasonin g ability or even destroy the model due to extensive training with the reverse l oss. Existing works introduce additional data from downstream tasks to prevent t he model from losing capabilities, which requires downstream task awareness. Con trolling the tradeoff of unlearning existing knowledge while maintaining existin g capabilities is also challenging. To this end, we propose LaW (Large Scale Was hing), where we update the MLP layers in decoder-only large language models to p erform knowledge washing, as inspired by model editing methods. We derive a new objective with the knowledge to be unlearned to update the weights of certain ML P layers. Experimental results demonstrate the effectiveness of LaW in forgettin g target knowledge while maximally maintaining reasoning ability. The code will be open-sourced.
****
Riyaz Ahuja,Jeremy Avigad,Prasad Tetali,Sean Welleck
ImProver: Agent-Based Automated Proof Optimization
Large language models (LLMs) have been used to generate formal proofs of mathema tical theorems in proofs assistants such as Lean.
However, we often want to optimize a formal proof with respect to various criter ia, depending on its downstream use.
For example, we may want a proof to adhere to a certain style, be declaratively structured, or concise. Having suitably optimized proofs is also important for l earning tasks, especially since human-written proofs may not optimal for that pu rpose.
To this end, we study a new problem of automated proof optimization: rewriting a proof so that it is correct and optimizes for an arbitrary criterion, such as l ength or declarativity.
As a first method for automated proof optimization, we present ImProver, a large -language-model agent that rewrites proofs to optimize arbitrary user-defined me trics in Lean.
We find that naively applying LLMs to proof optimization falls short, and we inc orporate various improvements into ImProver, such as the use of symbolic Lean co ntext in a novel Chain-of-States technique, as well as error-correction and retr ieval. We test ImProver on rewriting real-world undergraduate, competition, and research-level mathematics theorems, finding that ImProver is capable of rewriti ng proofs so that they are substantially shorter and more declarative in structu re.

****
Song Duong,Florian Le Bronnec,Alexandre Allauzen,Vincent Guigue,Alberto Lumbrera s,Laure Soulier,Patrick Gallinari
SCOPE: A Self-supervised Framework for Improving Faithfulness in Conditional Tex t Generation
Large Language Models (LLMs), when used for conditional text generation, often p roduce hallucinations, i.e., information that is unfaithful or not grounded in t he input context. This issue arises in typical conditional text generation tasks , such as text summarization and data-to-text generation, where the goal is to p roduce fluent text based on contextual input. When fine-tuned on specific domain s, LLMs struggle to provide faithful answers to a given context, often adding in formation or generating errors. One underlying cause of this issue is that LLMs rely on statistical patterns learned from their training data. This reliance can interfere with the model's ability to stay faithful to a provided context, lead ing to the generation of ungrounded information. We build upon this observation and introduce a novel self-supervised method for generating a training set of un faithful samples. We then refine the model using a training process that encoura ges the generation of grounded outputs over unfaithful ones, drawing on preferen ce-based training. Our approach leads to significantly more grounded text genera tion, outperforming existing self-supervised techniques in faithfulness, as eval uated through automatic metrics, LLM-based assessments, and human evaluations.
****
Carolin Schmidt,Daniele Gammelli,James Harrison,Marco Pavone,Filipe Rodrigues
Offline Hierarchical Reinforcement Learning via Inverse Optimization
Hierarchical policies enable strong performance in many sequential decision-maki ng problems, such as those with high-dimensional action spaces, those requiring long-horizon planning, and settings with sparse rewards.
However, learning hierarchical policies from static offline datasets presents a significant challenge.
Crucially, actions taken by higher-level policies may not be directly observable within hierarchical controllers, and the offline dataset might have been genera ted using a different policy structure, hindering the use of standard offline le arning algorithms.
In this work, we propose $\textit{OHIO}$: a framework for offline reinforcement learning (RL) of hierarchical policies.
Our framework leverages knowledge of the policy structure to solve the $\textit{ inverse problem}$, recovering the unobservable high-level actions that likely ge nerated the observed data under our hierarchical policy.
This approach constructs a dataset suitable for off-the-shelf offline training.
We demonstrate our framework on robotic and network optimization problems and sh ow that it substantially outperforms end-to-end RL methods and improves robustne ss. We investigate a variety of instantiations of our framework, both in direct deployment of policies trained offline and when online fine-tuning is performed . Code and data are available at https://ohio-offline-hierarchical-rl.github.io.
****
Qitai Wang,Lue Fan,Yuqi Wang,Yuntao Chen,Zhaoxiang Zhang
FreeVS: Generative View Synthesis on Free Driving Trajectory
Existing reconstruction-based novel view synthesis methods for driving scenes fo cus on synthesizing camera views along the recorded trajectory of the ego vehicl e.
Their image rendering performance will severely degrade on viewpoints falling ou t of the recorded trajectory, where camera rays are untrained.
We propose FreeVS, a novel fully generative approach that can synthesize camera views on free new trajectories in real driving scenes.
To control the generation results to be 3D consistent with the real scenes and a ccurate in viewpoint pose, we propose the pseudo-image representation of view pr iors to control the generation process.
Viewpoint translation simulation is applied on pseudo-images to simulate camera movement in each direction.
Once trained, FreeVS can be applied to any validation sequences without reconstr

uction process and synthesis views on novel trajectories.
Moreover, we propose two new challenging benchmarks tailored to driving scenes, which are novel camera synthesis and novel trajectory synthesis, emphasizing the freedom of viewpoints.
Given that no ground truth images are available on novel trajectories, we also propose to evaluate the consistency of images synthesized on novel trajectories with 3D perception models.
Experiments on the Waymo Open Dataset show that FreeVS has a strong image synthesis performance on both the recorded trajectories and novel trajectories.
The code is released. Project page: https://freevs24.github.io/.
****

Jeffrey Willette,Heejun Lee,Youngwan Lee,Myeongjae Jeon,Sung Ju Hwang
Training Free Exponential Context Extension via Cascading KV Cache
The transformer's context window is vital for tasks such as few-shot learning and conditional generation as it preserves previous tokens for active memory. However, as the context lengths increase, the computational costs grow quadratically, hindering the deployment of large language models (LLMs) in real-world, long sequence scenarios. Although some recent key-value caching (KV Cache) methods offer linear inference complexity, they naively manage the stored context, prematurely evicting tokens and losing valuable information. Moreover, they lack an optimized prefill/prompt stage strategy, resulting in higher latency than even quadratic attention for realistic context sizes. In response, we introduce a novel mechanism that leverages cascading sub-cache buffers to selectively retain the most relevant tokens, enabling the model to maintain longer context histories without increasing the cache size. Our approach outperforms linear caching baselines across key benchmarks, including streaming perplexity, question answering, book summarization, and passkey retrieval, where it retains better retrieval accuracy at 1M tokens after four doublings of the cache size of 65K. Additionally, our method reduces prefill stage latency by a factor of 6.8 when compared to flash attention on 1M tokens. These innovations not only enhance the computational efficiency of LLMs but also pave the way for their effective deployment in resource-constrained environments, enabling large-scale, real-time applications with significantly reduced latency.
****

Yanming Liu,Xinyue Peng,Jiannan Cao,Shi Bo,Yuwei Zhang,Xuhong Zhang,Sheng Cheng,Xun Wang,Jianwei Yin,Tianyu Du
Tool-Planner: Task Planning with Clusters across Multiple Tools
Large language models (LLMs) have demonstrated exceptional reasoning capabilities, enabling them to solve various complex problems. Recently, this ability has been applied to the paradigm of tool learning. Tool learning involves providing examples of tool usage and their corresponding functions, allowing LLMs to formulate plans and demonstrate the process of invoking and executing each tool. LLMs can address tasks that they cannot complete independently, thereby enhancing their potential across different tasks. However, this approach faces two key challenges. First, redundant error correction leads to unstable planning and long execution time. Additionally, designing a correct plan among multiple tools is also a challenge in tool learning. To address these issues, we propose Tool-Planner, a task-processing framework based on toolkits. Tool-Planner groups tools based on the API functions with the same function into a toolkit and allows LLMs to implement planning across the various toolkits. When a tool error occurs, the language model can reselect and adjust tools based on the toolkit. Experiments show that our approach demonstrates a high pass and win rate across different datasets and optimizes the planning scheme for tool learning in models such as GPT-4 and Claude 3, showcasing the potential of our method. Our code is public at https://github.com/OceannTwT/Tool-Planner.
****

Lukas Rauch,Raphael Schwinger,Moritz Wirth,René Heinrich,Denis Huseljic,Marek Herde,Jonas Lange,Stefan Kahl,Bernhard Sick,Sven Tomforde,Christoph Scholz
BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics
Deep learning (DL) has greatly advanced audio classification, yet the field is l

imited by the scarcity of large-scale benchmark datasets that have propelled pro gress in other domains. While AudioSet is a pivotal step to bridge this gap as a universal-domain dataset, its restricted accessibility and limited range of eva luation use cases challenge its role as the sole resource. Therefore, we introdu ce BirdSet, a large-scale benchmark data set for audio classification focusing o n avian bioacoustics. BirdSet surpasses AudioSet with over 6,800 recording hours ($\uparrow17\%$) from nearly 10,000 classes ($\uparrow18\times$) for training a nd more than 400 hours ($\uparrow7\times$) across eight strongly labeled evaluat ion datasets. It serves as a versatile resource for use cases such as multi-labe l classification, covariate shift or self-supervised learning. We benchmark six well-known DL models in multi-label classification across three distinct trainin g scenarios and outline further evaluation use cases in audio classification. We host our dataset on Hugging Face for easy accessibility and offer an extensive codebase to reproduce our results.
****
Jinnan Chen,Chen Li,Jianfeng Zhang,Lingting Zhu,Buzhen Huang,Hanlin Chen,Gim Hee Lee
Generalizable Human Gaussians from Single-View Image
In this work, we tackle the task of learning 3D human Gaussians from a single im age, focusing on recovering detailed appearance and geometry including unobserve d regions. We introduce a single-view generalizable Human Gaussian Model (HGM), which employs a novel generate-then-refine pipeline with the guidance from human body prior and diffusion prior. Our approach uses a ControlNet to refine render ed back-view images from coarse predicted human Gaussians, then uses the refined image along with the input image to reconstruct refined human Gaussians. To mit igate the potential generation of unrealistic human poses and shapes, we incorpo rate human priors from the SMPL-X model as a dual branch, propagating image feat ures from the SMPL-X volume to the image Gaussians using sparse convolution and attention mechanisms. Given that the initial SMPL-X estimation might be inaccura te, we gradually refine it with our HGM model. We validate our approach on sever al publicly available datasets. Our method surpasses previous methods in both no vel view synthesis and surface reconstruction. Our approach also exhibits strong generalization for cross-dataset evaluation and in-the-wild images.
****
Minhyuk Seo,Hyunseo Koh,Jonghyun Choi
Budgeted Online Continual Learning by Adaptive Layer Freezing and Frequency-base d Sampling
The majority of online continual learning (CL) advocates single-epoch training a nd imposes restrictions on the size of replay memory. However, single-epoch trai ning would incur a different amount of computations per CL algorithm, and the ad ditional storage cost to store logit or model in addition to replay memory is la rgely ignored in calculating the storage budget. Arguing different computational and storage budgets hinder fair comparison among CL algorithms in practice, we propose to use floating point operations (FLOPs) and total memory size in Byte a s a metric for computational and memory budgets, respectively, to compare and de velop CL algorithms in the same ‘total resource budget.’ To improve a CL method in a limited total budget, we propose adaptive layer freezing that does not upda te the layers for less informative batches to reduce computational costs with a negligible loss of accuracy. In addition, we propose a memory retrieval method t hat allows the model to learn the same amount of knowledge as using random retri eval in fewer iterations. Empirical validations on the CIFAR-10/100, CLEAR-10/10 0, and ImageNet-1K datasets demonstrate that the proposed approach outperforms t he state-of-the-art methods within the same total budget. Furthermore, we valida te its effectiveness in the Multi-modal Concept incremental Learning setup with multimodal large language models, such as LLaVA-1.5-7B. Code is available at htt ps://github.com/snumprlab/budgeted-cl.
****
Peiqi Wang,Barbara D. Lam,Yingcheng Liu,Ameneh Asgari-Targhi,Rameswar Panda,Will iam M Wells,Tina Kapur,Polina Golland
Calibrating Expressions of Certainty

We present a novel approach to calibrating linguistic expressions of certainty, e.g., "Maybe" and "Likely". Unlike prior work that assigns a single score to each certainty phrase, we model uncertainty as distributions over the simplex to capture their semantics more accurately. To accommodate this new representation of certainty, we generalize existing measures of miscalibration and introduce a novel post-hoc calibration method. Leveraging these tools, we analyze the calibration of both humans (e.g., radiologists) and computational models (e.g., language models) and provide interpretable suggestions to improve their calibration.
****

Nayoung Kim,Seongsu Kim,Minsu Kim,Jinkyoo Park,Sungsoo Ahn
MOFFlow: Flow Matching for Structure Prediction of Metal-Organic Frameworks
Metal-organic frameworks (MOFs) are a class of crystalline materials with promising applications in many areas such as carbon capture and drug delivery. In this work, we introduce MOFFlow, the first deep generative model tailored for MOF structure prediction. Existing approaches, including ab initio calculations and even deep generative models, struggle with the complexity of MOF structures due to the large number of atoms in the unit cells. To address this limitation, we propose a novel Riemannian flow matching framework that reduces the dimensionality of the problem by treating the metal nodes and organic linkers as rigid bodies, capitalizing on the inherent modularity of MOFs. By operating in the $SE(3)$ space, MOFFlow effectively captures the roto-translational dynamics of these rigid components in a scalable way. Our experiment demonstrates that MOFFlow accurately predicts MOF structures containing several hundred atoms, significantly outperforming conventional methods and state-of-the-art machine learning baselines while being much faster. Code available at https://github.com/nayoung10/MOFFlow.
****

Junhua Chen,Lorenz Richter,Julius Berner,Denis Blessing,Gerhard Neumann,Anima Anandkumar
Sequential Controlled Langevin Diffusions
An effective approach for sampling from unnormalized densities is based on the idea of gradually transporting samples from an easy prior to the complicated target distribution. Two popular methods are (1) Sequential Monte Carlo (SMC), where the transport is performed through successive annealed densities via prescribed Markov chains and resampling steps, and (2) recently developed diffusion-based sampling methods, where a learned dynamical transport is used. Despite the common goal, both approaches have different, often complementary, advantages and drawbacks. The resampling steps in SMC allow focusing on promising regions of the space, often leading to robust performance. While the algorithm enjoys asymptotic guarantees, the lack of flexible, learnable transitions can lead to slow convergence. On the other hand, diffusion-based samplers are learned and can potentially better adapt themselves to the target at hand, yet often suffer from training instabilities. In this work, we present a principled framework for combining SMC with diffusion-based samplers by viewing both methods in continuous time and considering measures on path space. This culminates in the new Sequential Controlled Langevin Diffusion (SCLD) sampling method, which is able to utilize the benefits of both methods and reaches improved performance on multiple benchmark problems, in many cases using only 10% of the training budget of previous diffusion-based samplers.
****

Anvith Thudi,Chris J. Maddison
MixMax: Distributional Robustness in Function Space via Optimal Data Mixtures
Machine learning models are often required to perform well across several pre-defined settings, such as a set of user groups. Worst-case performance is a common metric to capture this requirement, and is the objective of group distributionally robust optimization (group DRO). Unfortunately, these methods struggle when the loss is non-convex in the parameters, or the model class is non-parametric. Here, we make a classical move to address this: we reparameterize group DRO from parameter space to function space, which results in a number of advantages. First, we show that group DRO over the space of bounded functions admits a minimax theorem. Second, for cross-entropy and mean squared error, we show that the mini

max optimal mixture distribution is the solution of a simple convex optimization problem. Thus, provided one is working with a model class of universal function approximators, group DRO can be solved by a convex optimization problem followed by a classical risk minimization problem. We call our method MixMax. In our experiments, we found that MixMax matched or outperformed the standard group DRO baselines, and in particular, MixMax improved the performance of XGBoost over the only baseline, data balancing, for variations of the ACSIncome and CelebA annotations datasets.
****
Seungjoo Shin,Jaesik Park,Sunghyun Cho
Locality-aware Gaussian Compression for Fast and High-quality Rendering
We present LocoGS, a locality-aware 3D Gaussian Splatting (3DGS) framework that exploits the spatial coherence of 3D Gaussians for compact modeling of volumetric scenes.
To this end, we first analyze the local coherence of 3D Gaussian attributes, and propose a novel locality-aware 3D Gaussian representation that effectively encodes locally-coherent Gaussian attributes using a neural field representation with a minimal storage requirement.
On top of the novel representation, LocoGS is carefully designed with additional components such as dense initialization, an adaptive spherical harmonics bandwidth scheme and different encoding schemes for different Gaussian attributes to maximize compression performance.
Experimental results demonstrate that our approach outperforms the rendering quality of existing compact Gaussian representations for representative real-world 3D datasets while achieving from 54.6$\times$ to 96.6$\times$ compressed storage size and from 2.1$\times$ to 2.4$\times$ rendering speed than 3DGS. Even our approach also demonstrates an averaged 2.4$\times$ higher rendering speed than the state-of-the-art compression method with comparable compression performance.
****
James Liu,Pragaash Ponnusamy,Tianle Cai,Han Guo,Yoon Kim,Ben Athiwaratkun
Training-Free Activation Sparsity in Large Language Models
Activation sparsity can enable practical inference speedups in large language models (LLMs) by reducing the compute and memory-movement required for matrix multiplications during the forward pass.
However, existing methods face limitations that inhibit widespread adoption. Some approaches are tailored towards older models with ReLU-based sparsity, while others require extensive continued pre-training on up to hundreds of billions of tokens.
This paper describes TEAL (**T**raining-Fre**e** **A**ctivation Sparsity in **L**LMs), a simple training-free method that applies magnitude-based activation sparsity to hidden states throughout the entire model. TEAL achieves 40-50\% model-wide sparsity with minimal performance degradation across Llama-2, Llama-3, and Mistral families, with sizes varying from 7B to 70B. We improve existing sparse kernels and demonstrate wall-clock decoding speed-ups of up to 1.53× and 1.8× at 40\% and 50\% model-wide sparsity. TEAL is compatible with weight quantization, enabling further efficiency gains.
****
Alan Baade,Puyuan Peng,David Harwath
SyllableLM: Learning Coarse Semantic Units for Speech Language Models
Language models require tokenized inputs. However, tokenization strategies for continuous data like audio and vision are often based on simple heuristics such as fixed sized convolutions or discrete clustering, which do not necessarily align with the semantic structure of the data. For speech in particular, the high resolution of waveforms (16,000 samples/second or more) presents a significant challenge as speech-based language models have had to use several times more tokens per word than text-based language models. In this work, we introduce a controllable self-supervised technique to merge speech representations into coarser syllable-like units while still preserving semantic information. We do this by 1) extracting noisy boundaries through analyzing correlations in pretrained encoder losses and 2) iteratively improving model representations with a novel distillati

on technique. Our method produces controllable-rate semantic units at as low as 5Hz and 60bps and achieves SotA in syllabic segmentation and clustering. Using these coarse tokens, we successfully train SyllableLM, a Speech Language Model (SpeechLM) that matches or outperforms current SotA SpeechLMs on a range of spoken language modeling tasks. SyllableLM also achieves significant improvements in efficiency with a 30x reduction in training compute and a 4x wall-clock inference speedup. Our code and checkpoints are available at https://www.github.com/alanbaade/SyllableLM
****
Blake Bordelon,Alexander Atanasov,Cengiz Pehlevan
How Feature Learning Can Improve Neural Scaling Laws
We develop a simple solvable model of neural scaling laws beyond the kernel limit. Theoretical analysis of this model predicts the performance scaling predictions with model size, training time and total amount of available data. From the scaling analysis we identify three relevant regimes: hard tasks, easy tasks, and super easy tasks. For easy and super-easy target functions, which are in the Hilbert space (RKHS) of the initial infinite-width neural tangent kernel (NTK), there is no change in the scaling exponents between feature learning models and models in the kernel regime. For hard tasks, which we define as tasks outside of the RKHS of the initial NTK, we show analytically and empirically that feature learning can improve the scaling with training time and compute, approximately doubling the exponent for very hard tasks. This leads to a new compute optimal scaling law for hard tasks in the feature learning regime. We support our finding that feature learning improves the scaling law for hard tasks with experiments of nonlinear MLPs fitting functions with power-law Fourier spectra on the circle and CNNs learning vision tasks.
****
Alex N Wang,Christopher Hoang,Yuwen Xiong,Yann LeCun,Mengye Ren
PooDLe■: Pooled and dense self-supervised learning from naturalistic videos
Self-supervised learning has driven significant progress in learning from single-subject, _iconic_ images.
However, there are still unanswered questions about the use of minimally-curated, naturalistic video data, which contain _dense_ scenes with many independent objects, imbalanced class distributions, and varying object sizes.
In this paper, we propose PooDLe, a self-supervised learning method that combines an invariance-based objective on pooled representations with a dense SSL objective that enforces equivariance to optical flow warping.
Our results show that a unified objective applied at multiple feature scales is essential for learning effective image representations from naturalistic videos.
We validate our method with experiments on the BDD100K driving video dataset and the Walking Tours first-person video dataset, demonstrating its ability to capture spatial understanding from a dense objective and semantic understanding via a pooled representation objective.
****
Chenze Shao,Fandong Meng,Jie Zhou
Beyond Next Token Prediction: Patch-Level Training for Large Language Models
The prohibitive training costs of Large Language Models (LLMs) have emerged as a significant bottleneck in the development of next-generation LLMs. In this paper, we show that it is possible to significantly reduce the training costs of LLMs without sacrificing their performance. Specifically, we introduce patch-level training for LLMs, in which multiple tokens are aggregated into a unit of higher information density, referred to as a `patch', to serve as the fundamental text unit for training LLMs. During patch-level training, we feed the language model shorter sequences of patches and train it to predict the next patch, thereby processing the majority of the training data at a significantly reduced cost. Following this, the model continues token-level training on the remaining training data to align with the inference mode. Experiments on a diverse range of models (370M-2.7B parameters) demonstrate that patch-level training can reduce the overall training costs to 0.5$\times$, without compromising the model performance compared to token-level training. Source code: \url{https://github.com/shaochenze/P

atchTrain}.
****
Jiecheng Lu,Yan Sun,Shihao Yang
In-context Time Series Predictor
Recent Transformer-based large language models (LLMs) demonstrate in-context learning ability to perform various functions based solely on the provided context, without updating model parameters. To fully utilize the in-context capabilities in time series forecasting (TSF) problems, unlike previous Transformer-based or LLM-based time series forecasting methods, we reformulate "time series forecasting tasks" as input tokens by constructing a series of (lookback, future) pairs within the tokens. This method aligns more closely with the inherent in-context mechanisms and is more parameter-efficient without the need of using pre-trained LLM parameters. Furthermore, it addresses issues such as overfitting in existing Transformer-based TSF models, consistently achieving better performance across full-data, few-shot, and zero-shot settings compared to previous architectures.
****
Jiaxin Wen,Jian Guan,Hongning Wang,Wei Wu,Minlie Huang
CodePlan: Unlocking Reasoning Potential in Large Language Models by Scaling Code-form Planning
Despite the remarkable success of large language models (LLMs) on traditional natural language processing tasks, their planning ability remains a critical bottleneck in tackling complex multi-step reasoning tasks. Existing approaches mainly rely on prompting or task-specific fine-tuning, often suffering from weak robustness and cross-task generalization. To address the limitation, we introduce CodePlan, a scalable paradigm that empowers LLMs to generate and follow code-form plans---pseudocode that outlines high-level, structured reasoning processes. By leveraging the structured and versatile nature of code, CodePlan effectively captures the rich semantics and control flows inherent to sophisticated reasoning. Importantly, CodePlan allows the automatic extraction of code-form plans from massive, wide-ranging text corpora without the need for curated, task-specific datasets. This enables it to scale up efficiently and improve reasoning capabilities across diverse scenarios. To train CodePlan, we construct a large-scale dataset of 2M examples that integrate code-form plans with standard prompt-response pairs from existing corpora. With minimal computation overhead during both training and inference, CodePlan achieves a 25.1\% relative improvement compared with directly generating responses, averaged across 13 challenging multi-step reasoning benchmarks, spanning mathematical reasoning, symbolic reasoning, instruction-following, multi-hop QA, and decision-making tasks. Further analysis reveals CodePlan's increasing performance gains on more complex reasoning tasks, as well as significant data efficiency thanks to its generalization ability.
****
Ryan Liu,Jiayi Geng,Joshua Peterson,Ilia Sucholutsky,Thomas L. Griffiths
Large Language Models Assume People are More Rational than We Really are
In order for AI systems to communicate effectively with people, they must understand how we make decisions. However, people's decisions are not always rational, so the implicit internal models of human decision-making in Large Language Models (LLMs) must account for this. Previous empirical evidence seems to suggest that these implicit models are accurate --- LLMs offer believable proxies of human behavior, acting how we expect humans would in everyday interactions. However, by comparing LLM behavior and predictions to a large dataset of human decisions, we find that this is actually not the case: when both simulating and predicting people's choices, a suite of cutting-edge LLMs (GPT-4o \& 4-Turbo, Llama-3-8B \& 70B, Claude 3 Opus) assume that people are more rational than we really are. Specifically, these models deviate from human behavior and align more closely with a classic model of rational choice --- expected value theory. Interestingly, people also tend to assume that other people are rational when interpreting their behavior. As a consequence, when we compare the inferences that LLMs and people draw from the decisions of others using another psychological dataset, we find that these inferences are highly correlated. Thus, the implicit decision-making models of LLMs appear to be aligned with the human expectation that other people

will act rationally, rather than with how people actually act.
****
Tuan Duc Ngo,Peiye Zhuang,Evangelos Kalogerakis,Chuang Gan,Sergey Tulyakov,Hsin-Ying Lee,Chaoyang Wang
DELTA: DENSE EFFICIENT LONG-RANGE 3D TRACKING FOR ANY VIDEO
Tracking dense 3D motion from monocular videos remains challenging, particularly when aiming for pixel-level precision over long sequences. We introduce DELTA, a novel method that efficiently tracks every pixel in 3D space, enabling accurate motion estimation across entire videos. Our approach leverages a joint global-local attention mechanism for reduced-resolution tracking, followed by a transformer-based upsampler to achieve high-resolution predictions. Unlike existing methods, which are limited by computational inefficiency or sparse tracking, DELTA delivers dense 3D tracking at scale, running over 8x faster than previous methods while achieving state-of-the-art accuracy. Furthermore, we explore the impact of depth representation on tracking performance and identify log-depth as the optimal choice. Extensive experiments demonstrate the superiority of DELTA on multiple benchmarks, achieving new state-of-the-art results in both 2D and 3D dense tracking tasks. Our method provides a robust solution for applications requiring fine-grained, long-term motion tracking in 3D space.
****
Mahalakshmi Sabanayagam,Lukas Gosch,Stephan Günnemann,Debarghya Ghoshdastidar
Exact Certification of (Graph) Neural Networks Against Label Poisoning
Machine learning models are highly vulnerable to label flipping, i.e., the adversarial modification (poisoning) of training labels to compromise performance. Thus, deriving robustness certificates is important to guarantee that test predictions remain unaffected and to understand worst-case robustness behavior. However, for Graph Neural Networks (GNNs), the problem of certifying label flipping has so far been unsolved. We change this by introducing an exact certification method, deriving both sample-wise and collective certificates. Our method leverages the Neural Tangent Kernel (NTK) to capture the training dynamics of wide networks enabling us to reformulate the bilevel optimization problem representing label flipping into a Mixed-Integer Linear Program (MILP). We apply our method to certify a broad range of GNN architectures in node classification tasks. Thereby, concerning the worst-case robustness to label flipping: (i) we establish hierarchies of GNNs on different benchmark graphs; (ii) quantify the effect of architectural choices such as activations, depth and skip-connections; and surprisingly, (iii) uncover a novel phenomenon of the robustness plateauing for intermediate perturbation budgets across all investigated datasets and architectures. While we focus on GNNs, our certificates are applicable to sufficiently wide NNs in general through their NTK. Thus, our work presents the first exact certificate to a poisoning attack ever derived for neural networks, which could be of independent interest. The code is available at https://github.com/saper0/qpcert.
****
Jiduan Wu,Rediet Abebe,Moritz Hardt,Ana-Andreea Stoica
Policy Design in Long-run Welfare Dynamics
Improving social welfare is a complex challenge requiring policymakers to optimize objectives across multiple time horizons. Evaluating the impact of such policies presents a fundamental challenge, as those that appear suboptimal in the short run may yield significant long-term benefits. We tackle this challenge by analyzing the long-term dynamics of two prominent policy frameworks: Rawlsian policies, which prioritize those with the greatest need, and utilitarian policies, which maximize immediate welfare gains. Conventional wisdom suggests these policies are at odds, as Rawlsian policies are assumed to come at the cost of reducing the average social welfare, which their utilitarian counterparts directly optimize. We challenge this assumption by analyzing these policies in a sequential decision-making framework where individuals' welfare levels stochastically decay over time, and policymakers can intervene to prevent this decay. Under reasonable assumptions, we prove that interventions following Rawlsian policies can outperform utilitarian policies in the long run, even when the latter dominate in the short run. We characterize the exact conditions under which Rawlsian policies ca

n outperform utilitarian policies. We further illustrate our theoretical finding s using simulations, which highlight the risks of evaluating policies based sole ly on their short-term effects. Our results underscore the necessity of consider ing long-term horizons in designing and evaluating welfare policies; the true ef ficacy of even well-established policies may only emerge over time.
****
Haoxian Chen,Hanyang Zhao,Henry Lam,David Yao,Wenpin Tang
MallowsPO: Fine-Tune Your LLM with Preference Dispersions
Direct Preference Optimization (DPO) has recently emerged as a popular approach to improve reinforcement learning from human feedback (RLHF), leading to better techniques to fine-tune large language models (LLM). A weakness of DPO, however, lies in its lack of capability to characterize the diversity of human preferenc es. Inspired by Mallows' theory of preference ranking, we develop in this paper a new approach, the *MallowsPO*. A distinct feature of this approach is a *disp ersion index*, which reflects the dispersion of human preference to prompts. We show that existing DPO models can be reduced to special cases of this dispersion index, thus unified with MallowsPO. More importantly, we demonstrate empiricall y how to use this dispersion index to enhance the performance of DPO in a broad array of benchmark tasks, from synthetic bandit selection to controllable genera tion and dialogues, while maintaining great generalization capabilities. Mallows PO is also compatible with other SOTA offline preference optimization methods, b oosting nearly 2\% extra LC win rate when used as a plugin for fine-tuning Llama 3-Instruct.
****
Taekyung Kim,Byeongho Heo,Dongyoon Han
Morphing Tokens Draw Strong Masked Image Models
Masked image modeling (MIM) has emerged as a promising approach for pre-training Vision Transformers (ViTs). MIMs predict masked tokens token-wise to recover ta rget signals that are tokenized from images or generated by pre-trained models l ike vision-language models. While using tokenizers or pre-trained models is viab le, they often offer spatially inconsistent supervision even for neighboring tok ens, hindering models from learning discriminative representations. Our pilot st udy identifies spatial inconsistency in supervisory signals and suggests that ad dressing it can improve representation learning. Building upon this insight, we introduce Dynamic Token Morphing (DTM), a novel method that dynamically aggregat es tokens while preserving context to generate contextualized targets, thereby l ikely reducing spatial inconsistency. DTM is compatible with various SSL framewo rks; we showcase significantly improved MIM results, barely introducing extra tr aining costs. Our method facilitates MIM training by using more spatially consis tent targets, resulting in improved training trends as evidenced by lower losses . Experiments on ImageNet-1K and ADE20K demonstrate DTM's superiority, which sur passes complex state-of-the-art MIM methods. Furthermore, the evaluation of tran sfer learning on downstream tasks like iNaturalist, along with extensive empiric al studies, supports DTM's effectiveness.
****
Hongkun Dou,Zeyu Li,Jinyang Du,Lijun Yang,Wen Yao,Yue Deng
Hybrid Regularization Improves Diffusion-based Inverse Problem Solving
Diffusion models, recognized for their effectiveness as generative priors, have become essential tools for addressing a wide range of visual challenges. Recentl y, there has been a surge of interest in leveraging Denoising processes for Regu larization (DR) to solve inverse problems. However, existing methods often face issues such as mode collapse, which results in excessive smoothing and diminishe d diversity. In this study, we perform a comprehensive analysis to pinpoint the root causes of gradient inaccuracies inherent in DR. Drawing on insights from di ffusion model distillation, we propose a novel approach called Consistency Regul arization (CR), which provides stabilized gradients without the need for ODE sim ulations. Building on this, we introduce Hybrid Regularization (HR), a unified f ramework that combines the strengths of both DR and CR, harnessing their synergi stic potential. Our approach proves to be effective across a broad spectrum of i nverse problems, encompassing both linear and nonlinear scenarios, as well as va

rious measurement noise statistics. Experimental evaluations on benchmark datase
ts, including FFHQ and ImageNet, demonstrate that our proposed framework not onl
y achieves highly competitive results compared to state-of-the-art methods but a
lso offers significant reductions in wall-clock time and memory consumption.
****

Joshua Engels,Eric J Michaud,Isaac Liao,Wes Gurnee,Max Tegmark
Not All Language Model Features Are One-Dimensionally Linear
Recent work has proposed that language models perform computation by manipulatin
g one-dimensional representations of concepts ("features") in activation space.
In contrast, we explore whether some language model representations may be inher
ently multi-dimensional. We begin by developing a rigorous definition of irreduc
ible multi-dimensional features based on whether they can be decomposed into eit
her independent or non-co-occurring lower-dimensional features. Motivated by the
se definitions, we design a scalable method that uses sparse autoencoders to aut
omatically find multi-dimensional features in GPT-2 and Mistral 7B. These auto-d
iscovered features include strikingly interpretable examples, e.g. $\textit{circ
ular}$ features representing days of the week and months of the year. We identif
y tasks where these exact circles are used to solve computational problems invol
ving modular arithmetic in days of the week and months of the year. Next, we pro
vide evidence that these circular features are indeed the fundamental unit of co
mputation in these tasks with intervention experiments on Mistral 7B and Llama 3
 8B, and we examine the continuity of the days of the week feature in Mistral 7B
. Overall, our work argues that understanding multi-dimensional features is nece
ssary to mechanistically decompose some model behaviors.
****

ZeMing Gong,Austin Wang,Xiaoliang Huo,Joakim Bruslund Haurum,Scott C. Lowe,Graha
m W. Taylor,Angel X Chang
CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale
Measuring biodiversity is crucial for understanding ecosystem health. While prio
r works have developed machine learning models for taxonomic classification of p
hotographic images and DNA separately, in this work, we introduce a multi-modal
approach combining both, using CLIP-style contrastive learning to align images,
barcode DNA, and text-based representations of taxonomic labels in a unified emb
edding space. This allows for accurate classification of both known and unknown
insect species without task-specific fine-tuning, leveraging contrastive learnin
g for the first time to fuse DNA and image data. Our method surpasses previous s
ingle-modality approaches in accuracy by over 8% on zero-shot learning tasks, sh
owcasing its effectiveness in biodiversity studies.
****

Shpresim Sadiku,Moritz Wagner,Sebastian Pokutta
GSE: Group-wise Sparse and Explainable Adversarial Attacks
Sparse adversarial attacks fool deep neural networks (DNNs) through minimal pixe
l perturbations, often regularized by the ℓ_0 norm. Recent efforts have rep
laced this norm with a structural sparsity regularizer, such as the nuclear grou
p norm, to craft group-wise sparse adversarial attacks. The resulting perturbati
ons are thus explainable and hold significant practical relevance, shedding ligh
t on an even greater vulnerability of DNNs. However, crafting such attacks poses
 an optimization challenge, as it involves computing norms for groups of pixels
within a non-convex objective. We address this by presenting a two-phase algorit
hm that generates group-wise sparse attacks within semantically meaningful areas
 of an image. Initially, we optimize a quasinorm adversarial loss using the $1/2
$-quasinorm proximal operator tailored for non-convex programming. Subsequently,
 the algorithm transitions to a projected Nesterov's accelerated gradient descen
t with 2-norm regularization applied to perturbation magnitudes. Rigorous eval
uations on CIFAR-10 and ImageNet datasets demonstrate a remarkable increase in g
roup-wise sparsity, e.g., 50.9% on CIFAR-10 and 38.4% on ImageNet (average
 case, targeted attack). This performance improvement is accompanied by signific
antly faster computation times, improved interpretability, and a 100% attack
success rate.
****

Jianxin Zhang,Josh Viktorov,Doosan Jung,Emily Pitler
Efficient Training of Neural Stochastic Differential Equations by Matching Finite Dimensional Distributions
Neural Stochastic Differential Equations (Neural SDEs) have emerged as powerful mesh-free generative models for continuous stochastic processes, with critical applications in fields such as finance, physics, and biology. Previous state-of-the-art methods have relied on adversarial training, such as GANs, or on minimizing distance measures between processes using signature kernels. However, GANs suffer from issues like instability, mode collapse, and the need for specialized training techniques, while signature kernel-based methods require solving linear PDEs and backpropagating gradients through the solver, whose computational complexity scales quadratically with the discretization steps. In this paper, we identify a novel class of strictly proper scoring rules for comparing continuous Markov processes. This theoretical finding naturally leads to a novel approach called Finite Dimensional Matching (FDM) for training Neural SDEs. Our method leverages the Markov property of SDEs to provide a computationally efficient training objective. This scoring rule allows us to bypass the computational overhead associated with signature kernels and reduces the training complexity from $O(D^2)$ to $O(D)$ per epoch, where D represents the number of discretization steps of the process. We demonstrate that FDM achieves superior performance, consistently outperforming existing methods in terms of both computational efficiency and generative quality.
****
Dengsheng Chen,Jie Hu,Xiaoming Wei,Enhua Wu
Denoising with a Joint-Embedding Predictive Architecture
Joint-embedding predictive architectures (JEPAs) have shown substantial promise in self-supervised representation learning, yet their application in generative modeling remains underexplored. Conversely, diffusion models have demonstrated significant efficacy in modeling arbitrary probability distributions. In this paper, we introduce Denoising with a Joint-Embedding Predictive Architecture (D-JEPA), pioneering the integration of JEPA within generative modeling. By recognizing JEPA as a form of masked image modeling, we reinterpret it as a generalized next-token prediction strategy, facilitating data generation in an auto-regressive manner. Furthermore, we incorporate diffusion loss to model the per-token probability distribution, enabling data generation in a continuous space. We also adapt flow matching loss as an alternative to diffusion loss, thereby enhancing the flexibility of D-JEPA. Empirically, with increased GFLOPs, D-JEPA consistently achieves lower FID scores with fewer training epochs, indicating its good scalability. Our base, large, and huge models outperform all previous generative models across all scales on ImageNet conditional generation benchmarks. Beyond image generation, D-JEPA is well-suited for other continuous data modeling, including video and audio.
****
Daniel Watson,Saurabh Saxena,Lala Li,Andrea Tagliasacchi,David J. Fleet
Controlling Space and Time with Diffusion Models
We present 4DiM, a cascaded diffusion model for 4D novel view synthesis (NVS), supporting generation with arbitrary camera trajectories and timestamps, in natural scenes, conditioned on one or more images. With a novel architecture and sampling procedure, we enable training on a mixture of 3D (with camera pose), 4D (pose+time) and video (time but no pose) data, which greatly improves generalization to unseen images and camera pose trajectories over prior works which generally operate in limited domains (e.g., object centric).
4DiM is the first-ever NVS method with intuitive metric-scale camera pose control enabled by our novel calibration pipeline for structure-from-motion-posed data. Experiments demonstrate that 4DiM outperforms prior 3D NVS models both in terms of
image fidelity and pose alignment, while also enabling the generation of scene dynamics. 4DiM provides a general framework for a variety of tasks including single-image-to-3D, two-image-to-video (interpolation and extrapolation), and pose-conditioned video-to-video translation, which we illustrate qualitatively on a va

riety of scenes.
See https://4d-diffusion.github.io for video samples.
****

Yuzhe Gu,Wenwei Zhang,Chengqi Lyu,Dahua Lin,Kai Chen
Mask-DPO: Generalizable Fine-grained Factuality Alignment of LLMs
Large language models (LLMs) exhibit hallucinations (i.e., unfaithful or nonsens ical information) when serving as AI assistants in various domains. Since halluc inations always come with truthful content in the LLM responses, previous factua lity alignment methods that conduct response-level preference learning inevitabl y introduced noises during training. Therefore, this paper proposes a fine-grain ed factuality alignment method based on Direct Preference Optimization (DPO), ca lled Mask-DPO. Incorporating sentence-level factuality as mask signals, Mask-DP O only learns from factually correct sentences in the preferred samples and prev ents the penalty on factual contents in the not preferred samples, which resolve s the ambiguity in the preference learning. Extensive experimental results demon strate that Mask-DPO can significantly improve the factuality of LLMs responses to questions from both in-domain and out-of-domain datasets, although these ques tions and their corresponding topics are unseen during training. Only trained on the ANAH train set, the score of Llama3.1-8B-Instruct on the ANAH test set is i mproved from 49.19% to 77.53%, even surpassing the score of Llama3.1-70B-Instruc t (53.44%), while its FactScore on the out-of-domain Biography dataset is also i mproved from 30.29% to 39.39%. We further study the generalization property of M ask-DPO using different training sample scaling strategies and find that scaling the number of topics in the dataset is more effective than the number of questi ons. We provide a hypothesis of what factual alignment is doing with LLMs, on th e implication of this phenomenon, and conduct proof-of-concept experiments to ve rify it. We hope the method and the findings pave the way for future research on scaling factuality alignment.
****

Yilin Wang,chuan guo,Yuxuan Mu,Muhammad Gohar Javed,Xinxin Zuo,Juwei Lu,Hai Jian g,Li cheng
MotionDreamer: One-to-Many Motion Synthesis with Localized Generative Masked Tra nsformer
Generative masked transformer have demonstrated remarkable success across variou s content generation tasks, primarily due to their ability to effectively model large-scale dataset distributions with high consistency. However, in the animati on domain, large datasets are not always available. Applying generative masked m odeling to generate diverse instances from a single MoCap reference may lead to overfitting, a challenge that remains unexplored. In this work, we present Motio nDreamer, a localized masked modeling paradigm designed to learn motion internal patterns from a given motion with arbitrary topology and duration. By embedding the given motion into quantized tokens with a novel distribution regularization method, MotionDreamer constructs a robust and informative codebook for local mo tion patterns. Moreover, a sliding window local attention is introduced in our m asked transformer, enabling the generation of natural yet diverse animations tha t closely resemble the reference motion patterns. As demonstrated through compre hensive experiments, MotionDreamer outperforms the state-of-the-art methods that are typically GAN or Diffusion-based in both faithfulness and diversity. Thanks to the consistency and robustness of quantization-based approach, MotionDreamer can also effectively perform downstream tasks such as temporal motion editing, crowd motion synthesis, and beat-aligned dance generation, all using a single re ference motion. Our implementation, learned models and results are to be made pu blicly available upon paper acceptance.
****

Ziheng Chen,Yue Song,Xiaojun Wu,Nicu Sebe
Gyrogroup Batch Normalization
Several Riemannian manifolds in machine learning, such as Symmetric Positive Def inite (SPD), Grassmann, spherical, and hyperbolic manifolds, have been proven to admit gyro structures, thus enabling a principled and effective extension of Eu clidean Deep Neural Networks (DNNs) to manifolds. Inspired by this, this study i

ntroduces a general Riemannian Batch Normalization (RBN) framework on gyrogroups , termed GyroBN. We identify the least requirements to guarantee GyroBN with the oretical control over sample statistics, referred to as \textit{pseudo-reduction } and \textit{gyroisometric gyrations}, which are satisfied by all the existing gyrogroups in machine learning. Besides, our GyroBN incorporates several existin g normalization methods, including the one on general Lie groups and different t ypes of RBN on the non-group SPD geometry. Lastly, we instantiate our GyroBN on the Grassmannian and hyperbolic spaces. Experiments on the Grassmannian and hype rbolic networks demonstrate the effectiveness of our GyroBN. The code is availab le at https://github.com/GitZH-Chen/GyroBN.git.
****

Ça■lar H■zl■,Ça■atay Y■ld■z,Matthias Bethge,S. T. John,Pekka Marttinen
Identifying latent state transitions in non-linear dynamical systems
This work aims to recover the underlying states and their time evolution in a la tent dynamical system from high-dimensional sensory measurements. Previous works on identifiable representation learning in dynamical systems focused on identif ying the latent states, often with linear transition approximations. As such, th ey cannot identify nonlinear transition dynamics, and hence fail to reliably pre dict complex future behavior. Inspired by the advances in nonlinear ICA, we prop ose a state-space modeling framework in which we can identify not just the laten t states but also the unknown transition function that maps the past states to t he present. Our identifiability theory relies on two key assumptions: (i) suffic ient variability in the latent noise, and (ii) the bijectivity of the augmented transition function. Drawing from this theory, we introduce a practical algorith m based on variational auto-encoders. We empirically demonstrate that it improve s generalization and interpretability of target dynamical systems by (i) recover ing latent state dynamics with high accuracy, (ii) correspondingly achieving hig h future prediction accuracy, and (iii) adapting fast to new environments. Addit ionally, for complex real-world dynamics, (iv) it produces state-of the-art futu re prediction results for long horizons, highlighting its usefulness for practic al scenarios.
****

Mohammad Omama,Po-han Li,Sandeep P. Chinchali
Exploiting Distribution Constraints for Scalable and Efficient Image Retrieval
Image retrieval is crucial in robotics and computer vision, with downstream appl ications in robot place recognition and vision-based product recommendations. Mo dern retrieval systems face two key challenges: scalability and efficiency.
State-of-the-art image retrieval systems train specific neural networks for each dataset, an approach that lacks scalability. Furthermore, since retrieval speed is directly proportional to embedding size, existing systems that use large emb eddings lack efficiency. To tackle scalability, recent works propose using off-t he-shelf foundation models. However, these models, though applicable across data sets, fall short in achieving performance comparable to that of dataset-specific models. Our key observation is that, while foundation models capture necessary subtleties for effective retrieval, the underlying distribution of their embeddi ng space can negatively impact cosine similarity searches. We introduce Autoenco ders with Strong Variance Constraints (AE-SVC), which, when used for projection, significantly improves the performance of foundation models. We provide an in-d epth theoretical analysis of AE-SVC. Addressing efficiency, we introduce Single-Shot Similarity Space Distillation ((SS)2D), a novel approach to learn embedding s with adaptive sizes that offers a better trade-off between size and performanc e. We conducted extensive experiments on four retrieval datasets, including Stan -
ford Online Products (SoP) and Pittsburgh30k, using four different off-the-shelf foundation models, including DinoV2 and CLIP. AE-SVC demonstrates up to a 16% i mprovement in retrieval performance, while (SS)2D shows a further 10% improvemen t for smaller embedding sizes.
****

Shahryar Zehtabi,Dong-Jun Han,Rohit Parasnis,Seyyedali Hosseinalipour,Christophe r Brinton

Decentralized Sporadic Federated Learning: A Unified Algorithmic Framework with Convergence Guarantees

Decentralized federated learning (DFL) captures FL settings where both (i) model updates and (ii) model aggregations are exclusively carried out by the clients without a central server. Existing DFL works have mostly focused on settings where clients conduct a fixed number of local updates between local model exchanges, overlooking heterogeneity and dynamics in communication and computation capabilities. In this work, we propose Decentralized Sporadic Federated Learning ($\texttt{DSpodFL}$), a DFL methodology built on a generalized notion of *sporadicity* in both local gradient and aggregation processes. $\texttt{DSpodFL}$ subsumes many existing decentralized optimization methods under a unified algorithmic framework by modeling the per-iteration (i) occurrence of gradient descent at each client and (ii) exchange of models between client pairs as arbitrary indicator random variables, thus capturing *heterogeneous and time-varying* computation/communication scenarios. We analytically characterize the convergence behavior of $\texttt{DSpodFL}$ for both convex and non-convex models and for both constant and diminishing learning rates, under mild assumptions on the communication graph connectivity, data heterogeneity across clients, and gradient noises. We show how our bounds recover existing results from decentralized gradient descent as special cases. Experiments demonstrate that $\texttt{DSpodFL}$ consistently achieves improved training speeds compared with baselines under various system settings.

****

Wei Yao,Haian Yin,Shangzhi Zeng,Jin Zhang

Overcoming Lower-Level Constraints in Bilevel Optimization: A Novel Approach with Regularized Gap Functions

Constrained bilevel optimization tackles nested structures present in constrained learning tasks like constrained meta-learning, adversarial learning, and distributed bilevel optimization.

However, existing bilevel optimization methods mostly are typically restricted to specific constraint settings, such as linear lower-level constraints.

In this work, we overcome this limitation and develop a new single-loop, Hessian-free constrained bilevel algorithm capable of handling more general lower-level constraints.

We achieve this by employing a doubly regularized gap function tailored to the constrained lower-level problem, transforming constrained bilevel optimization into an equivalent single-level optimization problem with a single smooth constraint.

We rigorously establish the non-asymptotic convergence analysis of the proposed algorithm under the convexity of lower-level problem, avoiding the need for strong convexity assumptions on the lower-level objective or coupling convexity assumptions on lower-level constraints found in existing literature.

Additionally, the generality of our method allows for its extension to bilevel optimization with minimax lower-level problem.

We evaluate the effectiveness and efficiency of our algorithm on various synthetic problems, typical hyperparameter learning tasks, and generative adversarial network.

****

Michael JQ Zhang,W. Bradley Knox,Eunsol Choi

Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions

Large language models (LLMs) must often respond to highly ambiguous user requests. In such cases, the LLM's best response may be to ask a clarifying question to elicit more information. Existing LLMs often respond by presupposing a single interpretation of such ambiguous requests, frustrating users who intended a different interpretation. We speculate this is caused by current preference data labeling practice, where LLM responses are evaluated only on their prior contexts. To address this, we assign preference labels by simulating their expected outcomes in future turns. This allows LLMs to learn to ask clarifying questions when it can generate responses that are tailored to each user interpretation in future turns. On open-domain QA datasets with multiple annotations, we evaluate systems

based on their ability to ask clarifying questions to recover each user's interpretation and expected answer. We compare systems trained using our proposed preference labeling methods against standard methods, which assign preferences based on only prior context. Our method achieves a 5% improvement in F1 measured against the answer set from different interpretations of each query, showing the value of modeling future conversation turns. We further demonstrate that our method can be used to train models to judiciously determine when to ask clarifying questions, directly answering the question when clarification is unnecessary. In our experiments, we find that our method achives a 3% improvement in accuracy of such judgments over existing methods.
****

Kaizheng Wang,Fabio Cuzzolin,Keivan Shariatmadar,David Moens,Hans Hallez
Credal Wrapper of Model Averaging for Uncertainty Estimation in Classification
This paper presents an innovative approach, called credal wrapper, to formulating a credal set representation of model averaging for Bayesian neural networks (BNNs) and deep ensembles (DEs), capable of improving uncertainty estimation in classification tasks. Given a finite collection of single predictive distributions derived from BNNs or DEs, the proposed credal wrapper approach extracts an upper and a lower probability bound per class, acknowledging the epistemic uncertainty due to the availability of a limited amount of distributions. Such probability intervals over classes can be mapped on a convex set of probabilities (a credal set) from which, in turn, a unique prediction can be obtained using a transformation called intersection probability transformation. In this article, we conduct extensive experiments on several out-of-distribution (OOD) detection benchmarks, encompassing various dataset pairs (CIFAR10/100 vs SVHN/Tiny-ImageNet, CIFAR10 vs CIFAR10-C, CIFAR100 vs CIFAR100-C and ImageNet vs ImageNet-O) and using different network architectures (such as VGG16, ResNet-18/50, EfficientNet B2, and ViT Base). Compared to the BNN and DE baselines, the proposed credal wrapper method exhibits superior performance in uncertainty estimation and achieves a lower expected calibration error on corrupted data.
****

Weiwei Lin,Chenhang HE
Continuous Autoregressive Modeling with Stochastic Monotonic Alignment for Speech Synthesis
We propose a novel autoregressive modeling approach for speech synthesis, combining a variational autoencoder (VAE) with a multi-modal latent space and an autoregressive model that uses Gaussian Mixture Models (GMM) as the conditional probability distribution. Unlike previous methods that rely on residual vector quantization, our model leverages continuous speech representations from the VAE's latent space, greatly simplifying the training and inference pipelines. We also introduce a stochastic monotonic alignment mechanism to enforce strict monotonic alignments. Our approach significantly outperforms the state-of-the-art autoregressive model VALL-E in both subjective and objective evaluations, achieving these results with only 10.3\% of VALL-E's parameters. This demonstrates the potential of continuous speech language models as a more efficient alternative to existing quantization-based speech language models. Sample audio can be found at \url{https://tinyurl.com/gmm-lm-tts}.
****

Haoran Xu,Kenton Murray,Philipp Koehn,Hieu Hoang,Akiko Eriguchi,Huda Khayrallah
X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale
Large language models (LLMs) have achieved remarkable success across various NLP tasks with a focus on English due to English-centric pre-training and limited multilingual data. In this work, we focus on the problem of translation, and while some multilingual LLMs claim to support for hundreds of languages, models often fail to provide high-quality responses for mid- and low-resource languages, leading to imbalanced performance heavily skewed in favor of high-resource languages. We introduce **X-ALMA**, a model designed to ensure top-tier performance across 50 diverse languages, regardless of their resource levels. X-ALMA surpasses state-of-the-art open-source multilingual LLMs, such as Aya-101 and Aya-23,

in every single translation direction on the FLORES-200 and WMT'23 test datasets according to COMET-22. This is achieved by plug-and-play language-specific module architecture to prevent language conflicts during training and a carefully designed training regimen with novel optimization methods to maximize the translation performance. After the final stage of training regimen, our proposed **A**daptive **R**ejection **P**reference **O**ptimization (**ARPO**) surpasses existing preference optimization methods in translation tasks.
****
Makoto Shing,Kou Misaki,Han Bao,Sho Yokoi,Takuya Akiba
TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models
Causal language models have demonstrated remarkable capabilities, but their size poses significant challenges for deployment in resource-constrained environments. Knowledge distillation, a widely-used technique for transferring knowledge from a large teacher model to a small student model, presents a promising approach for model compression.
A significant remaining issue lies in the major differences between teacher and student models, namely the substantial capacity gap, mode averaging, and mode collapse, which pose barriers during distillation.
To address these issues, we introduce $\textit{Temporally Adaptive Interpolated Distillation (TAID)}$, a novel knowledge distillation approach that dynamically interpolates student and teacher distributions through an adaptive intermediate distribution, gradually shifting from the student's initial distribution towards the teacher's distribution. We provide a theoretical analysis demonstrating TAID's ability to prevent mode collapse and empirically show its effectiveness in addressing the capacity gap while balancing mode averaging and mode collapse.
Our comprehensive experiments demonstrate TAID's superior performance across various model sizes and architectures in both instruction tuning and pre-training scenarios. Furthermore, we showcase TAID's practical impact by developing two state-of-the-art compact foundation models: $\texttt{TAID-LLM-1.5B}$ for language tasks and $\texttt{TAID-VLM-2B}$ for vision-language tasks.
These results demonstrate TAID's effectiveness in creating high-performing and efficient models, advancing the development of more accessible AI technologies.
****
Yongshuo Zong,Ondrej Bohdal,Timothy Hospedales
VL-ICL Bench: The Devil in the Details of Multimodal In-Context Learning
Large language models (LLMs) famously exhibit emergent in-context learning (ICL) - the ability to rapidly adapt to new tasks using few-shot examples provided as a prompt, without updating the model's weights. Built on top of LLMs, vision large language models (VLLMs) have advanced significantly in areas such as recognition, reasoning, and grounding. However, investigations into multimodal ICL have predominantly focused on few-shot visual question answering (VQA), and image captioning, which we will show neither exploit the strengths of ICL, nor test its limitations. The broader capabilities and limitations of multimodal ICL remain under-explored. In this study, we introduce a comprehensive benchmark VL-ICL Bench for multimodal in-context learning, encompassing a broad spectrum of tasks that involve both images and text as inputs and outputs, and different types of challenges, from {perception to reasoning and long context length}. We evaluate the abilities of state-of-the-art VLLMs against this benchmark suite, revealing their diverse strengths and weaknesses, and showing that even the most advanced models, such as GPT-4, find the tasks challenging. By highlighting a range of new ICL tasks, and the associated strengths and limitations of existing models, we hope that our dataset will inspire future work on enhancing the in-context learning capabilities of VLLMs, as well as inspire new applications that leverage VLLM ICL. Project page is at https://ys-zong.github.io/VL-ICL/
****
Wen Wen,Han Li,Tieliang Gong,Hong Chen
Towards Generalization Bounds of GCNs for Adversarially Robust Node Classification
Adversarially robust generalization of Graph Convolutional Networks (GCNs) has g

arnered significant attention in various security-sensitive application areas, d
riven by intrinsic adversarial vulnerability. Albeit remarkable empirical advanc
ement, theoretical understanding of the generalization behavior of GCNs subjecte
d to adversarial attacks remains elusive. To make progress on the mystery, we es
tablish unified high-probability generalization bounds for GCNs in the context o
f node classification, by leveraging adversarial Transductive Rademacher Complex
ity (TRC) and developing a novel contraction technique on graph convolution. Our
 bounds capture the interaction between generalization error and adversarial per
turbations, revealing the importance of key quantities in mitigating the negativ
e effects of perturbations, such as low-dimensional feature projection, perturba
tion-dependent norm regularization, normalized graph matrix, proper number of ne
twork layers, etc. Furthermore, we provide TRC-based bounds of popular GCNs with
 ℓ_r-norm-additive perturbations for arbitrary $r\geq 1$. A comparison of t
heoretical results demonstrates that specific network architectures (e.g., resid
ual connection) can help alleviate the cumulative effect of perturbations during
 the forward propagation of deep GCNs. Experimental results on benchmark dataset
s validate our theoretical findings.
****

Shashata Sawmya,Linghao Kong,Ilia Markov,Dan Alistarh,Nir N Shavit
Wasserstein Distances, Neuronal Entanglement, and Sparsity
Disentangling polysemantic neurons is at the core of many current approaches to
interpretability of large language models. Here we attempt to study how disentan
glement can be used to understand performance, particularly under weight sparsit
y, a leading post-training optimization technique. We suggest a novel measure fo
r estimating neuronal entanglement: the Wasserstein distance of a neuron's outpu
t distribution to a Gaussian. Moreover, we show the existence of a small number
of highly entangled "Wasserstein Neurons" in each linear layer of an LLM, charac
terized by their highly non-Gaussian output distributions, their role in mapping
 similar inputs to dissimilar outputs, and their significant impact on model acc
uracy. To study these phenomena, we propose a new experimental framework for dis
entangling polysemantic neurons. Our framework separates each layer's inputs to
create a mixture of experts where each neuron's output is computed by a mixture
of neurons of lower Wasserstein distance, each better at maintaining accuracy wh
en sparsified without retraining. We provide strong evidence that this is becaus
e the mixture of sparse experts is effectively disentangling the input-output re
lationship of individual neurons, in particular the difficult Wasserstein neuron
s.
****

Jindou Jia,Zihan Yang,Meng Wang,Kexin Guo,Jianfei Yang,Xiang Yu,Lei Guo
Feedback Favors the Generalization of Neural ODEs
The well-known generalization problem hinders the application of artificial neur
al networks in continuous-time prediction tasks with varying latent dynamics. In
 sharp contrast, biological systems can neatly adapt to evolving environments be
nefiting from real-time feedback mechanisms. Inspired by the feedback philosophy
, we present feedback neural networks, showing that a feedback loop can flexibly
 correct the learned latent dynamics of neural ordinary differential equations (
neural ODEs), leading to a prominent generalization improvement. The feedback ne
ural network is a novel two-DOF neural network, which possesses robust performan
ce in unseen scenarios with no loss of accuracy performance on previous tasks. A
 linear feedback form is presented to correct the learned latent dynamics firstl
y, with a convergence guarantee. Then, domain randomization is utilized to learn
 a nonlinear neural feedback form. Finally, extensive tests including trajectory
 prediction of a real irregular object and model predictive control of a quadrot
or with various uncertainties, are implemented, indicating significant improveme
nts over state-of-the-art model-based and learning-based methods.
****

Qining Zhang,Lei Ying
Zeroth-Order Policy Gradient for Reinforcement Learning from Human Feedback with
out Reward Inference
Reward inference (learning a reward model from human preferences) is a critical

intermediate step in the Reinforcement Learning from Human Feedback (RLHF) pipel
ine for fine-tuning Large Language Models (LLMs). In practice, RLHF faces fundam
ental challenges such as distribution shift, reward model overfitting, and probl
em misspecification. An alternative approach is direct policy optimization witho
ut reward inference, such as Direct Preference Optimization (DPO), which provide
s a much simpler pipeline and has shown empirical success in LLM applications. H
owever, DPO utilizes the closed-form expression between the optimal policy and t
he reward function, which is only suitable under the bandit setting or determini
stic MDPs. This paper develops two RLHF algorithms without reward inference for
general RL problems beyond bandits and deterministic MDPs, and general preferenc
e models beyond the Bradley-Terry model. The key idea is to estimate the local v
alue function difference from human preferences and then approximate the policy
gradient with a zeroth-order gradient approximator. For both algorithms, we esta
blish polynomial convergence rates in terms of the number of policy gradient ite
rations, the number of trajectory samples, and human preference queries per iter
ation. Numerical experiments in stochastic environments validate the performance
of our proposed algorithms, outperforming popular RLHF baselines such as DPO an
d PPO. Our paper shows there exist provably efficient methods to solve general R
LHF problems without reward inference.
****

Sajad Movahedi,Antonio Orvieto,Seyed-Mohsen Moosavi-Dezfooli
Geometric Inductive Biases of Deep Networks: The Role of Data and Architecture
In this paper, we propose the *geometric invariance hypothesis (GIH)*, which arg
ues that the input space curvature of a neural network remains invariant under t
ransformation in certain architecture-dependent directions during training. We i
nvestigate a simple, non-linear binary classification problem residing on a plan
e in a high dimensional space and observe that—unlike MPLs—ResNets fai
l to generalize depending on the orientation of the plane. Motivated by this exa
mple, we define a neural network's **average geometry** and **average geometry e
volution** as compact *architecture-dependent* summaries of the model's input-ou
tput geometry and its evolution during training. By investigating the average ge
ometry evolution at initialization, we discover that the geometry of a neural ne
twork evolves according to the data covariance projected onto its average geomet
ry. This means that the geometry only changes in a subset of the input space whe
n the average geometry is low-rank, such as in ResNets. This causes an architect
ure-dependent invariance property in the input space curvature, which we dub GIH
. Finally, we present extensive experimental results to observe the consequences
of GIH and how it relates to generalization in neural networks.
****

Mingyang Zhao,Gaofeng Meng,Dong-ming Yan
Occlusion-aware Non-Rigid Point Cloud Registration via Unsupervised Neural Defor
mation Correntropy
Non-rigid alignment of point clouds is crucial for scene understanding, reconstr
uction, and various computer vision and robotics tasks. Recent advancements in i
mplicit deformation networks for non-rigid registration have significantly reduc
ed the reliance on large amounts of annotated training data. However, existing s
tate-of-the-art methods still face challenges in handling occlusion scenarios. T
o address this issue, this paper introduces an innovative unsupervised method ca
lled Occlusion-Aware Registration (OAR) for non-rigidly aligning point clouds. T
he key innovation of our method lies in the utilization of the adaptive correntr
opy function as a localized similarity measure, enabling us to treat individual
points distinctly. In contrast to previous approaches that solely minimize overa
ll deviations between two shapes, we combine unsupervised implicit neural repres
entations with the maximum correntropy criterion to optimize the deformation of
unoccluded regions. This effectively avoids collapsed, tearing, and other physic
ally implausible results. Moreover, we present a theoretical analysis and establ
ish the relationship between the maximum correntropy criterion and the commonly
used Chamfer distance, highlighting that the correntropy-induced metric can be s
erved as a more universal measure for point cloud analysis. Additionally, we int
roduce

locally linear reconstruction to ensure that regions lacking correspondences bet
ween shapes still undergo physically natural deformations. Our method achieves s
uperior or competitive performance compared to existing approaches, particularly
 when dealing with occluded geometries. We also demonstrate the versatility of o
ur method in challenging tasks such as large deformations, shape interpolation,
and shape completion under occlusion disturbances.
****
Naman Jain,King Han,Alex Gu,Wen-Ding Li,Fanjia Yan,Tianjun Zhang,Sida Wang,Arman
do Solar-Lezama,Koushik Sen,Ion Stoica
LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Mode
ls for Code
Large Language Models (LLMs) applied to code-related applications have emerged a
s a prominent field, attracting significant interest from academia and industry.
 However, as new and improved LLMs are developed, existing evaluation benchmarks
 (e.g., HumanEvla, MBPP) are no longer sufficient for assessing their capabiliti
es suffering from data contamination, overfitting, saturation, and focus on mere
ly code generation. In this work, we propose LiveCodeBench, a comprehensive and
contamination-free evaluation of LLMs for code, which collects new problems over
 time from contests across three competition platforms, Leetcode, Atcoder, and C
odeforces. Notably, our benchmark also focuses on a broader range of code-relate
d capabilities, such as self-repair, code execution, and test output prediction,
 beyond just code generation. Currently, LiveCodeBench hosts over six hundred co
ding problems that were published between May 2023 and Aug 2024. We evaluate ove
r 50 LLMs on LiveCodeBench (LCB for brevity) presenting the largest evaluation s
tudy of code LLMs on competition problems. Based on the study, we present novel
empirical findings on contamination, overfitting, and holistic evaluations. We d
emonstrate that time-segmented evaluations serve as a robust approach to evade c
ontamination; they are successful at detecting contamination across a wide range
 of open and closed models including GPT-4O, Claude, Deepseek, and Codestral. Ne
xt, we highlight overfitting and saturation of traditional coding benchmarks lik
e HumanEvla and demonstrate LCB allows more reliable evaluations. Finally, our h
olistic evaluation scenarios allow for measuring the different capabilities of p
rogramming agents in isolation.
****
Clement Neo,Luke Ong,Philip Torr,Mor Geva,David Krueger,Fazl Barez
Towards Interpreting Visual Information Processing in Vision-Language Models
Vision-Language Models (VLMs) are powerful tools for processing and understandin
g text and images. We study the processing of visual tokens in the language mode
l component of LLaVA, a prominent VLM. Our approach focuses on analyzing the loc
alization of object information, the evolution of visual token representations a
cross layers, and the mechanism of integrating visual information for prediction
s. Through ablation studies, we demonstrated that object identification accuracy
 drops by over 70\% when object-specific tokens are removed. We observed that vi
sual token representations become increasingly interpretable in the vocabulary s
pace across layers, suggesting an alignment with textual tokens corresponding to
 image content. Finally, we found that the model extracts object information fro
m these refined representations at the last token position for prediction, mirro
ring the process in text-only language models for factual association tasks. The
se findings provide crucial insights into how VLMs process and integrate visual
information, bridging the gap between our understanding of language and vision m
odels, and paving the way for more interpretable and controllable multimodal sys
tems.
****
Mingye Zhu,Yi Liu,Lei Zhang,Junbo Guo,Zhendong Mao
On-the-fly Preference Alignment via Principle-Guided Decoding
With the rapidly expanding landscape of large language models, aligning model ge
nerations with human values and preferences is becoming increasingly important.
Popular alignment methods, such as Reinforcement Learning from Human Feedback, h
ave shown significant success in guiding models with greater control. However, t
hese methods require considerable computational resources, which is inefficient,

and substantial collection of training data to accommodate the diverse and plur
alistic nature of human preferences, which is impractical. These limitations sig
nificantly constrain the scope and efficacy of both task-specific and general pr
eference alignment methods. In this work, we introduce On-the-fly Preference Ali
gnment via Principle-Guided Decoding (OPAD) to directly align
model outputs with human preferences during inference, eliminating the need for
fine-tuning. Our approach involves first curating a surrogate solution to an oth
erwise infeasible optimization problem and then designing a principle-guided rew
ard function based on this surrogate. The final decoding policy is derived by ma
ximizing this customized reward, which exploits the discrepancy between the
constrained policy and its unconstrained counterpart. OPAD directly modifies the
model’s predictions during inference, ensuring principle adherence without incu
rring the computational overhead of retraining or fine-tuning. Experiments show
that OPAD achieves competitive or superior performance in both general and perso
nalized alignment tasks, demonstrating its efficiency and effectiveness compared
to state-of-the-art baselines.
****
Chenjia Bai,Yang Zhang,Shuang Qiu,Qiaosheng Zhang,Kang Xu,Xuelong Li
Online Preference Alignment for Language Models via Count-based Exploration
Reinforcement Learning from Human Feedback (RLHF) has shown great potential in f
ine-tuning Large Language Models (LLMs) to align with human preferences. Existin
g methods perform preference alignment from a fixed dataset, which can be limite
d in data coverage and the resulting reward model is hard to generalize in out-o
f-distribution responses. Thus, online RLHF is more desirable to empower the LLM
to explore outside the support of the initial dataset by iteratively collecting
the prompt-response pairs. In this paper, we study the fundamental problem in o
nline RLHF, i.e., how to explore for LLM. We give a theoretical motivation in li
near reward assumption to show that an optimistic reward with an upper confidenc
e bound (UCB) term leads to a provably efficient RLHF policy. Then, we reformula
te our objective to direct preference optimization with an exploration term, whe
re the UCB-term can be converted to a count-based exploration bonus. We further
propose a practical algorithm, named Count-based Online Preference Optimization
(COPO), which leverages a simple coin-flip counting module to estimate the pseud
o-count of a prompt-response pair in previously collected data. COPO encourages
LLMs to balance exploration and preference optimization in an iterative manner,
which enlarges the exploration space and the entire data coverage of iterative L
LM policies. We conduct online RLHF experiments on Zephyr and Llama-3 models. Th
e results on instruction-following and standard academic benchmarks show that CO
PO significantly increases performance.
****
Si Chen,Richong Zhang,Xu Zhang
GeoILP: A Synthetic Dataset to Guide Large-Scale Rule Induction
Inductive logic programming (ILP) is a machine learning approach aiming to learn
explanatory rules from data.
While existing ILP systems can successfully solve small-scale tasks, large-s
cale applications with various language biases are rarely explored.
Besides, it is crucial for a large majority of current ILP systems to requir
e expert-defined language bias, which hampers the development of ILP towards bro
ader utilizations.
In this paper, we introduce GeoILP, a large-scale synthetic dataset of diver
se ILP tasks involving numerous aspects of language bias.
These tasks are built from geometry problems, at the level from textbook exe
rcise to regional International Mathematical Olympiad (IMO), with the help of a
deduction engine.
These problems are elaborately selected to cover all challenging language bi
ases, such as recursion, predicate invention, and high arity.
Experimental results show that no existing method can solve GeoILP tasks.
In addition, along with classic symbolic-form data, we provide image-form da
ta to boost the development of the joint learning of neural perception and symbo
lic rule induction.

****
Elias Abad Rocamora,Grigorios Chrysos,Volkan Cevher
Certified Robustness Under Bounded Levenshtein Distance
Text classifiers suffer from small perturbations, that if chosen adversarially, can dramatically change the output of the model. Verification methods can provide robustness certificates against such adversarial perturbations, by computing a sound lower bound on the robust accuracy. Nevertheless, existing verification methods incur in prohibitive costs and cannot practically handle Levenshtein distance constraints. We propose the first method for computing the Lipschitz constant of convolutional classifiers with respect to the Levenshtein distance. We use these Lipschitz constant estimates for training 1-Lipschitz classifiers. This enables computing the certified radius of a classifier in a single forward pass. Our method, LipsLev, is able to obtain 38.80% and 13.93% verified accuracy at distance 1 and 2 respectively in the AG-News dataset, while being 4 orders of magnitude faster than existing approaches. We believe our work can open the door to more efficient verification in the text domain.
****
Evan Frick,Tianle Li,Connor Chen,Wei-Lin Chiang,Anastasios Nikolas Angelopoulos, Jiantao Jiao,Banghua Zhu,Joseph E. Gonzalez,Ion Stoica
How to Evaluate Reward Models for RLHF
We introduce a new benchmark for reward models that quantifies their ability to produce strong language models through RLHF (Reinforcement Learning from Human Feedback).
The gold-standard approach is to run a full RLHF training pipeline and directly probe downstream LLM performance.
However, this process is prohibitively expensive.
To address this, we build a predictive model of downstream LLM performance by evaluating the reward model on proxy tasks.
These proxy tasks consist of a large-scale human preference and a verifiable correctness preference dataset, in which we measure 12 metrics across 12 domains.
To investigate which reward model metrics are most correlated to gold-standard RLHF outcomes, we launch an end-to-end RLHF experiment on a large-scale crowd-sourced human preference platform to view real reward model downstream performance as ground truth.
Ultimately, we compile our data and findings into Preference Proxy Evaluations (PPE), the first reward model benchmark explicitly linked to post-RLHF real-world human preference performance, which we opensource for public use and further development at https://github.com/lmarena/PPE.
****
Costin-Andrei Oncescu,Sanket Purandare,Stratos Idreos,Sham M. Kakade
Flash Inference: Near Linear Time Inference for Long Convolution Sequence Models and Beyond
While transformers have been at the core of most recent advancements in sequence generative models, their computational cost remains quadratic in sequence length.
Several subquadratic architectures have been proposed to address this computational issue. Some of them, including long convolution sequence models (LCSMs), such as Hyena, address this issue at training time but remain quadratic during inference. We propose a method for speeding up LCSMs' exact inference to quasilinear time, identify the key properties that make this possible, and propose a general framework that exploits these. Our approach, inspired by previous work on relaxed polynomial interpolation, is based on a tiling which helps decrease memory movement and share computation. It has the added benefit of allowing for almost complete parallelization across layers of the position-mixing part of the architecture. Empirically, we provide a proof of concept implementation for Hyena, which gets up to $7.8\times$ end-to-end improvement over standard inference by improving $110\times$ within the position-mixing part.
****
Nikita Kotelevskii,Vladimir Kondratyev,Martin Taká■,Eric Moulines,Maxim Panov
From Risk to Uncertainty: Generating Predictive Uncertainty Measures via Bayesia

n Estimation
There are various measures of predictive uncertainty in the literature, but their relationships to each other remain unclear. This paper uses a decomposition of statistical pointwise risk into components associated with different sources of predictive uncertainty: namely, aleatoric uncertainty (inherent data variability) and epistemic uncertainty (model-related uncertainty). Together with Bayesian methods applied as approximations, we build a framework that allows one to generate different predictive uncertainty measures.

We validate measures, derived from our framework on image datasets by evaluating its performance in detecting out-of-distribution and misclassified instances using the AUROC metric. The experimental results confirm that the measures derived from our framework are useful for the considered downstream tasks.
****
Simiao Li,Yun Zhang,Wei Li,Hanting Chen,Wenjia Wang,Bingyi Jing,Shaohui Lin,Jie Hu
Knowledge Distillation with Multi-granularity Mixture of Priors for Image Super-Resolution
Knowledge distillation (KD) is a promising yet challenging model compression approach that transmits rich learning representations from robust but resource-demanding teacher models to efficient student models. Previous methods for image super-resolution (SR) are often tailored to specific teacher-student architectures, limiting their potential for improvement and hindering broader applications. This work presents a novel KD framework for SR models, the multi-granularity Mixture of Priors Knowledge Distillation (MiPKD), which can be universally applied to a wide range of architectures at both feature and block levels. The teacher’s knowledge is effectively integrated with the student's feature via the Feature Prior Mixer, and the reconstructed feature propagates dynamically in the training phase with the Block Prior Mixer. Extensive experiments illustrate the significance of the proposed MiPKD technique.
****
Yiping Ji,Hemanth Saratchandran,Cameron Gordon,Zeyu Zhang,Simon Lucey
Efficient Learning with Sine-Activated Low-Rank Matrices
Low-rank decomposition has emerged as a vital tool for enhancing parameter efficiency in neural network architectures, gaining traction across diverse applications in machine learning. These techniques significantly lower the number of parameters, striking a balance between compactness and performance. However, a common challenge has been the compromise between parameter efficiency and the accuracy of the model, where reduced parameters often lead to diminished accuracy compared to their full-rank counterparts. In this work, we propose a novel theoretical framework that integrates a sinusoidal function within the low-rank decomposition process. This approach not only preserves the benefits of the parameter efficiency characteristic of low-rank methods but also increases the decomposition's rank, thereby enhancing model performance. Our method proves to be a plug in enhancement for existing low-rank models, as evidenced by its successful application in Vision Transformers (ViT), Large Language Models (LLMs), Neural Radiance Fields (NeRF) and 3D shape modelling.
****
Di Wu,Siyuan Li,Chen Feng,Lu Cao,Yue Zhang,Jie Yang,Mohamad Sawan
Towards Homogeneous Lexical Tone Decoding from Heterogeneous Intracranial Recordings
Recent advancements in brain-computer interfaces (BCIs) and deep learning have made decoding lexical tones from intracranial recordings possible, providing the potential to restore the communication ability of speech-impaired tonal language speakers. However, data heterogeneity induced by both physiological and instrumental factors poses a significant challenge for unified invasive brain tone decoding. Particularly, the existing heterogeneous decoding paradigm (training subject-specific models with individual data) suffers from the intrinsic limitation that fails to learn generalized neural representations and leverages data across subjects. To this end, we introduce Homogeneity-Heterogeneity Disentangled Learn

ing for Neural Representations (H2DiLR), a framework that disentangles and learn
s the homogeneity and heterogeneity from intracranial recordings of multiple sub
jects. To verify the effectiveness of H2DiLR, we collected stereoelectroencephal
ography (sEEG) from multiple participants reading Mandarin materials containing
407 syllables (covering nearly all Mandarin characters). Extensive experiments d
emonstrate that H2DiLR, as a unified decoding paradigm, outperforms the naive he
terogeneous decoding paradigm by a large margin. We also empirically show that H
2DiLR indeed captures homogeneity and heterogeneity during neural representation
 learning.
****
Zhaolin Gao,Wenhao Zhan,Jonathan Daniel Chang,Gokul Swamy,Kianté Brantley,Jason
D. Lee,Wen Sun
Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLH
F
Large Language Models (LLMs) have achieved remarkable success at tasks like summ
arization that involve a single turn of interaction. However, they can still str
uggle with multi-turn tasks like dialogue that require long-term planning. Previ
ous works on multi-turn dialogue extend single-turn reinforcement learning from
human feedback (RLHF) methods to the multi-turn setting by treating all prior di
alogue turns as a long context. Such approaches suffer from covariate shift: the
 conversations in the training set have previous turns generated by some referen
ce policy, which means that low training error may not necessarily correspond to
 good performance when the learner is actually in the conversation loop. In resp
onse, we introduce REgressing the RELative FUture (REFUEL), an efficient policy
optimization approach designed to address multi-turn RLHF in LLMs. REFUEL employ
s a single model to estimate Q-values and trains on self-generated data, addre
ssing the covariate shift issue. REFUEL frames the multi-turn RLHF problem as a
sequence of regression tasks on iteratively collected datasets, enabling ease of
 implementation. Theoretically, we prove that REFUEL can match the performance o
f any policy covered by the training set. Empirically, we evaluate our algorithm
 by using Llama-3.1-70B-it to simulate a user in conversation with our model. RE
FUEL consistently outperforms state-of-the-art methods such as DPO and REBEL acr
oss various settings. Furthermore, despite having only 8 billion parameters, Lla
ma-3-8B-it fine-tuned with REFUEL outperforms Llama-3.1-70B-it on long multi-tur
n dialogues. Implementation of REFUEL can be found at https://github.com/Zhaolin
Gao/REFUEL/, and models trained by REFUEL can be found at https://huggingface.co
/Cornell-AGI.
****
Alexander Tyurin
Tight Time Complexities in Parallel Stochastic Optimization with Arbitrary Compu
tation Dynamics
In distributed stochastic optimization, where parallel and asynchronous methods
are employed, we establish optimal time complexities under virtually any computa
tion behavior of workers/devices/CPUs/GPUs, capturing potential disconnections d
ue to hardware and network delays, time-varying computation powers, and any poss
ible fluctuations and trends of computation speeds. These real-world scenarios a
re formalized by our new universal computation model. Leveraging this model and
new proof techniques, we discover tight lower bounds that apply to virtually all
 synchronous and asynchronous methods, including Minibatch SGD, Asynchronous SGD
 (Recht et al., 2011), and Picky SGD (Cohen et al., 2021). We show that these lo
wer bounds, up to constant factors, are matched by the optimal Rennala SGD and M
alenia SGD methods (Tyurin & Richtárik, 2023).
****
Boris Knyazev,Abhinav Moudgil,Guillaume Lajoie,Eugene Belilovsky,Simon Lacoste-J
ulien
Accelerating Training with Neuron Interaction and Nowcasting Networks
Neural network training can be accelerated when a learnable update rule is used
in lieu of classic adaptive optimizers (e.g. Adam). However, learnable update ru
les can be costly and unstable to train and use. Recently, Jang et al. (2023) pr
oposed a simpler approach to accelerate training based on weight nowcaster netwo

rks (WNNs). In their approach, Adam is used for most of the optimization steps a
nd periodically, only every few steps, a WNN nowcasts (predicts near future) par
ameters. We improve WNNs by proposing neuron interaction and nowcasting (NiNo) n
etworks. In contrast to WNNs, NiNo leverages neuron connectivity and graph neura
l networks to more accurately nowcast parameters. We further show that in some n
etworks, such as Transformers, modeling neuron connectivity accurately is challe
nging. We address this and other limitations, which allows NiNo to accelerate Ad
am training by up to 50% in vision and language tasks.
****
Haofei Lu,Zhe Wu,Junliang Xing,Jianshu Li,Ruoyu Li,Zhe Li,Yuanchun Shi
BodyGen: Advancing Towards Efficient Embodiment Co-Design
Embodiment co-design aims to optimize a robot's morphology and control policy si
multaneously.
While prior work has demonstrated its potential for generating environment-adapt
ive robots, this field still faces persistent challenges in optimization efficie
ncy due to the (i) combinatorial nature of morphological search spaces and (ii)
intricate dependencies between morphology and control.
We prove that the ineffective morphology representation and unbalanced reward si
gnals between the design and control stages are key obstacles to efficiency.
To advance towards efficient embodiment co-design, we propose **BodyGen**, which
 utilizes (1) topology-aware self-attention for both design and control, enablin
g efficient morphology representation with lightweight model sizes; (2) a tempor
al credit assignment mechanism that ensures balanced reward signals for optimiza
tion. With our findings, BodyGen achieves an average **60.03%** performance impr
ovement against state-of-the-art baselines. We provide codes and more results on
 the website: https://genesisorigin.github.io.
****
Sifan Wang,Jacob H Seidman,Shyam Sankaran,Hanwen Wang,George J. Pappas,Paris Per
dikaris
CViT: Continuous Vision Transformer for Operator Learning
Operator learning, which aims to approximate maps between infinite-dimensional f
unction spaces, is an important area in scientific machine learning with applica
tions across various physical domains. Here we introduce the Continuous Vision T
ransformer (CViT), a novel neural operator architecture that leverages advances
in computer vision to address challenges in learning complex physical systems.
CViT combines a vision transformer encoder, a novel grid-based coordinate embedd
ing, and a query-wise cross-attention mechanism to effectively capture multi-sca
le dependencies. This design allows for flexible output representations and cons
istent evaluation at arbitrary resolutions. We demonstrate CViT's effectiveness
across a diverse range of partial differential equation (PDE) systems, including
 fluid dynamics, climate modeling, and reaction-diffusion processes. Our compreh
ensive experiments show that CViT achieves state-of-the-art performance on multi
ple benchmarks, often surpassing larger foundation models, even without extensiv
e pretraining and roll-out fine-tuning. Taken together, CViT exhibits robust han
dling of discontinuous solutions, multi-scale features, and intricate spatio-tem
poral dynamics. Our contributions can be viewed as a significant step towards ad
apting advanced computer vision architectures for building more flexible and acc
urate machine learning models in the physical sciences.
****
Guanting Dong,Keming Lu,Chengpeng Li,Tingyu Xia,Bowen Yu,Chang Zhou,Jingren Zhou
Self-play with Execution Feedback: Improving Instruction-following Capabilities
of Large Language Models
One core capability of large language models~(LLMs) is to follow natural languag
e instructions. However, the issue of automatically constructing high-quality tr
aining data to enhance the complex instruction-following abilities of LLMs witho
ut manual annotation remains unresolved. In this paper, we introduce AutoIF, the
 first scalable and reliable method for automatically generating instruction-fol
lowing training data. AutoIF transforms the validation of instruction-following
data quality into code verification, requiring LLMs to generate instructions, th
e corresponding code to verify the correctness of the instruction responses, and

unit test samples to cross-validate the code's correctness. Then, execution feedback-based rejection sampling can generate data for Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) training. AutoIF achieves significant improvements across three training algorithms, SFT, Offline DPO, and Online DPO, when applied to the advanced open-source LLMs, Qwen2 and LLaMA3, in self-alignment and strong-to-weak distillation settings. Using two widely-used and three challenging general instruction-following benchmarks, we demonstrate that AutoIF significantly improves LLM performance across a wide range of natural instruction constraints. Notably, AutoIF is the first to surpass 90\% accuracy in IFEval’s loose instruction accuracy, without compromising general, math and coding capabilities. Further analysis of quality, scaling, combination, and data efficiency highlights AutoIF's strong generalization and alignment potential. Our code are available at https://github.com/QwenLM/AutoIF
****

Alexander DeRieux,Walid Saad

eQMARL: Entangled Quantum Multi-Agent Reinforcement Learning for Distributed Cooperation over Quantum Channels

Collaboration is a key challenge in distributed multi-agent reinforcement learning (MARL) environments. Learning frameworks for these decentralized systems must weigh the benefits of explicit player coordination against the communication overhead and computational cost of sharing local observations and environmental data. Quantum computing has sparked a potential synergy between quantum entanglement and cooperation in multi-agent environments, which could enable more efficient distributed collaboration with minimal information sharing. This relationship is largely unexplored, however, as current state-of-the-art quantum MARL (QMARL) implementations rely on classical information sharing rather than entanglement over a quantum channel as a coordination medium. In contrast, in this paper, a novel framework dubbed entangled QMARL (eQMARL) is proposed. The proposed eQMARL is a distributed actor-critic framework that facilitates cooperation over a quantum channel and eliminates local observation sharing via a quantum entangled split critic. Introducing a quantum critic uniquely spread across the agents allows coupling of local observation encoders through entangled input qubits over a quantum channel, which requires no explicit sharing of local observations and reduces classical communication overhead. Further, agent policies are tuned through joint observation-value function estimation via joint quantum measurements, thereby reducing the centralized computational burden. Experimental results show that eQMARL with Ψ^{+} entanglement converges to a cooperative strategy up to $17.8\\%$ faster and with a higher overall score compared to split classical and fully centralized classical and quantum baselines. The results also show that eQMARL achieves this performance with a constant factor of 25-times fewer centralized parameters compared to the split classical baseline.
****

Khyathi Chandu,Linjie Li,Anas Awadalla,Ximing Lu,Jae Sung Park,Jack Hessel,Lijuan Wang,Yejin Choi

CertainlyUncertain: A Benchmark and Metric for Multimodal Epistemic and Aleatoric Awareness

The ability to acknowledge the inevitable uncertainty in their knowledge and reasoning is a prerequisite for AI systems to be truly truthful and reliable. In this paper, we present a taxonomy of uncertainty specific to vision-language AI systems, distinguishing between epistemic uncertainty (arising from a lack of information) and aleatoric uncertainty (due to inherent unpredictability), and further explore finer categories within. Based on this taxonomy, we synthesize a benchmark dataset, CertainlyUncertain, featuring 178K visual question answering (VQA) samples as contrastive pairs. This is achieved by 1) inpainting images to make previously answerable questions into unanswerable ones; and 2) using image captions to prompt large language models for both answerable and unanswerable questions. Additionally, we introduce a new metric confidence-weighted accuracy, that is well correlated with both accuracy and calibration error, to address the shortcomings of existing metrics. Despite the recent rapid progress in vision-language models (VLMs), evaluations on our benchmark show that they perform poorly in

uncertain scenarios. Further experiments demonstrate that supervised fine-tuning with CertainlyUncertain enhances the performance of VLMs, and reduces the calibration error. These improvements extend beyond our benchmark to existing refusal-oriented datasets and show positive results on reducing hallucinations, while maintaining performance on standard VQA benchmarks. Our work underscores the importance of addressing uncertainty in vision-language AI systems to improve their reliability and trustworthiness in real-world applications.
****
Yanming Liu,Xinyue Peng,Jiannan Cao,Shi Bo,Yanxin Shen,Tianyu Du,Sheng Cheng,Xun Wang,Jianwei Yin,Xuhong Zhang
Bridging Context Gaps: Leveraging Coreference Resolution for Long Contextual Understanding
Large language models (LLMs) have shown remarkable capabilities in natural language processing; however, they still face difficulties when tasked with understanding lengthy contexts and executing effective question answering. These challenges often arise due to the complexity and ambiguity present in longer texts. To enhance the performance of LLMs in such scenarios, we introduce the Long Question Coreference Adaptation (LQCA) method. This innovative framework focuses on coreference resolution tailored to long contexts, allowing the model to identify and manage references effectively. The LQCA method encompasses four key steps: resolving coreferences within sub-documents, computing the distances between mentions, defining a representative mention for coreference, and answering questions through mention replacement. By processing information systematically, the framework provides easier-to-handle partitions for LLMs, promoting better understanding. Experimental evaluations on a range of LLMs and datasets have yielded positive results, with a notable improvements on OpenAI-o1-mini and GPT-4o models, highlighting the effectiveness of leveraging coreference resolution to bridge context gaps in question answering. Our code is public at https://github.com/OceannTwT/LQCA.
****
Peng Xu,Wei Ping,Xianchao Wu,Chejian Xu,Zihan Liu,Mohammad Shoeybi,Bryan Catanzaro
ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities
In this work, we introduce ChatQA 2, an Llama 3.0-based model with a 128K
context window, designed to bridge the gap between open-source LLMs and
leading proprietary models (e.g., GPT-4-Turbo-2024-04-09) in long context un-
derstanding and retrieval-augmented generation (RAG) capabilities. These two
capabilities are complementary to each other and essential for LLMs to process
large volumes of information that cannot fit into a single prompt. We present
a detailed continued training recipe to extend the context window of Llama3-
70B-base from 8K to 128K tokens, along with a three-stage instruction tun-
ing process to enhance the model’s instruction-following, RAG performance,
and long-context understanding capabilities. Our results demonstrate that the
Llama3-ChatQA-2-70B model outperforms most existing state-of-the-art models,
including GPT-4-Turbo-2024-04-09, Qwen2-72B-Instruct, and Llama3.1-70B-
Instruct, on ultra-long tasks beyond 100K tokens, as well as on the RAG benchmark
using only a 4K context window, showing the strong long context capability across
varying sequence lengths. We further provide extensive comparisons between
direct long-context and RAG solutions using the same state-of-the-art long-context
LLMs. Interestingly, we find that the performance of strong long-context LLMs
using RAG improves when retrieving a larger number of chunks. With a large set
of top-k chunks, RAG consistently outperforms direct long-context solution using
the same state-of-the-art long-context models (e.g., Llama3-ChatQA-2-70B and
Qwen2-72B-Instruct) on both 32K and 128K benchmarks. We open-source the
model weights, training data, and the evaluation setup for the for the community:

https://chatqa2-project.github.io/
****
Amin Nejatbakhsh,Victor Geadah,Alex H Williams,David Lipshutz
Comparing noisy neural population dynamics using optimal transport distances
Biological and artificial neural systems form high-dimensional neural representa
tions that underpin their computational capabilities. Methods for quantifying ge
ometric similarity in neural representations have become a popular tool for iden
tifying computational principles that are potentially shared across neural syste
ms. These methods generally assume that neural responses are deterministic and s
tatic. However, responses of biological systems, and some artificial systems, ar
e noisy and dynamically unfold over time. Furthermore, these characteristics can
 have substantial influence on a system’s computational capabilities. Here, we d
emonstrate that existing metrics can fail to capture key differences between neu
ral systems with noisy dynamic responses. We then propose a metric for comparing
 the geometry of noisy neural trajectories, which can be derived as an optimal t
ransport distance between Gaussian processes. We use the metric to compare model
s of neural responses in different regions of the motor system and to compare th
e dynamics of latent diffusion models for text-to-image synthesis.
****
Hoin Jung,Junyi Chai,Xiaoqian Wang
Adversarial Latent Feature Augmentation for Fairness
Achieving fairness in machine learning remains a critical challenge, especially
due to the opaque effects of data augmentation on input spaces within nonlinear
neural networks. Nevertheless, current approaches that emphasize augmenting late
nt features, rather than input spaces, offer limited insights into their ability
 to detect and mitigate bias. In response, we introduce the concept of the "unfa
ir region" in the latent space, a subspace that highlights areas where misclassi
fication rates for certain demographic groups are disproportionately high, leadi
ng to unfair prediction results. To address this, we propose Adversarial Latent
Feature Augmentation (ALFA), a method that leverages adversarial fairness attack
s to perturb latent space features, which are then used as data augmentation for
 fine-tuning. ALFA intentionally shifts latent features into unfair regions, and
 the last layer of the network is fine-tuned with these perturbed features, lead
ing to a corrected decision boundary that enhances fairness in classification in
 a cost-effective manner. We present a theoretical framework demonstrating that
our adversarial fairness objective reliably generates biased feature perturbatio
ns, and that fine-tuning on samples from these unfair regions ensures fairness i
mprovements. Extensive experiments across diverse datasets, modalities, and back
bone networks validate that training with these adversarial features significant
ly enhances fairness while maintaining predictive accuracy in classification tas
ks.
****
Junghun Oh,Sungyong Baik,Kyoung Mu Lee
Find A Winning Sign: Sign Is All We Need to Win the Lottery
The Lottery Ticket Hypothesis (LTH) posits the existence of a sparse subnetwork
(a.k.a. winning ticket) that can generalize comparably to its over-parameterized
 counterpart when trained from scratch.
The common approach to finding a winning ticket is to preserve the original stro
ng generalization through Iterative Pruning (IP) and transfer information useful
 for achieving the learned generalization by applying the resulting sparse mask
to an untrained network.
However, existing IP methods still struggle to generalize their observations bey
ond ad-hoc initialization and small-scale architectures or datasets, or they byp
ass these challenges by applying their mask to trained weights instead of initia
lized ones.
In this paper, we demonstrate that the parameter sign configuration plays a cruc
ial role in conveying useful information for generalization to any randomly init
ialized network.
Through linear mode connectivity analysis, we observe that a sparse network trai
ned by an existing IP method can retain its basin of attraction if its parameter

signs and normalization layer parameters are preserved.
To take a step closer to finding a winning ticket, we alleviate the reliance on normalization layer parameters by preventing high error barriers along the linear path between the sparse network trained by our method and its counterpart with initialized normalization layer parameters.
Interestingly, across various architectures and datasets, we observe that any randomly initialized network can be optimized to exhibit low error barriers along the linear path to the sparse network trained by our method by inheriting its sparsity and parameter sign information, potentially achieving performance comparable to the original.
The code is available at https://github.com/JungHunOh/AWS_ICLR2025.git.
****

Kexun Zhang,Weiran Yao,Zuxin Liu,Yihao Feng,Zhiwei Liu,Rithesh R N,Tian Lan,Lei Li,Renze Lou,Jiacheng Xu,Bo Pang,Yingbo Zhou,Shelby Heinecke,Silvio Savarese,Huan Wang,Caiming Xiong
Diversity Empowers Intelligence: Integrating Expertise of Software Engineering Agents
Large language model (LLM) agents have shown great potential in solving real-world software engineering (SWE) problems. The most advanced open-source SWE agent can resolve over 27% of real GitHub issues in SWE-Bench Lite. However, these sophisticated agent frameworks exhibit varying strengths, excelling in certain tasks while underperforming in others. To fully harness the diversity of these agents, we propose DEI (Diversity Empowered Intelligence), a framework that leverages their unique expertise. DEI functions as a meta-module atop existing SWE agent frameworks, managing agent collectives for enhanced problem-solving. Experimental results show that a DEI-guided committee of agents is able to surpass the best individual agent's performance by a large margin. For instance, a group of open-source SWE agents, with a maximum individual resolve rate of 27.3% on SWE-Bench Lite, can achieve a 34.3% resolve rate with DEI, making a 25% improvement and beating most closed-source solutions. Our best-performing group excels with a 55% resolve rate, securing the highest ranking on SWE-Bench Lite. Our findings contribute to the growing body of research on collaborative AI systems and their potential to solve complex software engineering challenges.
****

Nadav Timor,Jonathan Mamou,Daniel Korat,Moshe Berchansky,Oren Pereg,Moshe Wasserblat,Tomer Galanti,Michal Gordon-Kiwkowitz,David Harel
Distributed Speculative Inference (DSI): Speculation Parallelism for Provably Faster Lossless Language Model Inference
This paper introduces *distributed speculative inference (DSI)*, a novel inference algorithm that is provably faster than speculative inference (SI) [leviathan2023, chen2023, miao2024, sun2025, timor2025] and standard autoregressive inference (non-SI). Like other SI algorithms, DSI operates on frozen language models (LMs), requiring no training or architectural modifications, and it preserves the target distribution. Prior studies on SI have demonstrated empirical speedups over non-SI—but rely on sufficiently fast and accurate drafters, which are often unavailable in practice. We identify a gap where SI can be slower than non-SI if drafters are too slow or inaccurate. We close this gap by proving that DSI is faster than both SI and non-SI—given any drafters. DSI is therefore not only faster than SI, but also unlocks the acceleration of LMs for which SI fails. DSI leverages *speculation parallelism (SP)*, a novel type of task parallelism, to orchestrate target and drafter instances that overlap in time, establishing a new foundational tradeoff between computational resources and latency. Our simulations show that DSI is 1.29-1.92x faster than SI in single-node setups for various off-the-shelf LMs and tasks. We open-source all our code.
****

Chenhang Cui,An Zhang,Yiyang Zhou,Zhaorun Chen,Gelei Deng,Huaxiu Yao,Tat-Seng Chua
Fine-Grained Verifiers: Preference Modeling as Next-token Prediction in Vision-Language Alignment
The recent advancements in large language models (LLMs) and pre-trained vision m

odels have accelerated the development of vision-language large models (VLLMs), enhancing the interaction between visual and linguistic modalities. Despite thei r notable success across various domains, VLLMs face challenges in modality alig nment, which can lead to issues like hallucinations and unsafe content generatio n. Current alignment techniques often rely on coarse feedback and external datas ets, limiting scalability and performance. In this paper, we propose FiSAO (Fine -Grained Self-Alignment Optimization), a novel self-alignment method that utiliz es the model’s own visual encoder as a fine-grained verifier to improve vision-l anguage alignment without the need for additional data. By leveraging token-leve l feedback from the vision encoder, FiSAO significantly improves vision-language alignment, even surpassing traditional preference tuning methods that require a dditional data. Through both theoretical analysis and experimental validation, w e demonstrate that FiSAO effectively addresses the misalignment problem in VLLMs , marking the first instance of token-level rewards being applied to such models . Our code is avaliable at \url{https://anonymous.4open.science/r/FISAO-57F0/}.
****
RISHI HAZRA,Alkis Sygkounas,Andreas Persson,Amy Loutfi,Pedro Zuidberg Dos Martir es
REvolve: Reward Evolution with Large Language Models using Human Feedback
Designing effective reward functions is crucial to training reinforcement learni ng (RL) algorithms. However, this design is non-trivial, even for domain experts , due to the subjective nature of certain tasks that are hard to quantify explic itly. In recent works, large language models (LLMs) have been used for reward ge neration from natural language task descriptions, leveraging their extensive ins truction tuning and commonsense understanding of human behavior. In this work, w e hypothesize that LLMs, guided by human feedback, can be used to formulate rewa rd functions that reflect human implicit knowledge. We study this in three chall enging settings -- autonomous driving, humanoid locomotion, and dexterous manipu lation -- wherein notions of ``good" behavior are tacit and hard to quantify. To this end, we introduce REvolve, a truly evolutionary framework that uses LLMs f or reward design in RL. REvolve generates and refines reward functions by utiliz ing human feedback to guide the evolution process, effectively translating impli cit human knowledge into explicit reward functions for training (deep) RL agents . Experimentally, we demonstrate that agents trained on REvolve-designed rewards outperform other state-of-the-art baselines.
****
Sungwon Kim,Yoonho Lee,Yunhak Oh,Namkyeong Lee,Sukwon Yun,Junseok Lee,Sein Kim,C arl Yang,Chanyoung Park
Subgraph Federated Learning for Local Generalization
Federated Learning (FL) on graphs enables collaborative model training to enhanc e performance without compromising the privacy of each client. However, existing methods often overlook the mutable nature of graph data, which frequently intro duces new nodes and leads to shifts in label distribution. Since they focus sole ly on performing well on each client's local data, they are prone to overfitting to their local distributions (i.e., local overfitting), which hinders their abi lity to generalize to unseen data with diverse label distributions. In contrast, our proposed method, FedLoG, effectively tackles this issue by mitigating local overfitting. Our model generates global synthetic data by condensing the reliab le information from each class representation and its structural information acr oss clients. Using these synthetic data as a training set, we alleviate the loca l overfitting problem by adaptively generalizing the absent knowledge within eac h local dataset. This enhances the generalization capabilities of local models, enabling them to handle unseen data effectively. Our model outperforms baselines in our proposed experimental settings, which are designed to measure generaliza tion power to unseen data in practical scenarios.
Our code is available at https://github.com/sung-won-kim/FedLoG
****
Guangxuan Xiao,Jiaming Tang,Jingwei Zuo,junxian guo,Shang Yang,Haotian Tang,Yao Fu,Song Han
DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming

Heads
Deploying long-context large language models (LLMs) is essential but poses signi
ficant computational and memory challenges.
Caching all Key and Value (KV) states across all attention heads consumes substa
ntial memory.
Existing KV cache pruning methods either damage the long-context capabilities of
 LLMs or offer only limited efficiency improvements.
In this paper, we identify that only a fraction of attention heads, a.k.a, Retri
eval Heads, are critical for processing long contexts and require full attention
 across all tokens.
In contrast, all other heads, which primarily focus on recent tokens and attenti
on sinks—referred to as Streaming Heads—do not require full attention.
Based on this insight, we introduce DuoAttention, a framework that only applies
a full KV cache to retrieval heads while using a light-weight, constant-length K
V cache for streaming heads, which reduces both LLM's decoding and pre-filling m
emory and latency without compromising its long-context abilities.
DuoAttention uses a lightweight, optimization-based algorithm with synthetic dat
a to identify retrieval heads accurately.
Our method significantly reduces long-context inference memory by up to 2.55$\ti
mes$ for MHA and 1.67$\times$ for GQA models while speeding up decoding by up to
 2.18$\times$ and 1.50$\times$ and accelerating pre-filling by up to 1.73$\times
$ and 1.63$\times$ for MHA and GQA models, respectively, with minimal accuracy l
oss compared to full attention.
Notably, combined with quantization, DuoAttention enables Llama-3-8B decoding wi
th 3.33 million context length measured on a single A100 GPU. Code is provided i
n https://github.com/mit-han-lab/duo-attention.
****
Tao Liu,Kai Wang,Senmao Li,Joost van de Weijer,Fahad Shahbaz Khan,Shiqi Yang,Yax
ing Wang,Jian Yang,Ming-Ming Cheng
One-Prompt-One-Story: Free-Lunch Consistent Text-to-Image Generation Using a Sin
gle Prompt
Text-to-image generation models can create high-quality images from input prompt
s. However, they struggle to support the consistent generation of identity-prese
rving requirements for storytelling. Existing approaches to this problem typical
ly require extensive training in large datasets or additional modifications to t
he original model architectures. This limits their applicability across differen
t domains and diverse diffusion model configurations. In this paper, we first ob
serve the inherent capability of language models, coined $\textit{context consis
tency}$, to comprehend identity through context with a single prompt. Drawing in
spiration from the inherent $\textit{context consistency}$, we propose a novel $
\textit{training-free}$ method for consistent text-to-image (T2I) generation, te
rmed "One-Prompt-One-Story" ($\textit{1Prompt1Story}$). Our approach $\textit{1P
rompt1Story}$ concatenates all prompts into a single input for T2I diffusion mod
els, initially preserving character identities. We then refine the generation pr
ocess using two novel techniques: $\textit{Singular-Value
Reweighting}$ and $\textit{Identity-Preserving Cross-Attention}$, ensuring bette
r alignment with the input description for each frame. In our experiments, we co
mpare our method against various existing consistent T2I generation approaches t
o demonstrate its effectiveness, through quantitative metrics and qualitative as
sessments. Code is available at https://github.com/byliutao/1Prompt1Story.
****
Jack Brady,Julius von Kügelgen,Sebastien Lachapelle,Simon Buchholz,Thomas Kipf,W
ieland Brendel
Interaction Asymmetry: A General Principle for Learning Composable Abstractions
Learning disentangled representations of concepts and re-composing them in unsee
n ways is crucial for generalizing to out-of-domain situations. However, the und
erlying properties of concepts that enable such disentanglement and compositiona
l generalization remain poorly understood. In this work, we propose the principl
e of interaction asymmetry which states: "Parts of the same concept have more co
mplex interactions than parts of different concepts". We formalize this via bloc

k diagonality conditions on the $(n+1)$th order derivatives of the generator map
ping concepts to observed data, where different orders of "complexity" correspon
d to different n. Using this formalism, we prove that interaction asymmetry en
ables both disentanglement and compositional generalization. Our results unify r
ecent theoretical results for learning concepts of objects, which we show are re
covered as special cases with $n=0$ or 1. We provide results for up to $n=2$,
thus extending these prior works to more flexible generator functions, and conje
cture that the same proof strategies generalize to larger n. Practically, our
theory suggests that, to disentangle concepts, an autoencoder should penalize it
s latent capacity and the interactions between concepts during decoding. We prop
ose an implementation of these criteria using a flexible Transformer-based VAE,
with a novel regularizer on the attention weights of the decoder. On synthetic i
mage datasets consisting of objects, we provide evidence that this model can ach
ieve comparable object disentanglement to existing models that use more explicit
 object-centric priors.
****
Sachit Gaudi,Gautam Sreekumar,Vishnu Boddeti
CoInD: Enabling Logical Compositions in Diffusion Models
How can we learn generative models to sample data with arbitrary logical composi
tions of statistically independent attributes? The prevailing solution is to sam
ple from distributions expressed as a composition of attributes' conditional mar
ginal distributions under the assumption that they are statistically independent
. This paper shows that standard conditional diffusion models violate this assum
ption, even when all attribute compositions are observed during training. And, t
his violation is significantly more severe when only a subset of the composition
s is observed. We propose CoInD to address this problem. It explicitly enforces
statistical independence between the conditional marginal distributions by minim
izing Fisher’s divergence between the joint and marginal distributions. The theo
retical advantages of CoInD are reflected in both qualitative and quantitative e
xperiments, demonstrating a significantly more faithful and controlled generatio
n of samples for arbitrary logical compositions of attributes. The benefit is mo
re pronounced for scenarios that current solutions relying on the assumption of
conditionally independent marginals struggle with, namely, logical compositions
involving the NOT operation and when only a subset of compositions are observed
during training.
****
Ge Ya Luo,Gian Mario Favero,ZhiHao Luo,Alexia Jolicoeur-Martineau,Christopher Pa
l
Beyond FVD: An Enhanced Evaluation Metrics for Video Generation Distribution Qua
lity
The Fréchet Video Distance (FVD) is a widely adopted metric for evaluating video
 generation distribution quality. However, its effectiveness relies on critical
assumptions. Our analysis reveals three significant limitations: (1) the non-Gau
ssianity of the Inflated 3D Convnet (I3D) feature space; (2) the insensitivity o
f I3D features to temporal distortions; (3) the impractical sample sizes require
d for reliable estimation. These findings undermine FVD's reliability and show t
hat FVD falls short as a standalone metric for video generation evaluation. Afte
r extensive analysis of a wide range of metrics and backbone architectures, we p
ropose JEDi, the JEPA Embedding Distance,
based on features derived from a Joint Embedding Predictive Architecture, measur
ed using Maximum Mean Discrepancy with polynomial kernel. Our experiments on mu
ltiple open-source datasets show clear evidence that it is a superior alternativ
e to the widely used FVD metric, requiring only 16% of the samples to reach its
steady value, while increasing alignment with human evaluation by 34%, on averag
e.

Project page: https://oooolga.github.io/JEDi.github.io/.
****
Jeffrey Gu,Serena Yeung-Levy
Foundation Models Secretly Understand Neural Network Weights: Enhancing Hypernet

work Architectures with Foundation Models
Large pre-trained models, or foundation models, have shown impressive performanc e when adapted to a variety of downstream tasks, often out-performing specialize d models. Hypernetworks, neural networks that generate some or all of the parame ters of another neural network, have become an increasingly important technique for conditioning and generalizing implicit neural representations (INRs), which represent signals or objects such as audio or 3D shapes using a neural network. However, despite the potential benefits of incorporating foundation models in hy pernetwork methods, this research direction has not been investigated, likely du e to the dissimilarity of the weight generation task with other visual tasks. To address this gap, we (1) show how foundation models can improve hypernetworks w ith Transformer-based architectures, (2) provide an empirical analysis of the be nefits of foundation models for hypernetworks through the lens of the generaliza ble INR task, showing that leveraging foundation models improves performance, ge neralizability, and data efficiency across a variety of algorithms and modalitie s. We also provide further analysis in examining the design space of foundation model-based hypernetworks, including examining the choice of foundation models, algorithms, and the effect of scaling foundation models.
****
Chuan Liu,Chunshu Wu,shihui cao,Mingkai Chen,James Chenhao Liang,Ang Li,Michael Huang,Chuang Ren,Ying Nian Wu,Dongfang Liu,Tong Geng
Diff-PIC: Revolutionizing Particle-In-Cell Nuclear Fusion Simulation with Diffus ion Models
The rapid development of AI highlights the pressing need for sustainable energy, a critical global challenge for decades. Nuclear fusion, generally seen as a pr omising solution, has been the focus of intensive research for nearly a century, with investments reaching hundreds of billions of dollars. Recent advancements in Inertial Confinement Fusion (ICF) have drawn significant attention to fusion research, in which Laser-Plasma Interaction (LPI) is critical for ensuring fusio n stability and efficiency. However, the complexity of LPI makes analytical appr oaches impractical, leaving researchers dependent on extremely computationally i ntensive Particle-in-Cell (PIC) simulations to generate data, posing a significa nt bottleneck to the advancement of fusion research. In response, this work intr oduces Diff-PIC, a novel framework that leverages conditional diffusion models a s a computationally efficient alternative to PIC simulations for generating high -fidelity scientific LPI data. In this work, physical patterns captured by PIC s imulations are distilled into diffusion models associated with two tailored enha ncements: (1) To effectively capture the complex relationships between physical parameters and their corresponding outcomes, the parameters are encoded in a phy sically informed manner. (2) To further enhance efficiency while maintaining phy sical validity, the rectified flow technique is employed to transform our model into a one-step conditional diffusion model. Experimental results show that Diff -PIC achieves a $\sim$16,200$\times$ speedup compared to traditional PIC on a 10 0 picosecond simulation, while delivering superior accuracy compared to other da ta generation approaches.
****
Depen Morwani,Itai Shapira,Nikhil Vyas,eran malach,Sham M. Kakade,Lucas Janson
A New Perspective on Shampoo's Preconditioner
Shampoo, a second-order optimization algorithm that uses a Kronecker product pre conditioner, has recently received increasing attention from the machine learnin g community. Despite the increasing popularity of Shampoo, the theoretical found ations of its effectiveness are not well understood. The preconditioner used by Shampoo can be viewed as either an approximation of the Gauss--Newton component of the Hessian or the covariance matrix of the gradients maintained by Adagrad. Our key contribution is providing an explicit and novel connection between the o ptimal Kronecker product approximation of these matrices and the approximation made by Shampoo. Our connection highlights a subtle but common misconception abo ut Shampoo’s approximation. In particular, the square of the approximation used by the Shampoo optimizer is equivalent to a single step of the power iteration algorithm for computing the aforementioned optimal Kronecker product a

pproximation. Across a variety of datasets and architectures we empirically demonstrate that this is close to the optimal Kronecker product approximation. We also study the impact of batch gradients and empirical Fisher on the quality of Hessian approximation. Our findings not only advance the theoretical understanding of Shampoo but also illuminate potential pathways for enhancing its practical performance.
****
Siddharth Joshi,Jiayi Ni,Baharan Mirzasoleiman
Dataset Distillation via Knowledge Distillation: Towards Efficient Self-Supervised Pre-training of Deep Networks
Dataset distillation (DD) generates small synthetic datasets that can efficiently train deep networks with a limited amount of memory and compute. Despite the success of DD methods for supervised learning, DD for self-supervised pre-training of deep models has remained unaddressed. Pre-training on unlabeled data is crucial for efficiently generalizing to downstream tasks with limited labeled data. In this work, we propose the first effective DD method for SSL pre-training. First, we show, theoretically and empirically, that naiive application of supervised DD methods to SSL fails, due to the high variance of the SSL gradient. Then, we address this issue by relying on insights from knowledge distillation (KD) literature. Specifically, we train a small student model to match the representations of a larger teacher model trained with SSL. Then, we generate a small synthetic dataset by matching the training trajectories of the student models. As the KD objective has considerably lower variance than SSL, our approach can generate synthetic datasets that can successfully pre-train high-quality encoders. Through extensive experiments, we show that our distilled sets lead to up to 13% higher accuracy than prior work, on a variety of downstream tasks, in the presence of limited labeled data. Code at https://github.com/BigML-CS-UCLA/MKDT.
****
Xingzhuo Guo,Yu Zhang,Baixu Chen,Haoran Xu,Jianmin Wang,Mingsheng Long
Dynamical Diffusion: Learning Temporal Dynamics with Diffusion Models
Diffusion models have emerged as powerful generative frameworks by progressively adding noise to data through a forward process and then reversing this process to generate realistic samples. While these models have achieved strong performance across various tasks and modalities, their application to temporal predictive learning remains underexplored. Existing approaches treat predictive learning as a conditional generation problem, but often fail to fully exploit the temporal dynamics inherent in the data, leading to challenges in generating temporally coherent sequences. To address this, we introduce Dynamical Diffusion (DyDiff), a theoretically sound framework that incorporates temporally aware forward and reverse processes. Dynamical Diffusion explicitly models temporal transitions at each diffusion step, establishing dependencies on preceding states to better capture temporal dynamics. Through the reparameterization trick, Dynamical Diffusion achieves efficient training and inference similar to any standard diffusion model. Extensive experiments across scientific spatiotemporal forecasting, video prediction, and time series forecasting demonstrate that Dynamical Diffusion consistently improves performance in temporal predictive tasks, filling a crucial gap in existing methodologies. Code is available at this repository: https://github.com/thuml/dynamical-diffusion.
****
Jerry Yao-Chieh Hu,Weimin Wu,Yi-Chen Lee,Yu-Chao Huang,Minshuo Chen,Han Liu
On Statistical Rates of Conditional Diffusion Transformers: Approximation, Estimation and Minimax Optimality
We investigate the approximation and estimation rates of conditional diffusion transformers (DiTs) with classifier-free guidance. We present a comprehensive analysis for “in-context” conditional DiTs under various common assumptions: generic and strong Hölder, linear latent (subspace), and Lipschitz score function assumptions. Importantly, we establish minimax optimality of DiTs by leveraging score function regularity. Specifically, we discretize the input domains into infinitesimal grids and then perform term-by-term Taylor expansions on the conditional diffusion score function under the Hölder smooth data assumption. This enables

fine-grained use of transformers' universal approximation through a more detaile
d piecewise constant approximation, and hence obtains tighter bounds. Additional
ly, we extend our analysis to latent settings. Our findings establish statistica
l limits for DiTs and offer practical guidance toward more efficient and accurat
e designs.
****
Yuanchao Wang,Zhao-Rong Lai,Tianqi Zhong
Out-of-distribution Generalization for Total Variation based Invariant Risk Mini
mization
Invariant risk minimization is an important general machine learning framework t
hat has recently been interpreted as a total variation model (IRM-TV). However,
how to improve out-of-distribution (OOD) generalization in the IRM-TV setting re
mains unsolved. In this paper, we extend IRM-TV to a Lagrangian multiplier model
 named OOD-TV-IRM. We find that the autonomous TV penalty hyperparameter is exac
tly the Lagrangian multiplier. Thus OOD-TV-IRM is essentially a primal-dual opti
mization model, where the primal optimization minimizes the entire invariant ris
k and the dual optimization strengthens the TV penalty. The objective is to reac
h a semi-Nash equilibrium where the balance between the training loss and OOD ge
neralization is maintained. We also develop a convergent primal-dual algorithm t
hat facilitates an adversarial learning scheme. Experimental results show that O
OD-TV-IRM outperforms IRM-TV in most situations.
****
Xinyu Wang,Bohan Zhuang,Qi Wu
Are Large Vision Language Models Good Game Players?
Large Vision Language Models (LVLMs) have demonstrated remarkable abilities in u
nderstanding and reasoning about both visual and textual information. However, e
xisting evaluation methods for LVLMs, primarily based on benchmarks like Visual
Question Answering and image captioning, often fail to capture the full scope of
 LVLMs' capabilities. These benchmarks are limited by issues such as inadequate
assessment of detailed visual perception, data contamination, and a lack of focu
s on multi-turn reasoning. To address these challenges, we propose LVLM-Playgrou
nd, a game-based evaluation framework designed to provide a comprehensive assess
ment of LVLMs' cognitive and reasoning skills in structured environments. LVLM-P
layground uses a set of games to evaluate LVLMs on four core tasks: Perceiving,
Question Answering, Rule Following, and End-to-End Playing, with each target tas
k designed to assess specific abilities, including visual perception, reasoning,
 decision-making, etc. Based on this framework, we conduct extensive experiments
 that explore the limitations of current LVLMs, such as handling long structured
 outputs and perceiving detailed and dense elements. Code and data are publicly
available at https://github.com/xinke-wang/LVLM-Playground.
****
Jiahao Wu,Rui Peng,Zhiyan Wang,Lu Xiao,Luyang Tang,Jinbo Yan,Kaiqiang Xiong,Rong
gang Wang
Swift4D: Adaptive divide-and-conquer Gaussian Splatting for compact and efficien
t reconstruction of dynamic scene
Novel view synthesis has long been a practical but challenging task, although th
e introduction of numerous methods to solve this problem, even combining advance
d representations like 3D Gaussian Splatting, they still struggle to recover hig
h-quality results and often consume too much storage memory and training time.
In this paper we propose Swift4D, a divide-and-conquer 3D Gaussian Splatting met
hod that can handle static and dynamic primitives separately, achieving a good t
rade-off between rendering quality and efficiency, motivated by the fact that mo
st of the scene is the static primitive and does not require additional dynamic
properties. Concretely, we focus on modeling dynamic transformations only for th
e dynamic primitives which benefits both efficiency and quality. We first employ
 a learnable decomposition strategy to separate the primitives, which relies on
an additional parameter to classify primitives as static or dynamic. For the dyn
amic primitives, we employ a compact multi-resolution 4D Hash mapper to transfor
m these primitives from canonical space into deformation space at each timestamp
, and then mix the static and dynamic primitives to produce the final output. Th

is divide-and-conquer method facilitates efficient training and reduces storage redundancy. Our method not only achieves state-of-the-art rendering quality whil e being 20× faster in training than previous SOTA methods with a minimum storage requirement of only 30MB on real-world datasets.
****
Vlad Sobal,Mark Ibrahim,Randall Balestriero,Vivien Cabannes,Diane Bouchacourt,Pi etro Astolfi,Kyunghyun Cho,Yann LeCun
$\mathbb{X}$-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs
Learning good representations involves capturing the diverse ways in which data samples relate. Contrastive loss—an objective matching related samples—underlies methods from self-supervised to multimodal learning. Contrastive losses, howeve r, can be viewed more broadly as modifying a similarity graph to indicate how sa mples should relate in the embedding space. This view reveals a shortcoming in contrastive learning: the similarity graph is binary, as only one sample is the related positive sample. Crucially, similarities \textit{across} samples are ign ored.
Based on this observation, we revise the standard contrastive loss to explicitly encode how a sample relates to others. We experiment with this new objective, c alled $\mathbb{X}$-Sample Contrastive, to train vision models based on similarit ies in class or text caption descriptions.
Our study spans three scales: ImageNet-1k with 1 million, CC3M with 3 million, a nd CC12M with 12 million samples. The representations learned via our objective outperform both contrastive self-supervised and vision-language models trained o n the same data across a range of tasks. When training on CC12M, we outperform C LIP by 0.6% on both ImageNet and ImageNet Real. Our objective appears to work particularly well in lower-data regimes, with gains over CLIP of 17.2% on Im ageNet and 18.0% on ImageNet Real when training with CC3M. Finally, our objec tive encourages the model to learn representations that separate objects from th eir attributes and backgrounds, with gains of 3.3-5.6\% over CLIP on ImageNe t9. The proposed method takes a step towards developing richer learning objectiv es for understanding sample relations in foundation models.
****
Yu Wang,Ryan A. Rossi,Namyong Park,Huiyuan Chen,Nesreen K. Ahmed,Puja Trivedi,Fr anck Dernoncourt,Danai Koutra,Tyler Derr
A Large-scale Training Paradigm for Graph Generative Models
Large Generative Models (LGMs) such as GPT, Stable Diffusion, Sora, and Suno are trained on a huge amount of texts, images, videos, and audio that are extremely diverse from numerous domains. This large-scale training paradigm on diverse we ll-curated data enhances the creativity and diversity of the generated content. However, all previous graph-generative models (e.g., GraphRNN, MDVAE, MoFlow, GD SS, and DiGress) have been trained only on one dataset each time, which cannot r eplicate the revolutionary success achieved by LGMs in other fields. To remedy t his crucial gap, we propose a large-scale training paradigm that uses a large co rpus of graphs (over 5000 graphs) from 13 domains, leading to the development of large graph generative models (LGGMs). We empirically demonstrate that the pre-trained LGGMs have superior zero-shot generative capability to existing graph ge nerative models. Furthermore, our pre-trained LGGMs can be easily fine-tuned wit h graphs from target domains and demonstrate even better performance than those directly trained from scratch, behaving as a solid starting point for real-world customization. Inspired by Stable Diffusion, we further equip LGGMs with the Te xt-to-Graph generation capability, such as providing the description of the netw ork name and domain (i.e., "The power-1138-bus graph represents a network of bus es in a power distribution system.") and network statistics (i.e., "The graph ha s a low average degree, suitable for modeling social media interactions."). This Text-to-Graph capability integrates the extensive world knowledge in the underl ying language model, offering users fine-grained control of the generated graphs . We release the code, the model checkpoint, and the datasets at https://github. com/KINDLab-Fly/LGGM.
****

Sara Oblak,Despoina Paschalidou,Sanja Fidler,Matan Atzmon
ReMatching Dynamic Reconstruction Flow
Reconstructing a dynamic scene from image inputs is a fundamental computer vision task with many downstream applications. Despite recent advancements, existing approaches still struggle to achieve high-quality reconstructions from unseen viewpoints and timestamps. This work introduces the ReMatching framework, designed to improve reconstruction quality by incorporating deformation priors into dynamic reconstruction models. Our approach advocates for velocity-field based priors, for which we suggest a matching procedure that can seamlessly supplement existing dynamic reconstruction pipelines. The framework is highly adaptable and can be applied to various dynamic representations. Moreover, it supports integrating multiple types of model priors and enables combining simpler ones to create more complex classes. Our evaluations on popular benchmarks involving both synthetic and real-world dynamic scenes demonstrate that augmenting current state-of-the-art methods with our approach leads to a clear improvement in reconstruction accuracy.
****
Xuan Son Nguyen,Shuo Yang,Aymeric Histace
Neural networks on Symmetric Spaces of Noncompact Type
Recent works have demonstrated promising performances of neural networks on hyperbolic spaces and symmetric positive definite (SPD) manifolds. These spaces belong to a family of Riemannian manifolds referred to as symmetric spaces of noncompact type. In this paper, we propose a novel approach for developing neural networks on such spaces. Our approach relies on a unified formulation of the distance from a point to a hyperplane on the considered spaces. We show that some existing formulations of the point-to-hyperplane distance can be recovered by our approach under specific settings. Furthermore, we derive a closed-form expression for the point-to-hyperplane distance in higher-rank symmetric spaces of noncompact type equipped with G-invariant Riemannian metrics. The derived distance then serves as a tool to design fully-connected (FC) layers and an attention mechanism for neural networks on the considered spaces. Our approach is validated on challenging benchmarks for image classification, electroencephalogram (EEG) signal classification, image generation, and natural language inference.
****
Nicolas Zilberstein,Morteza Mardani,Santiago Segarra
Repulsive Latent Score Distillation for Solving Inverse Problems
Score Distillation Sampling (SDS) has been pivotal for leveraging pre-trained diffusion models in downstream tasks such as inverse problems, but it faces two major challenges: (i) mode collapse and (ii) latent space inversion, which become more pronounced in high-dimensional data.
To address mode collapse, we introduce a novel variational framework for posterior sampling.
Utilizing the Wasserstein gradient flow interpretation of SDS, we propose a multimodal variational approximation with a \emph{repulsion} mechanism that promotes diversity among particles by penalizing pairwise kernel-based similarity.
This repulsion acts as a simple regularizer, encouraging a more diverse set of solutions.
To mitigate latent space ambiguity, we extend this framework with an \emph{augmented} variational distribution that disentangles the latent and data.
This repulsive augmented formulation balances computational efficiency, quality, and diversity.
Extensive experiments on linear and nonlinear inverse tasks with high-resolution images (512×512) using pre-trained Stable Diffusion models demonstrate the effectiveness of our approach.
****
Tianyuan Jin,Qin Zhang,Dongruo Zhou
Breaking the $\log(1/\Delta_2)$ Barrier: Better Batched Best Arm Identification with Adaptive Grids
We investigate the problem of batched best arm identification in multi-armed bandits, where we want to find the best arm from a set of n arms while minimizing

both the number of samples and batches. We introduce an algorithm that achieves near-optimal sample complexity and features an instance-sensitive batch complexity, which breaks the $\log(1/\Delta_2)$ barrier. The main contribution of our algorithm is a novel sample allocation scheme that effectively balances exploration and exploitation for batch sizes. Experimental results indicate that our approach is more batch-efficient across various setups. We also extend this framework to the problem of batched best arm identification in linear bandits and achieve similar improvements.
****
Kecheng Chen,Pingping Zhang,Tiexin Qin,Shiqi Wang,Hong Yan,Haoliang Li
Test-time Adaptation for Image Compression with Distribution Regularization
Current test- or compression-time adaptation image compression (TTA-IC) approaches, which leverage both latent and decoder refinements as a two-step adaptation scheme, have potentially enhanced the rate-distortion (R-D) performance of learned image compression models on cross-domain compression tasks, \textit{e.g.,} from natural to screen content images. However, compared with the emergence of various decoder refinement variants, the latent refinement, as an inseparable ingredient, is barely
tailored to cross-domain scenarios. To this end, we are interested in developing an advanced latent refinement method by extending the effective hybrid latent refinement (HLR) method, which is designed for \textit{in-domain} inference improvement but shows noticeable degradation of the rate cost in \textit{cross-domain} tasks. Specifically, we first provide theoretical analyses, in a cue of marginalization approximation from in- to cross-domain scenarios, to uncover that the vanilla HLR suffers from an underlying mismatch between refined Gaussian conditional and hyperprior distributions, leading to deteriorated joint probability approximation of marginal distribution with increased rate consumption. To remedy this issue, we introduce a simple Bayesian approximation-endowed \textit{distribution regularization} to encourage learning a better joint probability approximation in a plug-and-play manner. Extensive experiments on six in- and cross-domain datasets demonstrate that our proposed method not only improves the R-D performance compared with other latent refinement counterparts, but also can be flexibly integrated into existing TTA-IC methods with incremental benefits.
****
Steffen Schotthöfer,Emanuele Zangrando,Gianluca Ceruti,Francesco Tudisco,Jonas Kusch
GeoLoRA: Geometric integration for parameter efficient fine-tuning
Low-Rank Adaptation (LoRA) has become a widely used method for parameter-efficient fine-tuning of large-scale, pre-trained neural networks. However, LoRA and its extensions face several challenges, including the need for rank adaptivity, robustness, and computational efficiency during the fine-tuning process. We introduce GeoLoRA, a novel approach that addresses these limitations by leveraging dynamical low-rank approximation theory. GeoLoRA requires only a single backpropagation pass over the small-rank adapters, significantly reducing computational cost as compared to similar dynamical low-rank training methods and making it faster than popular baselines such as AdaLoRA. This allows GeoLoRA to efficiently adapt the allocated parameter budget across the model, achieving smaller low-rank adapters compared to heuristic methods like AdaLoRA and LoRA, while maintaining critical convergence, descent, and error-bound theoretical guarantees. The resulting method is not only more efficient but also more robust to varying hyperparameter settings. We demonstrate the effectiveness of GeoLoRA on several state-of-the-art benchmarks, showing that it outperforms existing methods in both
accuracy and computational efficiency
****
Xinpeng Liu,Junxuan Liang,Zili Lin,Haowen Hou,Yong-Lu Li,Cewu Lu
ImDy: Human Inverse Dynamics from Imitated Observations
Inverse dynamics (ID), which aims at reproducing the driven torques from human kinematic observations, has been a critical tool for gait analysis. However, it is hindered from wider application to general motion due to its limited scalability. Conventional optimization-based ID requires expensive laboratory setups, res

tricting its availability. To alleviate this problem, we propose to exploit the recently progressive human motion imitation algorithms to learn human inverse dynamics in a data-driven manner. The key insight is that the human ID knowledge is implicitly possessed by motion imitators, though not directly applicable. In light of this, we devise an efficient data collection pipeline with state-of-the-art motion imitation algorithms and physics simulators, resulting in a large-scale human inverse dynamics benchmark as Imitated Dynamics (ImDy). ImDy contains over 150 hours of motion with joint torque and full-body ground reaction force data. With ImDy, we train a data-driven human inverse dynamics solver ImDyS(olver) in a fully supervised manner, which conducts ID and ground reaction force estimation simultaneously. Experiments on ImDy and real-world data demonstrate the impressive competency of ImDyS in human inverse dynamics and ground reaction force estimation. Moreover, the potential of ImDy(-S) as a fundamental motion analysis tool is exhibited with downstream applications. The project page is https://foruck.github.io/ImDy.
****
Zhongshen Zeng,Pengguang Chen,Shu Liu,Haiyun Jiang,Jiaya Jia
MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation
In this work, we introduce a novel evaluation paradigm for Large Language Models (LLMs) that compels them to transition from a traditional question-answering role,
akin to a student, to a solution-scoring role, akin to a teacher. This paradigm, focusing on "reasoning about reasoning," termed meta-reasoning, shifts the emphasis
from result-oriented assessments, which often neglect the reasoning process, to a
more comprehensive evaluation that effectively distinguishes between the cognitive
capabilities of different models. Our meta-reasoning process mirrors "system-2" slow thinking, requiring careful examination of assumptions, conditions, calculations, and logic to identify mistakes. This paradigm enables one to transform
existed saturated, non-differentiating benchmarks that might be leaked in data pretraining stage to evaluation tools that are both challenging and robust against data
contamination. To prove our point, we applied our paradigm to GSM8K dataset and developed the MR-GSM8K benchmark. Our extensive analysis includes several
state-of-the-art models from both open-source and commercial domains, uncovering fundamental deficiencies in their training and evaluation methodologies.
Specifically, we found the OpenAI o1 models which possess characteristics of
"system-2" thinking excel the other SOTA models by more than 20 absolute points
in our benchmark, supporting our deficiency hypothesis.
****
Liangzu Peng,Juan Elenter,Joshua Agterberg,Alejandro Ribeiro,Rene Vidal
TSVD: Bridging Theory and Practice in Continual Learning with Pre-trained Models
The goal of continual learning (CL) is to train a model that can solve multiple
tasks presented sequentially. Recent CL approaches have achieved strong performance
by leveraging large pre-trained models that generalize well to downstream
tasks. However, such methods lack theoretical guarantees, making them prone to
unexpected failures. Conversely, principled CL approaches often fail to achieve
competitive performance. In this work, we aim to bridge this gap between theory
and practice by designing a simple CL method that is theoretically sound and
highly performant. Specifically, we lift pre-trained features into a higher dimensional
space and formulate an over-parametrized minimum-norm least-squares
problem. We find that the lifted features are highly ill-conditioned, potentially
leading to large training errors (numerical instability) and increased generalization
errors. We address these challenges by continually truncating the singular value

decomposition (SVD) of the lifted features. Our approach, termed TSVD, is stable with respect to the choice of hyperparameters, can handle hundreds of tasks, and outperforms state-of-the-art CL methods on multiple datasets. Importantly, our method satisfies a recurrence relation throughout its continual learning process,
which allows us to prove it maintains small training and generalization errors by
appropriately truncating a fraction of SVD factors. This results in a stable continual
learning method with strong empirical performance and theoretical guarantees. Code available: https://github.com/liangzu/tsvd.
****

Christopher Liao,Christian So,Theodoros Tsiligkaridis,Brian Kulis
Multimodal Unsupervised Domain Generalization by Retrieving Across the Modality Gap
Domain generalization (DG) is an important problem that learns a model which generalizes to unseen test domains leveraging one or more source domains, under the assumption of shared label spaces. However, most DG methods assume access to abundant source data in the target label space, a requirement that proves overly stringent for numerous real-world applications, where acquiring the same label space as the target task is prohibitively expensive. For this setting, we tackle the multimodal version of the unsupervised domain generalization (MUDG) problem, which uses a large task-agnostic unlabeled source dataset during finetuning. Our framework does not explicitly assume any relationship between the source dataset and target task. Instead, it relies only on the premise that the source dataset can be accurately and efficiently searched in a joint vision-language space. We make three contributions in the MUDG setting. Firstly, we show theoretically that cross-modal approximate nearest neighbor search suffers from low recall due to the large distance between text queries and the image centroids used for coarse quantization. Accordingly, we propose paired k-means, a simple clustering algorithm that improves nearest neighbor recall by storing centroids in query space instead of image space. Secondly, we propose an adaptive text augmentation scheme for target labels designed to improve zero-shot accuracy and diversify retrieved image data. Lastly, we present two simple but effective components to further improve downstream target accuracy. We compare against state-of-the-art name-only transfer, source-free DG and zero-shot (ZS) methods on their respective benchmarks and show consistent improvement in accuracy on 20 diverse datasets. Code is available: https://github.com/Chris210634/mudg
****

Hong Qian,Yiyi Zhu,Xiang Shu,Shuo Liu,Yaolin Wen,Xin An,Huakang Lu,Aimin Zhou,Ke Tang,Yang Yu
SOO-Bench: Benchmarks for Evaluating the Stability of Offline Black-Box Optimization
Black-box optimization aims to find the optima through building a model close to the black-box objective function based on function value evaluation. However, in many real-world tasks, such as the design of molecular formulas and mechanical structures, it is perilous, costly, or even infeasible to evaluate the objective function value of an actively sampled solution. In this situation, optimization can only be conducted via utilizing offline historical data, which yields offline black-box optimization. Different from the traditional goal that is to pursue the optimal solution, this paper emphasizes that the goal of offline optimization is to stably surpass the offline dataset during optimization procedure. Although benchmarks called Design-Bench already exist in this emerging field, it can hardly evaluate the stability of offline optimization and mainly provides real-world offline tasks and the corresponding offline datasets. To this end, this paper proposes benchmarks named SOO-Bench (i.e., Stable Offline Optimization Benchmarks) for offline black-box optimization algorithms, so as to systematically evaluate the stability of surpassing the offline dataset under different data distributions. Along with SOO-Bench, we also propose a stability indicator to measure the degree of stability. Specifically, SOO-Bench includes various real-world o

ffline optimization tasks and offline datasets under different data distribution
s, involving the fields of satellites, materials science, structural mechanics,
and automobile manufacturing. Empirically, baseline and state-of-the-art algorit
hms are tested and analyzed on SOO-Bench. Hopefully, SOO-Bench is expected to se
rve as a catalyst for the rapid developments of more novel and stable offline op
timization methods. The code is available at \url{https://github.com/zhuyiyi-123
/SOO-Bench}.
****
Francesco D'Angelo,Francesco Croce,Nicolas Flammarion
Selective induction Heads: How Transformers Select Causal Structures in Context
Transformers have exhibited exceptional capabilities in sequence modelling tasks
, leveraging self-attention and in-context learning. Critical to this success ar
e induction heads, attention circuits that enable copying tokens based on their
previous occurrences. In this work, we introduce a novel synthetic framework des
igned to enable the theoretical analysis of transformers' ability to dynamically
 handle causal structures. Existing works rely on Markov Chains to study the for
mation of induction heads, revealing how transformers capture causal dependencie
s and learn transition probabilities in-context. However, they rely on a fixed c
ausal structure that fails to capture the complexity of natural languages, where
 the relationship between tokens dynamically changes with context. To this end,
 our framework varies the causal structure through interleaved Markov chains wit
h different lags while keeping the transition probabilities fixed. This setting
unveils the formation of Selective Induction Heads, a new circuit that endows tr
ansformers with the ability to select the correct causal structure in-context. W
e empirically demonstrate that attention-only transformers learn this mechanism
to predict the next token by identifying the correct lag and copying the corresp
onding token from the past. We provide a detailed construction of a 3-layer tran
sformer to implement the selective induction head, and a theoretical analysis pr
oving that this mechanism asymptotically converges to the maximum likelihood sol
ution. Our findings advance the theoretical understanding of how transformers se
lect causal structures, providing new insights into their functioning and interp
retability.
****
Litu Rout,Yujia Chen,Nataniel Ruiz,Abhishek Kumar,Constantine Caramanis,Sanjay S
hakkottai,Wen-Sheng Chu
RB-Modulation: Training-Free Stylization using Reference-Based Modulation
We propose Reference-Based Modulation (RB-Modulation), a new plug-and-play solut
ion for training-free personalization of diffusion models.
Existing training-free approaches exhibit difficulties in (a) style extraction f
rom reference images in the absence of additional style or content text descript
ions, (b) unwanted content leakage from reference style images, and (c) effectiv
e composition of style and content.
RB-Modulation is built on a novel stochastic optimal controller where a style de
scriptor encodes the desired attributes through a terminal cost.
The resulting drift not only overcomes the difficulties above, but also ensures
high fidelity to the reference style and adheres to the given text prompt.
We also introduce a cross-attention-based feature aggregation scheme that allows
 RB-Modulation to decouple content and style from the reference image.
With theoretical justification and empirical evidence, our test-time optimizatio
n framework demonstrates precise extraction and control of *content* and *style*
 in a training-free manner.
Further, our method allows a seamless composition of content and style, which ma
rks a departure from the dependency on external adapters or ControlNets. See pro
ject page: https://rb-modulation.github.io/ for code and further details.
****
Oskar van der Wal,Pietro Lesci,Max Müller-Eberstein,Naomi Saphra,Hailey Schoelko
pf,Willem Zuidema,Stella Biderman
PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Run
s
The stability of language model pre-training and its effects on downstream perfo

rmance are still understudied. Prior work shows that the training process can yi
eld significantly different results in response to slight variations in initial
conditions, e.g., the random seed.
Crucially, the research community still lacks sufficient resources and tools to
systematically investigate pre-training stability, particularly for decoder-only
language models. We introduce the PolyPythias, a set of 45 new training runs fo
r the Pythia model suite: 9 new seeds across 5 model sizes, from 14M to 410M par
ameters, resulting in about 7k new checkpoints that we release. Using these new
45 training runs, in addition to the 5 already available, we study the effects o
f different initial conditions determined by the seed---i.e., parameters' initia
lisation and data order---on (i) downstream performance, (ii) learned linguistic
representations, and (iii) emergence of training phases. In addition to common
scaling behaviours, our analyses generally reveal highly consistent training dyn
amics across both model sizes and initial conditions. Further, the new seeds for
each model allow us to identify outlier training runs and delineate their chara
cteristics.
Our findings show the potential of using these methods to predict training stabi
lity.
****
DiJia Su,Sainbayar Sukhbaatar,Michael Rabbat,Yuandong Tian,Qinqing Zheng
Dualformer: Controllable Fast and Slow Thinking by Learning with Randomized Reas
oning Traces
In cognition theory, human thinking is governed by two systems: the fast and int
uitive System 1 and the slower but more deliberative System 2. Analogously, Larg
e Language Models (LLMs) can operate in two reasoning modes: outputting only the
solutions (\emph{fast mode}) or both the reasoning chain and the final solution
(\emph{slow mode}). We present \dualformer, a single Transformer model that se
amlessly integrates both the fast and slow reasoning modes by training on random
ized reasoning traces, where different parts of the traces are strategically dro
pped during training. At inference time, \dualformer can be easily configured to
execute in either fast or slow mode, or automatically decide which mode to enga
ge (\emph{auto mode}). It outperforms baselines in both performance and computa
tional efficiency across all three modes: \textbf{(1)} in slow mode, \dualformer
achieves 97.6% optimal rate on unseen 30×30 maze tasks, surpassing
the \searchformer baseline (93.3\%) trained on data with complete reasoning trac
es, with 45.5% fewer reasoning steps; \textbf{(2)} in fast mode, \dualformer
achieves 80% optimal rate, significantly outperforming the Solution-Only mod
el trained on solution-only data, which has an optimal rate of only 30\%; \tex
tbf{(3)} in auto mode, \dualformer achieves 96.6% optimal rate with 59.9%
fewer steps than \searchformer. For math reasoning problems, our techniques hav
e also achieved improved performance with LLM fine-tuning, demonstrating its gen
eralization beyond task-specific models. We open source our code at https://gith
ub.com/facebookresearch/dualformer.
****
Vahideh Sanjaroonpouri,Pouria Ramazi
Linear SCM Identification in the Presence of Confounders and Gaussian Noise
Noisy linear structural causal models (SCMs) in the presence of confounding vari
ables are known to be identifiable if all confounding and noise variables are no
n-Gaussian and unidentifiable if all are Gaussian.

The identifiability when only some are Gaussian remains concealed.

We show that, in the presence of Gaussian noise, a linear SCM is uniquely id
entifiable provided that \emph{(i)} the number of confounders is at most the num
ber of the observed variables, \emph{(ii)} the confounders do not have a Gaussia
n component, and \emph{(iii)} the causal structure of the SCM is known.

If the third condition is relaxed, the SCM becomes finitely identifiable; mo
re specifically, it belongs to a set of at most $n!$ linear SCMS, where n is t
he number of observed variables.

The confounders in all of these $n!$ SCMs share the same joint probability d
istribution function (PDF), which we obtain analytically.

For the case where both the noise and confounders are Gaussian, we provide f

urther insight into the existing counter-example-based unidentifiability result and demonstrate that every SCM with confounders can be represented as an SCM wit hout confounders but with the same joint PDF.
****

Yixuan Weng,Minjun Zhu,Guangsheng Bao,Hongbo Zhang,Jindong Wang,Yue Zhang,Linyi Yang

CycleResearcher: Improving Automated Research via Automated Review

The automation of scientific discovery has been a long-standing goal within the research community, driven by the potential to accelerate knowledge creation. Wh ile significant progress has been made using commercial large language models (L LMs) as research assistants or idea generators, the possibility of automating th e entire research process with open-source LLMs remains largely unexplored. This paper explores the feasibility of using open-source post-trained LLMs as autono mous agents capable of performing the full cycle of automated research and revie w, from literature review and manuscript preparation to peer review and paper re finement. Our iterative preference training framework consists of CycleResearche r, which conducts research tasks, and CycleReviewer, which simulates the peer re view process, providing iterative feedback via reinforcement learning. To train these models, we develop two new datasets, Review-5k and Research-14k, reflectin g real-world machine learning research and peer review dynamics. Our results dem onstrate that CycleReviewer achieves promising performance with a 26.89\% reduct ion in mean absolute error (MAE) compared to individual human reviewers in predi cting paper scores, indicating the potential of LLMs to effectively assist exper t-level research evaluation. In research, the papers generated by the CycleResea rcher model achieved a score of 5.36 in simulated peer reviews, showing some com petitiveness in terms of simulated review scores compared to the preprint level of 5.24 from human experts, while still having room for improvement compared to the accepted paper level of 5.69. This work represents a significant step toward fully automated scientific inquiry, providing ethical safeguards and exploring AI-driven research capabilities. The code, dataset and model weight are released at https://wengsyx.github.io/Researcher.
****

Fabrizio Garuti,Enver Sangineto,Simone Luetto,Lorenzo Forni,Rita Cucchiara

Diffusion Transformers for Tabular Data Time Series Generation

Tabular data generation has recently attracted a growing interest due to its dif ferent application scenarios. However,
generating time series of tabular data, where each element of the series depends on the others,
remains a largely unexplored domain.
This gap is probably due to the difficulty of jointly solving different problems , the main of which are the heterogeneity of tabular data (a problem common to n on-time-dependent approaches) and the variable length of a time series.
In this paper, we propose a Diffusion Transformers (DiTs) based approach for tab ular data series generation. Inspired by the recent success of DiTs in image and video generation, we extend this framework to deal with heterogeneous data and variable-length sequences.
Using extensive experiments on six datasets, we show that the proposed approach outperforms previous work by a large margin.
****

Xiaogeng Liu,Peiran Li,G. Edward Suh,Yevgeniy Vorobeychik,Zhuoqing Mao,Somesh Jh a,Patrick McDaniel,Huan Sun,Bo Li,Chaowei Xiao

AutoDAN-Turbo: A Lifelong Agent for Strategy Self-Exploration to Jailbreak LLMs

Jailbreak attacks serve as essential red-teaming tools, proactively assessing wh ether LLMs can behave responsibly and safely in adversarial environments. Despit e diverse strategies (e.g., cipher, low-resource language, persuasions, and so o n) that have been proposed and shown success, these strategies are still manuall y designed, limiting their scope and effectiveness as a red-teaming tool. In thi s paper, we propose AutoDAN-Turbo, a black-box jailbreak method that can automat ically discover as many jailbreak strategies as possible from scratch, without a ny human intervention or predefined scopes (e.g., specified candidate strategies

), and use them for red-teaming. As a result, AutoDAN-Turbo can significantly ou
tperform baseline methods, achieving a 74.3% higher average attack success rate
on public benchmarks. Notably, AutoDAN-Turbo achieves an 88.5 attack success rat
e on GPT-4-1106-turbo. In addition, AutoDAN-Turbo is a unified framework that ca
n incorporate existing human-designed jailbreak strategies in a plug-and-play ma
nner. By integrating human-designed strategies, AutoDAN-Turbo can even achieve a
 higher attack success rate of 93.4 on GPT-4-1106-turbo.
****
Binwei Yao,Zefan Cai,Yun-Shiuan Chuang,Shanglin Yang,Ming Jiang,Diyi Yang,Junjie
 Hu
No Preference Left Behind: Group Distributional Preference Optimization
Preferences within a group of people are not uniform but follow a distribution.
While existing alignment methods like Direct Preference Optimization (DPO) attem
pt to steer models to reflect human preferences, they struggle to capture the di
stributional pluralistic preferences within a group. These methods often skew to
ward dominant preferences, overlooking the diversity of opinions, especially whe
n conflicting preferences arise. To address this issue, we propose Group Distrib
utional Preference Optimization (GDPO), a novel framework that aligns language m
odels with the distribution of preferences within a group by incorporating the c
oncept of beliefs that shape individual preferences. GDPO calibrates a language
model using statistical estimation of the group's belief distribution and aligns
 the model with belief-conditioned preferences, offering a more inclusive alignm
ent framework than traditional methods. In experiments using both synthetic cont
rollable opinion generation and real-world movie review datasets, we show that D
PO fails to align with the targeted belief distributions, while GDPO consistentl
y reduces this alignment gap during training. Additionally, our evaluation metri
cs demonstrate that GDPO outperforms existing approaches in aligning with group
distributional preferences, marking a significant advance in pluralistic alignme
nt.
****
Weifeng Lin,Xinyu Wei,Ruichuan An,Peng Gao,Bocheng Zou,Yulin Luo,Siyuan Huang,Sh
anghang Zhang,Hongsheng Li
Draw-and-Understand: Leveraging Visual Prompts to Enable MLLMs to Comprehend Wha
t You Want
In this paper, we present the Draw-and-Understand framework, exploring how to in
tegrate visual prompting understanding capabilities into Multimodal Large Langua
ge Models (MLLMs). Visual prompts allow users to interact through multi-modal in
structions, enhancing the models' interactivity and fine-grained image comprehen
sion. In this framework, we propose a general architecture adaptable to differen
t pre-trained MLLMs, enabling it to recognize various types of visual prompts (s
uch as points, bounding boxes, and free-form shapes) alongside language understa
nding. Additionally, we introduce MDVP-Instruct-Data, a multi-domain dataset fea
turing 1.2 million image-visual prompt-text triplets, including natural images,
document images, scene text images, mobile/web screenshots, and remote sensing i
mages. Building on this dataset, we introduce MDVP-Bench, a challenging benchmar
k designed to evaluate a model's ability to understand visual prompting instruct
ions. The experimental results demonstrate that our framework can be easily and
effectively applied to various MLLMs, such as SPHINX-X and LLaVA. After training
 with MDVP-Instruct-Data and image-level instruction datasets, our models exhibi
t impressive multimodal interaction capabilities and pixel-level understanding,
while maintaining their image-level visual perception performance.
****
Jun-Yu Ma,Hong Wang,Hao-Xiang Xu,Zhen-Hua Ling,Jia-Chen Gu
Perturbation-Restrained Sequential Model Editing
Model editing is an emerging field that focuses on updating the knowledge embedd
ed within large language models (LLMs) without extensive retraining. However, cu
rrent model editing methods significantly compromise the general abilities of LL
Ms as the number of edits increases, and this trade-off poses a substantial chal
lenge to the continual learning of LLMs. In this paper, we first theoretically a
nalyze that the factor affecting the general abilities in sequential model editi

ng lies in the condition number of the edited matrix. The condition number of a matrix represents its numerical sensitivity, and therefore can be used to indicate the extent to which the original knowledge associations stored in LLMs are perturbed after editing. Subsequently, statistical findings demonstrate that the value of this factor becomes larger as the number of edits increases, thereby exacerbating the deterioration of general abilities. To this end, a framework termed Perturbation Restraint on Upper bouNd for Editing (PRUNE) is proposed, which applies the condition number restraints in sequential editing. These restraints can lower the upper bound on perturbation to edited models, thus preserving the general abilities.
Systematically, we conduct experiments employing three editing methods on three LLMs across four downstream tasks.
The results show that PRUNE can preserve general abilities while maintaining the editing performance effectively in sequential model editing. The code are available at https://github.com/mjy1111/PRUNE.
****
Wuchao Li,Kai Zheng,Defu Lian,Qi Liu,Wentian Bao,Yun En Yu,Yang Song,Han Li,Kun Gai
Making Transformer Decoders Better Differentiable Indexers
Retrieval aims to find the top-k items most relevant to a query/user from a large dataset. Traditional retrieval models represent queries/users and items as embedding vectors and use Approximate Nearest Neighbor (ANN) search for retrieval. Recently, researchers have proposed a generative-based retrieval method that represents items as token sequences and uses a decoder model for autoregressive training. Compared to traditional methods, this approach uses more complex models and integrates index structure during training, leading to better performance. However, these methods remain two-stage processes, where index construction is separate from the retrieval model, limiting the model's overall capacity. Additionally, existing methods construct indices by clustering pre-trained item representations in Euclidean space. However, real-world scenarios are more complex, making this approach less accurate. To address these issues, we propose a \underline{U}nified framework for \underline{R}etrieval and \underline{I}ndexing, termed \textbf{URI}. URI ensures strong consistency between index construction and the retrieval model, typically a Transformer decoder. URI simultaneously builds the index and trains the decoder, constructing the index through the decoder itself. It no longer relies on one-sided item representations in Euclidean space but constructs the index within the interactive space between queries and items. Experimental comparisons on three real-world datasets show that URI significantly outperforms existing methods.
****
Andrew Jesson,Nicolas Beltran-Velez,David Blei
Can Generative AI Solve Your In-Context Learning Problem? A Martingale Perspective
This work is about estimating when a conditional generative model (CGM) can solve an in-context learning (ICL) problem. An in-context learning (ICL) problem comprises a CGM, a dataset, and a prediction task. The CGM could be a multi-modal foundation model; the dataset, a collection of patient histories, test results, and recorded diagnoses; and the prediction task to communicate a diagnosis to a new patient. A Bayesian interpretation of ICL assumes that the CGM computes a posterior predictive distribution over an unknown Bayesian model defining a joint distribution over latent explanations and observable data. From this perspective, Bayesian model criticism is a reasonable approach to assess the suitability of a given CGM for an ICL problem. However, such approaches---like posterior predictive checks (PPCs)---often assume that we can sample from the likelihood and posterior defined by the Bayesian model, which are not explicitly given for contemporary CGMs. To address this, we show when ancestral sampling from the predictive distribution of a CGM is equivalent to sampling datasets from the posterior predictive of the assumed Bayesian model. Then we develop the generative predictive p-value, which enables PPCs and their cousins for contemporary CGMs. The generative predictive p-value can be used in a statistical decision procedure to d

etermine when the model is appropriate for an ICL problem. Our method only requi
res generating queries and responses from a CGM and evaluating its response log
probability. Using large language models, we empirically evaluate our method on
tasks involving tabular data, imaging data, and natural language data.
****

Ziyu Lu,Wuwei Zhang,Trung Le,Hao Wang,Uygar Sümbül,Eric Todd SheaBrown,Lu Mi
NetFormer: An interpretable model for recovering dynamical connectivity in neuro
nal population dynamics
Neuronal dynamics are highly nonlinear and nonstationary. Traditional methods fo
r extracting the underlying network structure from neuronal activity recordings
mainly concentrate on modeling static connectivity, without accounting for key n
onstationary aspects of biological neural systems, such as ongoing synaptic plas
ticity and neuronal modulation. To bridge this gap, we introduce the NetFormer m
odel, an interpretable approach applicable to such systems. In NetFormer, the ac
tivity of each neuron across a series of historical time steps is defined as a t
oken. These tokens are then linearly mapped through a query and key mechanism to
 generate a state- (and hence time-) dependent attention matrix that directly en
codes nonstationary connectivity structures. We analyze our formulation from the
 perspective of nonstationary and nonlinear networked dynamical systems, and sho
w both via an analytical expansion and targeted simulations how it can approxima
te the underlying ground truth. Next, we demonstrate NetFormer's ability to mo
del a key feature of biological networks, spike-timing-dependent plasticity, whe
reby connection strengths continually change in response to local activity patte
rns. We further demonstrate that NetFormer can capture task-induced connectivity
 patterns on activity generated by task-trained recurrent neural networks. Thus
informed, we apply NetFormer to a multi-modal dataset of real neural recordings,
 which contains neural activity, cell type, and behavioral state information. W
e show that the NetFormer effectively predicts neural dynamics and identifies ce
ll-type specific, state-dependent dynamic connectivity that matches patterns mea
sured in separate ground-truth physiology experiments, demonstrating its ability
 to help decode complex neural interactions based on population activity observa
tions alone.
****

Tong Chen,Hao Fang,Patrick Xia,Xiaodong Liu,Benjamin Van Durme,Luke Zettlemoyer,
Jianfeng Gao,Hao Cheng
Generative Adapter: Contextualizing Language Models in Parameters with A Single
Forward Pass
Large language models (LLMs) acquire substantial knowledge during pretraining bu
t often need adaptation to new contexts, tasks, or domains, typically achieved t
hrough fine-tuning or prompting. However, fine-tuning incurs significant trainin
g costs, while prompting increases inference overhead. Inspired by fast weight m
emory, we introduce GenerativeAdapter, an effective and efficient adaptation met
hod that encode test-time context into language model parameters with a single f
orward pass.
GenerativeAdapter augments a frozen pretrained LM with a lightweight adapter gen
erator, trained via self-supervised learning, to produce parameter-efficient ada
pters.
Notably, our generator is general-purpose, i.e., one generator can adapt the cor
responding base model for all langauge processing scenarios.
We apply GenerativeAdapter to two pretrained LMs (Mistral-7B-Instruct and Llama2
-7B-Chat) and evaluate the adapted models across knowledge acquisition from doc
uments, learning from demonstrations, and personalization for users.
In StreamingQA, our approach is effective in injecting knowledge into the LM's p
arameters, achieving a 63.5\% improvement in F1 score over the model with superv
ised fine-tuning (from 19.5 to 31.5) for contexts as long as 32K tokens.
In the MetaICL in-context learning evaluation, our method achieves an average ac
curacy of 44.9 across 26 tasks, outperforming the base model.
On MSC, our method proves to be highly competitive in memorizing user informatio
n from conversations with a 4x reduction in computation and memory costs compare
d to

prompting with full conversation history.
Overall, GenerativeAdapter provides a viable solution for adapting large LMs to evolving information and providing tailored user experience, while reducing training and inference costs relative to traditional fine-tuning and prompting techniques.
****
Yiming Zhang,Athul Paul Jacob,Vivian Lai,Daniel Fried,Daphne Ippolito
Human-Aligned Chess With a Bit of Search
Chess has long been a testbed for AI's quest to match human intelligence, and in recent years, chess AI systems have surpassed the strongest humans at the game. However, these systems are *not human-aligned*; they are unable to match the skill levels of all human partners or model human-like behaviors beyond piece movement.
In this paper, we introduce Allie, a chess-playing AI designed to bridge the gap between artificial and human intelligence in this classic game.
Allie is trained on log sequences of real chess games to model the behaviors of human chess players across the skill spectrum, including non-move behaviors such as pondering times and resignations
In offline evaluations, we find that Allie exhibits humanlike behavior: it outperforms the existing state-of-the-art in human chess move prediction and ``ponders'' at critical positions.
The model learns to reliably assign reward at each game state, which can be used at inference as a reward function in a novel *time-adaptive* Monte-Carlo tree search (MCTS) procedure, where the amount of search depends on how long humans would think in the same positions.
Adaptive search enables remarkable *skill calibration*; in a large-scale online evaluation against players with ratings from 1000 to 2600 Elo, our adaptive search method leads to a skill gap of only 49 Elo on average, substantially outperforming search-free and standard MCTS baselines.
Against grandmaster-level (2500 Elo) opponents, Allie with adaptive search exhibits the strength of a fellow grandmaster, all while learning *exclusively from humans*.
****
Onkar Kishor Susladkar,Jishu Sen Gupta,Chirag Sehgal,Sparsh Mittal,Rekha Singhal
MotionAura: Generating High-Quality and Motion Consistent Videos using Discrete Diffusion
The spatio-temporal complexity of video data presents significant challenges in tasks such as compression, generation, and inpainting. We present four key contributions to address the challenges of spatiotemporal video processing. First, we introduce the 3D Mobile Inverted Vector-Quantization Variational Autoencoder (3D-MBQ-VAE), which combines Variational Autoencoders (VAEs) with masked modeling to enhance spatiotemporal video compression. The model achieves superior temporal consistency and state-of-the-art (SOTA) reconstruction quality by employing a novel training strategy with full frame masking. Second, we present MotionAura, a text-to-video generation framework that utilizes vector-quantized diffusion models to discretize the latent space and capture complex motion dynamics, producing temporally coherent videos aligned with text prompts. Third, we propose a spectral transformer-based denoising network that processes video data in the frequency domain using the Fourier Transform. This method effectively captures global context and long-range dependencies for high-quality video generation and denoising. Lastly, we introduce a downstream task of Sketch Guided Video Inpainting. This task leverages Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. Our models achieve SOTA performance on a range of benchmarks. Our work offers robust frameworks for spatiotemporal modeling and user-driven video content manipulation.
****
Kiho Park,Yo Joong Choe,Yibo Jiang,Victor Veitch
The Geometry of Categorical and Hierarchical Concepts in Large Language Models
The linear representation hypothesis is the informal idea that semantic concepts are encoded as linear directions in the representation spaces of large language

models (LLMs). Previous work has shown how to make this notion precise for representing binary concepts that have natural contrasts (e.g., {male, female}) as _directions_ in representation space. However, many natural concepts do not have natural contrasts (e.g., whether the output is about an animal). In this work, we show how to extend the formalization of the linear representation hypothesis to represent features (e.g., is_animal) as _vectors_. This allows us to immediately formalize the representation of categorical concepts as polytopes in the representation space. Further, we use the formalization to prove a relationship between the hierarchical structure of concepts and the geometry of their representations. We validate these theoretical results on the Gemma and LLaMA-3 large language models, estimating representations for 900+ hierarchically related concepts using data from WordNet.
****
Jie Zhang,Zhongqi Wang,Mengqi Lei,Zheng Yuan,Bei Yan,Shiguang Shan,Xilin Chen
Dysca: A Dynamic and Scalable Benchmark for Evaluating Perception Ability of LVLMs
Currently many benchmarks have been proposed to evaluate the perception ability of the Large Vision-Language Models (LVLMs).
However, most benchmarks conduct questions by selecting images from existing datasets, resulting in the potential data leakage. Besides, these benchmarks merely focus on evaluating LVLMs on the realistic style images and clean scenarios, leaving the multi-stylized images and noisy scenarios unexplored. In response to these challenges, we propose a dynamic and scalable benchmark named Dysca for evaluating LVLMs by leveraging synthesis images. Specifically, we leverage Stable Diffusion and design a rule-based method to dynamically generate novel images, questions and the corresponding answers. We consider 51 kinds of image styles and evaluate the perception capability in 20 subtasks. Moreover, we conduct evaluations under 4 scenarios (i.e., Clean, Corruption, Print Attacking and Adversarial Attacking) and 3 question types (i.e., Multi-choices, True-or-false and Free-form). Thanks to the generative paradigm, Dysca serves as a scalable benchmark for easily adding new subtasks and scenarios. A total of 24 advanced open-source LVLMs and 2 close-source LVLMs are evaluated on Dysca, revealing the drawbacks of current LVLMs. The benchmark is released in anonymous github page \url{https://github.com/Benchmark-Dysca/Dysca}.
****
Aniket Rajiv Didolkar,Andrii Zadaianchuk,Anirudh Goyal,Michael Curtis Mozer,Yoshua Bengio,Georg Martius,Maximilian Seitzer
On the Transfer of Object-Centric Representation Learning
The goal of object-centric representation learning is to decompose visual scenes into a structured representation that isolates the entities into individual vectors. Recent successes have shown that object-centric representation learning can be scaled to real-world scenes by utilizing features from pre-trained foundation models like DINO. However, so far, these object-centric methods have mostly been applied in-distribution, with models trained and evaluated on the same dataset. This is in contrast to the underlying foundation models, which have been shown to be applicable to a wide range of data and tasks. Thus, in this work, we answer the question of whether current real-world capable object-centric methods exhibit similar levels of transferability by introducing a benchmark comprising seven different synthetic and real-world datasets. We analyze the factors influencing performance under transfer and find that training on diverse real-world images improves generalization to unseen scenarios. Furthermore, inspired by the success of task-specific fine-tuning in foundation models, we introduce a novel fine-tuning strategy to adapt pre-trained vision encoders for the task of object discovery. We find that the proposed approach results in state-of-the-art performance for unsupervised object discovery, exhibiting strong zero-shot transfer to unseen datasets.
****
Binghao Liu,Han Yang,Fang Wan,Fei Gu
Prototype antithesis for biological few-shot class-incremental learning
Deep learning has become essential in the biological species recognition task. H

owever, a significant challenge is the ability to continuously learn new or muta
ted species with limited annotated samples. Since species within the same family
 typically share similar traits, distinguishing between new and existing (old) s
pecies during incremental learning often faces the issue of species confusion. T
his can result in "catastrophic forgetting" of old species and poor learning of
new ones. To address this issue, we propose a Prototype Antithesis (PA) method,
which leverages the hierarchical structures in biological taxa to reduce confusi
on between new and old species. PA operates in two steps: Residual Prototype Lea
rning (RPL) and Residual Prototype Mixing (RPM). RPL enables the model to learn
unique prototypes for each species alongside residual prototypes representing sh
ared traits within families. RPM generates synthetic samples by blending feature
s of new species with residual prototypes of old species, encouraging the model
to focus on species-unique traits and minimize species confusion. By integrating
 RPL and RPM, the proposed PA method mitigates "catastrophic forgetting" while i
mproving generalization to new species. Extensive experiments on CUB200, PlantVi
llage, and Tree-of-Life datasets demonstrate that PA significantly reduces inter
-species confusion and achieves state-of-the-art performance, highlighting its p
otential for deep learning in biological data analysis.
****

Yunju Cho,Jay-Yoon Lee

CoMRes: Semi-Supervised Time Series Forecasting Utilizing Consensus Promotion of
 Multi-Resolution

Long-term time series forecasting poses significant challenges due to the comple
x dynamics and temporal variations, particularly when dealing with unseen patter
ns and data scarcity. Traditional supervised learning approaches, which rely on
cleaned and labeled data, struggle to capture these unseen characteristics, limi
ting their effectiveness in real-world applications. In this study, we propose a
 semi-supervised approach that leverages multi-view setting on augmented data wi
thout requiring explicit future values as labels to address these limitations. B
y introducing a consensus promotion framework, our method enhances agreement amo
ng multiple single-view models on unseen augmented data. This approach not only
improves forecasting accuracy but also mitigates error accumulation in long-hori
zon predictions. Furthermore, we explore the impact of autoregressive and non-au
toregressive decoding schemes on error propagation, demonstrating the robustness
 of our model in extending prediction horizons. Experimental results show that o
ur proposed method not only surpasses traditional supervised models in accuracy
but also exhibits greater robustness when extending the prediction horizon.
****

Sreyan Ghosh,Sonal Kumar,Zhifeng Kong,Rafael Valle,Bryan Catanzaro,Dinesh Manoch
a

Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Dat
a

We present Synthio, a novel approach for augmenting small-scale audio classifica
tion datasets with synthetic data. Our goal is to improve audio classification a
ccuracy with limited labeled data. Traditional data augmentation techniques, whi
ch apply artificial transformations (e.g., adding random noise or masking segmen
ts), struggle to create data that captures the true diversity present in real-wo
rld audios. To address this shortcoming, we propose to augment the dataset with
synthetic audio generated from text-to-audio (T2A) diffusion models. However, sy
nthesizing effective augmentations is challenging because not only should the ge
nerated data be acoustically consistent with the underlying small-scale dataset,
 but they should also have sufficient compositional diversity. To overcome the f
irst challenge, we align the generations of the T2A model with the small-scale d
ataset using preference optimization. This ensures that the acoustic characteris
tics of the generated data remain consistent with the small-scale dataset. To ad
dress the second challenge, we propose a novel caption generation technique that
 leverages the reasoning capabilities of Large Language Models to (1) generate d
iverse and meaningful audio captions and (2) iteratively refine their quality. T
he generated captions are then used to prompt the aligned T2A model. We extensiv
ely evaluate Synthio on ten datasets and four simulated limited-data settings. R

esults indicate our method consistently outperforms all baselines by 0.1%-39% us
ing a T2A model trained only on weakly-captioned AudioSet.
****

Mengjing Wu,Junyu Xuan,Jie Lu
Bridging the Gap between Variational Inference and Stochastic Gradient MCMC in F
unction Space
Traditional parameter-space posterior inference for Bayesian neural networks fac
es several challenges, such as the difficulty in specifying meaningful prior, th
e potential pathologies in deep models and the intractability for multi-modal po
sterior. To address these issues, functional variational inference (fVI) and fun
ctional Markov Chain Monte Carlo (fMCMC) are two recently emerged Bayesian infer
ence schemes that perform posterior inference directly in function space by inco
rporating more informative functional priors. Similar to their parameter-space c
ounterparts, fVI and fMCMC have their own strengths and weaknesses. For instance
, fVI is computationally efficient but imposes strong distributional assumptions
, while fMCMC is asymptotically exact but suffers from slow mixing in high dimen
sions. To inherit the complementary benefits of both schemes, this work proposes
 a novel hybrid inference method for functional posterior inference. Specificall
y, it combines fVI and fMCMC successively by an elaborate linking mechanism to f
orm an alternating approximation process. We also provide theoretical justificat
ion for the soundness of such a hybrid inference through the lens of Wasserstein
 gradient flows in the function space. We evaluate our method on several benchma
rk tasks and observe improvements in both predictive accuracy and uncertainty qu
antification compared to parameter/function-space VI and MCMC.
****

Yupei Yang,Biwei Huang,Fan Feng,Xinyue Wang,Shikui Tu,Lei Xu
Towards Generalizable Reinforcement Learning via Causality-Guided Self-Adaptive
Representations
General intelligence requires quick adaptation across tasks. While existing rein
forcement learning (RL) methods have made progress in generalization, they typic
ally assume only distribution changes between source and target domains. In this
 paper, we explore a wider range of scenarios where not only the distribution bu
t also the environment spaces may change. For example, in the CoinRun environmen
t, we train agents from easy levels and generalize them to difficulty levels whe
re there could be new enemies that have never occurred before. To address this c
hallenging setting, we introduce a causality-guided self-adaptive representation
-based approach, called CSR, that equips the agent to generalize effectively acr
oss tasks with evolving dynamics. Specifically, we employ causal representation
learning to characterize the latent causal variables within the RL system. Such
compact causal representations uncover the structural relationships among variab
les, enabling the agent to autonomously determine whether changes in the environ
ment stem from distribution shifts or variations in space, and to precisely loca
te these changes. We then devise a three-step strategy to fine-tune the causal m
odel under different scenarios accordingly. Empirical experiments show that CSR
efficiently adapts to the target domains with only a few samples and outperforms
 state-of-the-art baselines on a wide range of scenarios, including our simulate
d environments, CartPole, CoinRun and Atari games.
****

Chi Zhang,Huaping Zhong,Kuan Zhang,Chengliang Chai,Rui Wang,Xinlin Zhuang,Tianyi
 Bai,Qiu Jiantao,Lei Cao,Ju Fan,Ye Yuan,Guoren Wang,Conghui He
Harnessing Diversity for Important Data Selection in Pretraining Large Language
Models
Data selection is of great significance in pretraining large language models, g
iven the variation in quality within the large-scale available training corpora
.
To achieve this, researchers are currently investigating the use of data influen
ce to measure the importance of data instances, $i.e.,$ a high influence score i
ndicates that incorporating this instance to the training set is likely to enhan
ce the model performance. Consequently, they select the top-k instances with t
he highest scores. However, this approach has several limitations.

(1) Calculating the accurate influence of all available data is time-consuming.
(2) The selected data instances are not diverse enough, which may hinder the pre trained model's ability to generalize effectively to various downstream tasks.
In this paper, we introduce $\texttt{Quad}$, a data selection approach that cons iders both quality and diversity by using data influence to achieve state-of-the -art pretraining results.
To compute the influence ($i.e.,$ the quality) more accurately and efficiently, we incorporate the attention layers to capture more semantic details, which can be accelerated through the Kronecker product.
For the diversity, $\texttt{Quad}$ clusters the dataset into similar data instan ces within each cluster and diverse instances across different clusters. For eac h cluster, if we opt to select data from it, we take some samples to evaluate th e influence to prevent processing all instances. Overall, we favor clusters with highly influential instances (ensuring high quality) or clusters that have been selected less frequently (ensuring diversity), thereby well balancing between q uality and diversity. Experiments on Slimpajama and FineWeb over 7B large langu age models demonstrate that $\texttt{Quad}$ significantly outperforms other data selection methods with a low FLOPs consumption. Further analysis also validates the effectiveness of our influence calculation.
****
Liliang Ren,Yang Liu,Yadong Lu,yelong shen,Chen Liang,Weizhu Chen
Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling
Efficiently modeling sequences with infinite context length has long been a chal lenging problem. Previous approaches have either suffered from quadratic computa tional complexity or limited extrapolation ability in length generalization. In this
work, we present Samba, a simple hybrid architecture that layer-wise combines Mamba, a selective State Space Model (SSM), with Sliding Window Attention (SWA). Samba selectively compresses a given sequence into recurrent hidden states while still maintaining the ability to precisely recall recent memories w ith the
attention mechanism. We scale Samba up to 3.8B parameters with 3.2T training tokens and demonstrate that it significantly outperforms state-of-the-art models across a variety of benchmarks. Pretrained on sequences of 4K length, Samba shows improved perplexity in context lengths of up to 1M in zero-shot. When finetuned on 4K-length sequences, Samba efficiently extrapolates to a 256K conte xt length with perfect memory recall on the Passkey Retrieval task, and exhibits superior retrieval extrapolation on the challenging Phonebook task compared to full-attention models. As a linear-time sequence model, Samba achieves a 3.73× higher throughput compared to Transformers with grouped-query attention for user prompts of 128K length, and a 3.64× speedup when generating 64K tokens with unlimited streaming.
****
Zhiqi Bu,Shiyun Xu
Gradient descent with generalized Newton’s method
We propose the generalized Newton's method (GeN) --- a Hessian-informed approach that applies to any optimizer such as SGD and Adam, and covers the Newton-Raphs on method as a sub-case. Our method automatically and dynamically selects the le arning rate that accelerates the convergence, without the intensive tuning of th e learning rate scheduler. In practice, our method is easily implementable, sinc e it only requires additional forward passes with almost zero computational over head (in terms of training time and memory cost), if the overhead is amortized o ver many iterations. We present extensive experiments on language and vision tas ks (e.g. GPT and ResNet) to showcase that GeN optimizers match the state-of-the- art performance, which was achieved with carefully tuned learning rate scheduler s.
****
Paul Garnier,Vincent Lannelongue,Jonathan Viquerat,Elie Hachem
MeshMask: Physics-Based Simulations with Masked Graph Neural Networks

We introduce a novel masked pre-training technique for graph neural networks (GN
Ns) applied to computational fluid dynamics (CFD) problems. By randomly masking
up to 40\% of input mesh nodes during pre-training, we force the model to learn
robust representations of complex fluid dynamics. We pair this masking strategy
with an asymmetric encoder-decoder architecture and gated multi-layer perceptron
s to further enhance performance. The proposed method achieves state-of-the-art
results on seven CFD datasets, including a new challenging dataset of 3D intracr
anial aneurysm simulations with over 250,000 nodes per mesh. Moreover, it signif
icantly improves model performance and training efficiency across such diverse r
ange of fluid simulation tasks. We demonstrate improvements of up to 60\% in lon
g-term prediction accuracy compared to previous best models, while maintaining s
imilar computational costs. Notably, our approach enables effective pre-training
 on multiple datasets simultaneously, significantly reducing the time and data r
equired to achieve high performance on new tasks.
Through extensive ablation studies, we provide insights into the optimal masking
 ratio, architectural choices, and training strategies.
****

Yanjun Zhao,Sizhe Dang,Haishan Ye,Guang Dai,Yi Qian,Ivor Tsang
Second-Order Fine-Tuning without Pain for LLMs: A Hessian Informed Zeroth-Order
Optimizer
Fine-tuning large language models (LLMs) is necessary for specific downstream ta
sks, but classic first-order optimizer entails prohibitive GPU memory because of
 the back propagation. Recent works such as MeZO have turned to zeroth-order opt
imizers for fine-tuning, which reduce substantial memory by using two forward pa
sses. However, heterogeneous curvatures across different parameter dimensions in
 LLMs often cause model convergence instability or even failure. In this work, w
e propose HiZOO, a diagonal Hessian informed Zeroth-Order Optimizer , which is t
he first work to leverage the diagonal Hessian to enhance ZOO for fine-tuning LL
Ms. We provide theoretical proof for HiZOO and visualize the optimization trajec
tories on test functions to illustrate how it improves convergence in handling h
eterogeneous curvatures. Extensive experiments on various models (RoBERTa, OPT,
Phi-2 and LLama3, with 350M$\sim$66B parameters) indicate that HiZOO significant
ly reduces training steps and enhances model accuracy, while keeping the memory
advantage of ZOO. For example, on SST2 task HiZOO achieves $8\times$ speedup and
 better accuracy over MeZO across different models. We also propose HiZOO-L, whi
ch reduces the Hessian memory cost to 10\% of the MeZO, while maintaining almost
 same performance. Compared with ZO-Adam, HiZOO-L achieves a 4.3\% improvement,
just using 50\% of the GPU memory. Code is available at https://anonymous.4open.
science/r/HiZOO-27F8.
****

Prajwal Koirala,Zhanhong Jiang,Soumik Sarkar,Cody Fleming
Latent Safety-Constrained Policy Approach for Safe Offline Reinforcement Learnin
g
In safe offline reinforcement learning, the objective is to develop a policy tha
t maximizes cumulative rewards while strictly adhering to safety constraints, ut
ilizing only offline data. Traditional methods often face difficulties in balanc
ing these constraints, leading to either diminished performance or increased saf
ety risks. We address these issues with a novel approach that begins by learning
 a conservatively safe policy through the use of Conditional Variational Autoenc
oders, which model the latent safety constraints. Subsequently, we frame this as
 a Constrained Reward-Return Maximization problem, wherein the policy aims to op
timize rewards while complying with the inferred latent safety constraints. This
 is achieved by training an encoder with a reward-Advantage Weighted Regression
objective within the latent constraint space. Our methodology is supported by th
eoretical analysis, including bounds on policy performance and sample complexity
. Extensive empirical evaluation on benchmark datasets, including challenging au
tonomous driving scenarios, demonstrates that our approach not only maintains sa
fety compliance but also excels in cumulative reward optimization, surpassing ex
isting methods. Additional visualizations provide further insights into the effe
ctiveness and underlying mechanisms of our approach.

****
Asmar Nadeem,Faegheh Sardari,Robert Dawes,Syed Sameed Husain,Adrian Hilton,Armin Mustafa
NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative
Existing video captioning benchmarks and models lack causal-temporal narrative, which is sequences of events linked through cause and effect, unfolding over time and driven by characters or agents. This lack of narrative restricts models' ability to generate text descriptions that capture the causal and temporal dynamics inherent in video content. To address this gap, we propose NarrativeBridge, an approach comprising of: (1) a novel Causal-Temporal Narrative (CTN) captions benchmark generated using a large language model and few-shot prompting, explicitly encoding cause-effect temporal relationships in video descriptions; and (2) a Cause-Effect Network (CEN) with separate encoders for capturing cause and effect dynamics, enabling effective learning and generation of captions with causal-temporal narrative. Extensive experiments demonstrate that CEN significantly outperforms state-of-the-art models in articulating the causal and temporal aspects of video content: 17.88 and 17.44 CIDEr on the MSVD-CTN and MSRVTT-CTN datasets, respectively. Cross-dataset evaluations further showcase CEN's strong generalization capabilities. The proposed framework understands and generates nuanced text descriptions with intricate causal-temporal narrative structures present in videos, addressing a critical limitation in video captioning. For project details, visit https://narrativebridge.github.io/.
****
Junseo Park,Hyeryung Jang
I^2AM: Interpreting Image-to-Image Latent Diffusion Models via Bi-Attribution Maps
Large-scale diffusion models have made significant advances in image generation, particularly through cross-attention mechanisms. While cross-attention has been well-studied in text-to-image tasks, their interpretability in image-to-image (I2I) diffusion models remains underexplored. This paper introduces Image-to-Image Attribution Maps $(\textbf{I}^2\textbf{AM})$, a method that enhances the interpretability of I2I models by visualizing bidirectional attribution maps, from the reference image to the generated image and vice versa. I^2AM aggregates cross-attention scores across time steps, attention heads, and layers, offering insights into how critical features are transferred between images. We demonstrate the effectiveness of I^2AM across object detection, inpainting, and super-resolution tasks. Our results demonstrate that I^2AM successfully identifies key regions responsible for generating the output, even in complex scenes. Additionally, we introduce the Inpainting Mask Attention Consistency Score (IMACS) as a novel evaluation metric to assess the alignment between attribution maps and inpainting masks, which correlates strongly with existing performance metrics. Through extensive experiments, we show that I^2AM enables model debugging and refinement, providing practical tools for improving I2I model's performance and interpretability.
****
Chen-Yu Liu,Chao-Han Huck Yang,Hsi-Sheng Goan,Min-Hsiu Hsieh
A Quantum Circuit-Based Compression Perspective for Parameter-Efficient Learning
Quantum-centric supercomputing presents a compelling framework for large-scale hybrid quantum-classical tasks. Although quantum machine learning (QML) offers theoretical benefits in various applications, challenges such as large-size data encoding in the input stage and the reliance on quantum resources in the inference stage limit its practicality for tasks like fine-tuning large language models (LLMs). Quantum parameter generation, a novel approach of QML, addresses these limitations by using quantum neural networks (QNNs) to generate classical model weights (parameters) exclusively during training, thereby decoupling inference from quantum hardware. In this work, we introduce Quantum Parameter Adaptation (QPA) in the framework of quantum parameter generation, which integrates QNNs with a classical multi-layer perceptron mapping model to generate parameters for fine-tuning methods. Using Gemma-2 and GPT-2 as case studies, QPA demonstrates significant parameter reduction for parameter-efficient fine-tuning methods, such as

Low-Rank Adaptation (LoRA), while maintaining comparable or improved performance in text generation tasks. Specifically, QPA reduces the number of parameters to 52.06% of the original LoRA for GPT-2 with a slight performance gain of 0.75%, and to 16.84% for Gemma-2, with a marginal performance improvement of 0.07%. These results highlight QPA's ability to achieve efficient parameter reduction without sacrificing performance in the quantum parameter generation framework. This work showcases the potential of quantum-enhanced parameter reduction, offering a scalable quantum-classical solution for fine-tuning LLMs while preserving the feasibility of inference on classical hardware.
****

Dang Nguyen,Wenhan Yang,Rathul Anand,Yu Yang,Baharan Mirzasoleiman
Mini-batch Coresets for Memory-efficient Language Model Training on Data Mixtures
Training with larger mini-batches improves the convergence rate and can yield superior performance. However, training with large mini-batches becomes prohibitive for Large Language Models (LLMs), due to the large GPU memory requirement. To address this problem, an effective approach is finding small mini-batch coresets that closely match the gradient of larger mini-batches. However, this approach becomes infeasible and ineffective for LLMs, due to the highly imbalanced mixture of sources in language data, use of the Adam optimizer, and the very large gradient dimensionality of LLMs. In this work, we address the above challenges by proposing *Coresets for Training LLMs* (CoLM). First, we show that mini-batch coresets found by gradient matching do not contain representative examples of the small sources w.h.p., and thus including all examples of the small sources in the mini-batch coresets is crucial for optimal performance. Second, we normalize the gradients by their historical exponential to find mini-batch coresets for training with Adam. Finally, we leverage zeroth-order methods to find smooth gradient of the last *V*-projection matrix and sparsify it to keep the dimensions with the largest normalized gradient magnitude. We apply CoLM to fine-tuning Phi-2, Phi-3, Zephyr, and Llama-3 models with LoRA on MathInstruct and SuperGLUE benchmark. Remarkably, CoLM reduces the memory requirement of fine-tuning by 2x and even outperforms training with 4x larger mini-batches. Moreover, CoLM seamlessly integrates with existing memory-efficient training methods like LoRA, further reducing the memory requirements of training LLMs.
****

Francesco S. Carzaniga,Gary Tom Hoppeler,Michael Hersche,Kaspar Schindler,Abbas Rahimi
The Case for Cleaner Biosignals: High-fidelity Neural Compressor Enables Transfer from Cleaner iEEG to Noisier EEG
All data modalities are not created equal, even when the signal they measure comes from the same source. In the case of the brain, two of the most important data modalities are the scalp electroencephalogram (EEG), and the intracranial electroencephalogram (iEEG). iEEG benefits from a higher signal-to-noise ratio (SNR), as it measures the electrical activity directly in the brain, while EEG is noisier and has lower spatial and temporal resolutions. Nonetheless, both EEG and iEEG are important sources of data for human neurology, from healthcare to brain–machine interfaces. They are used by human experts, supported by deep learning (DL) models, to accomplish a variety of tasks, such as seizure detection and motor imagery classification. Although the differences between EEG and iEEG are well understood by human experts, the performance of DL models across these two modalities remains under-explored. To help characterize the importance of clean data on the performance of DL models, we propose BrainCodec, a high-fidelity EEG and iEEG neural compressor. We find that training BrainCodec on iEEG and then transferring to EEG yields higher reconstruction quality than training on EEG directly. In addition, we also find that training BrainCodec on both EEG and iEEG improves fidelity when reconstructing EEG. Our work indicates that data sources with higher SNR, such as iEEG, provide better performance across the board also in the medical time-series domain. This finding is consistent with reports coming from natural language processing, where clean data sources appear to have an outsized effect on the performance of the DL model overall. BrainCodec also achieves u

p to a 64x compression on iEEG and EEG without a notable decrease in quality. Br
ainCodec markedly surpasses current state-of-the-art compression models both in
final compression ratio and in reconstruction fidelity. We also evaluate the fid
elity of the compressed signals objectively on a seizure detection and a motor i
magery task performed by standard DL models. Here, we find that BrainCodec achie
ves a reconstruction fidelity high enough to ensure no performance degradation o
n the downstream tasks. Finally, we collect the subjective assessment of an expe
rt neurologist, that confirms the high reconstruction quality of BrainCodec in a
 realistic scenario. The code is available at https://github.com/IBM/eeg-ieeg-br
ain-compressor.
****

Xinyan Chen,Jianfei Yang
X-Fi: A Modality-Invariant Foundation Model for Multimodal Human Sensing
Human sensing, which employs various sensors and advanced deep learning technolo
gies to accurately capture and interpret human body information, has significant
ly impacted fields like public security and robotics. However, current human sen
sing primarily depends on modalities such as cameras and LiDAR, each of which ha
s its own strengths and limitations. Furthermore, existing multimodal fusion sol
utions are typically designed for fixed modality combinations, requiring extensi
ve retraining when modalities are added or removed for diverse scenarios. In thi
s paper, we propose a modality-invariant foundation model for all modalities, X-
Fi, to address these issues. X-Fi enables the independent or combinatory use of
sensor modalities without additional training by utilizing a transformer structu
re to accommodate variable input sizes and incorporating a novel "X-fusion" mech
anism to preserve modality-specific features during multimodal integration. This
 approach not only enhances adaptability but also facilitates the learning of co
mplementary features across modalities. Extensive experiments conducted on the M
M-Fi and XRF55 datasets, employing six distinct modalities, demonstrate that X-F
i achieves state-of-the-art performance in human pose estimation (HPE) and human
 activity recognition (HAR) tasks. The findings indicate that our proposed model
 can efficiently support a wide range of human sensing applications, ultimately
contributing to the evolution of scalable, multimodal sensing technologies.
****

Seyedmorteza Sadat,Manuel Kansy,Otmar Hilliges,Romann M. Weber
No Training, No Problem: Rethinking Classifier-Free Guidance for Diffusion Model
s
Classifier-free guidance (CFG) has become the standard method for enhancing the
quality of conditional diffusion models. However, employing CFG requires either
training an unconditional model alongside the main diffusion model or modifying
the training procedure by periodically inserting a null condition. There is als
o no clear extension of CFG to unconditional models. In this paper, we revisit t
he core principles of CFG and introduce a new method, independent condition guid
ance (ICG), which provides the benefits of CFG without the need for any special
training procedures. Our approach streamlines the training process of conditiona
l diffusion models and can also be applied during inference on any pre-trained c
onditional model. Additionally, by leveraging the time-step information encoded
in all diffusion networks, we propose an extension of CFG, called time-step guid
ance (TSG), which can be applied to *any* diffusion model, including uncondition
al ones. Our guidance techniques are easy to implement and have the same samplin
g cost as CFG. Through extensive experiments, we demonstrate that ICG matches th
e performance of standard CFG across various conditional diffusion models. Moreo
ver, we show that TSG improves generation quality in a manner similar to CFG, wi
thout relying on any conditional information.
****

Wonjoon Chang,Myeongjin Lee,Jaesik Choi
Rethinking Shapley Value for Negative Interactions in Non-convex Games
We study causal interactions for payoff allocation in cooperative game theory, i
ncluding quantifying feature attribution for deep learning models. Most feature
attribution methods mainly stem from the criteria of the Shapley value, which as
signs fair payoffs to players based on their expected contribution in a cooperat

ive game. However, interactions between players in the game do not explicitly ap
pear in the original formulation of the Shapley value. In this work, we reformul
ate the Shapley value to clarify the role of interactions and discuss implicit a
ssumptions from a game-theoretical perspective. Our theoretical analysis demonst
rates that when negative interactions exist—common in deep learning models—the e
fficiency axiom can lead to the undervaluation of attributions or payoffs. We su
ggest a new allocation rule that decomposes contributions into interactions and
aggregates positive parts for non-convex games. Furthermore, we propose an appro
ximation algorithm to reduce the cost of interaction computation which can be ap
plied to differentiable functions such as deep learning models. Our approach mit
igates counterintuitive attribution outcomes observed in existing methods, ensur
ing that features critical to a model’s decision receive appropriate attribution
.
****

Xiaoyu Yang,Jie Lu,En Yu

Adapting Multi-modal Large Language Model to Concept Drift From Pre-training Onw
ards

Multi-modal Large Language Models (MLLMs) frequently face challenges from concep
t drift when dealing with real-world streaming data, wherein distributions chang
e unpredictably. This mainly includes gradual drift due to long-tailed data and
sudden drift from Out-Of-Distribution (OOD) data, both of which have increasingl
y drawn the attention of the research community. While these issues have been ex
tensively studied in the individual domain of vision or language, their impacts
on MLLMs in concept drift settings remain largely underexplored. In this paper,
we reveal the susceptibility and vulnerability of Vision-Language (VL) models to
 significant biases arising from gradual drift and sudden drift, particularly in
 the pre-training. To effectively address these challenges, we propose a unified
 framework that extends concept drift theory to the multi-modal domain, enhancin
g the adaptability of the VL model to unpredictable distribution changes. Additi
onally, a T-distribution based drift adapter is proposed to effectively mitigate
 the bias induced by the gradual drift, which also facilitates the model in dist
inguishing sudden distribution changes through explicit distribution modeling. E
xtensive experiments demonstrate our method enhances the efficiency and accuracy
 of image-text alignment in the pre-training of VL models, particularly in the c
oncept drift scenario. Moreover, various downstream tasks exhibit significant im
provements in our model's ability to adapt to the long-tailed open world. Furthe
rmore, we create a set of multi-modal datasets called OpenMMlo, specifically tai
lored for the long-tailed open-world setting, to validate our findings. To foste
r the development of the multi-modal community, we have made both OpenMMlo datas
ets and our code publicly available at: https://github.com/XiaoyuYoung/ConceptDr
iftMLLMs.
****

Juan A. Rodriguez,Xiangru Jian,Siba Smarak Panigrahi,Tianyu Zhang,Aarash Feizi,A
bhay Puri,Akshay Kalkunte Suresh,François Savard,Ahmed Masry,Shravan Nayak,Rabiu
l Awal,Mahsa Massoud,Amirhossein Abaskohi,Zichao Li,Suyuchen Wang,Pierre-Andre N
oel,Mats Leon Richter,Saverio Vadacchino,Shubham Agarwal,Sanket Biswas,Sara Shan
ian,Ying Zhang,Sathwik Tejaswi Madhusudhan,Joao Monteiro,Krishnamurthy Dj Dvijot
ham,Torsten Scholak,Nicolas Chapados,Sepideh Kharaghani,Sean Hughes,M. Özsu,Siva
 Reddy,Marco Pedersoli,Yoshua Bengio,Christopher Pal,Issam H. Laradji,Spandana G
ella,Perouz Taslakian,David Vazquez,Sai Rajeswar

BigDocs: An Open Dataset for Training Multimodal Models on Document and Code Tas
ks

Multimodal AI has the potential to significantly enhance document-understanding
tasks, such as processing receipts, understanding workflows, extracting data fro
m documents, and summarizing reports. Code generation tasks that require long-st
ructured outputs can also be enhanced by multimodality. Despite this, their use
in commercial applications is often limited due to limited access to relevant tr
aining data and restrictive licensing, which hinders open access. To address the
se limitations, we introduce BigDocs-7.5M, a high-quality, open-access dataset c
omprising 7.5 million multimodal documents across 30 tasks. We use an efficient

data curation process to ensure that our data is high quality and license-permis
sive. Our process emphasizes accountability, responsibility, and transparency th
rough filtering rules, traceable metadata, and careful content analysis. Additio
nally, we introduce BigDocs-Bench,, a benchmark suite with 10 novel tasks where
we carefully create datasets that reflect real-world use cases involving reasoni
ng over Graphical User Interfaces (GUI) and code generation from images. Our exp
eriments show that training with BigDocs-Bench, improves average performance up
to 25.8% over closed-source GPT-4o in document reasoning and structured output t
asks such as Screenshot2HTML or Image2Latex generation. Finally, human evaluatio
ns revealed that participants preferred the outputs from models trained with Big
Docs over those from GPT-4o. This suggests that BigDocs can help both academics
and the open-source community utilize and improve AI tools to enhance multimodal
 capabilities and document reasoning.
****
Ruijie Zheng,Yongyuan Liang,Shuaiyi Huang,Jianfeng Gao,Hal Daumé III,Andrey Kolo
bov,Furong Huang,Jianwei Yang
TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for General
ist Robotic Policies
Although large vision-language-action (VLA) models pretrained on extensive robot
 datasets offer promising generalist policies for robotic learning, they still s
truggle with spatial-temporal dynamics in interactive robotics, making them less
 effective in handling complex tasks, such as manipulation. In this work, we int
roduce visual trace prompting, a simple yet effective approach to facilitate VLA
 models' spatial-temporal awareness for action prediction by encoding state-acti
on trajectories visually. We develop a new TraceVLA model by finetuning
OpenVLA on our own collected dataset of 150K robot manipulation trajectories usi
ng visual trace prompting. Evaluations of TraceVLA across 137 configurations in
SimplerEnv and 4 tasks on a physical WidowX robot demonstrate state-of-the-art p
erformance, outperforming OpenVLA by 10% on SimplerEnv and 3.5x on real-robot ta
sks and exhibiting robust generalization across diverse embodiments and scenario
s. To further validate the effectiveness and generality of our method, we presen
t a compact VLA model based on 4B Phi-3-Vision, pretrained on the Open-X-Embodim
ent and finetuned on our dataset, rivals the 7B OpenVLA baseline while significa
ntly improving inference efficiency.
****
Ziyang Zheng,Shan Huang,Jianyuan Zhong,Zhengyuan Shi,Guohao Dai,Ningyi Xu,Qiang
Xu
DeepGate4: Efficient and Effective Representation Learning for Circuit Design at
 Scale
Circuit representation learning has become pivotal in electronic design automati
on, enabling critical tasks such as testability analysis, logic reasoning, power
 estimation, and SAT solving. However, existing models face significant challeng
es in scaling to large circuits due to limitations like over-squashing in graph
neural networks and the quadratic complexity of transformer-based models. To add
ress these issues, we introduce \textbf{DeepGate4}, a scalable and efficient gra
ph transformer specifically designed for large-scale circuits. DeepGate4 incorpo
rates several key innovations: (1) an update strategy tailored for circuit graph
s, which reduce memory complexity to sub-linear and is adaptable to any graph tr
ansformer; (2) a GAT-based sparse transformer with global and local structural e
ncodings for AIGs; and (3) an inference acceleration CUDA kernel that fully expl
oit the unique sparsity patterns of AIGs. Our extensive experiments on the ITC99
 and EPFL benchmarks show that DeepGate4 significantly surpasses state-of-the-ar
t methods, achieving 15.5\% and 31.1\% performance improvements over the next-be
st models. Furthermore, the Fused-DeepGate4 variant reduces runtime by 35.1\% an
d memory usage by 46.8\%, making it highly efficient for large-scale circuit ana
lysis. These results demonstrate the potential of DeepGate4 to handle complex ED
A tasks while offering superior scalability and efficiency.
****
Yangming Li,Max Ruiz Luyten,Mihaela van der Schaar
Risk-Sensitive Diffusion: Robustly Optimizing Diffusion Models with Noisy Sample

s
Diffusion models are mainly studied on image data. However, non-image data (e.g. , tabular data) are also prevalent in real applications and tend to be noisy due to some inevitable factors in the stage of data collection, degrading the generation quality of diffusion models. In this paper, we consider a novel problem setting where every collected sample is paired with a vector indicating the data quality: risk vector. This setting applies to many scenarios involving noisy data and we propose risk-sensitive SDE, a type of stochastic differential equation (SDE) parameterized by the risk vector, to address it. With some proper coefficients, risk-sensitive SDE can minimize the negative effect of noisy samples on the optimization of diffusion models. We conduct systematic studies for both Gaussian and non-Gaussian noise distributions, providing analytical forms of risk-sensitive SDE. To verify the effectiveness of our method, we have conducted extensive experiments on multiple tabular and time-series datasets, showing that risk-sensitive SDE permits a robust optimization of diffusion models with noisy samples and significantly outperforms previous baselines.
****
Longtao Zheng,Zhiyuan Huang,Zhenghai Xue,Xinrun Wang,Bo An,Shuicheng YAN
AgentStudio: A Toolkit for Building General Virtual Agents
General virtual agents need to handle multimodal observations, master complex action spaces, and self-improve in dynamic, open-domain environments. However, existing environments are often domain-specific and require complex setups, which limits agent development and evaluation in real-world settings. As a result, current evaluations lack in-depth analyses that decompose fundamental agent capabilities. We introduce AgentStudio, a trinity of environments, tools, and benchmarks to address these issues. AgentStudio provides a lightweight, interactive environment with highly generic observation and action spaces, e.g., video observations and GUI/API actions. It integrates tools for creating online benchmark tasks, annotating GUI elements, and labeling actions in videos. Based on our environment and tools, we curate an online task suite that benchmarks both GUI interactions and function calling with efficient auto-evaluation. We also reorganize existing datasets and collect new ones using our tools to establish three datasets: GroundUI, IDMBench, and CriticBench. These datasets evaluate fundamental agent abilities, including GUI grounding, learning from videos, and success detection, pointing to the desiderata for robust, general, and open-ended virtual agents.
****
Guipeng Lan,Shuai Xiao,Meng Xi,Jiabao Wen,Jiachen Yang
Inner Information Analysis Algorithm for Deep Neural Network based on Community
Deep learning has achieved advancements across a variety of forefront fields. However, its inherent 'black box' characteristic poses challenges to the comprehension and trustworthiness of the decision-making processes within neural networks. To mitigate these challenges, we introduce InnerSightNet, an inner information analysis algorithm designed to illuminate the inner workings of deep neural networks through the perspectives of community. This approach is aimed at deciphering the intricate patterns of neurons within deep neural networks, thereby shedding light on the networks' information processing and decision-making pathways. InnerSightNet operates in three primary phases, 'neuronization-aggregation-evaluation'. Initially, it transforms learnable units into a structured network of neurons. Subsequently, these neurons are aggregated into distinct communities according to representation attributes. The final phase involves the evaluation of these communities' roles and functionalities, to unpick the information flow and decision-making. By transcending focus on single-layer or individual neuron, InnerSightNet broadens the horizon for deep neural network interpretation. InnerSightNet offers a unique vantage point, enabling insights into the collective behavior of communities within the overarching architecture, thereby enhancing transparency and trust in deep learning systems.
****
Muhan Li,Lingji Kong,Sam Kriegman
Generating Freeform Endoskeletal Robots
The automatic design of embodied agents (e.g. robots) has existed for 31 years a

nd is experiencing a renaissance of interest in the literature. To date however, the field has remained narrowly focused on two kinds of anatomically simple robots: (1) fully rigid, jointed bodies; and (2) fully soft, jointless bodies. Here we bridge these two extremes with the open ended creation of terrestrial endoskeletal robots: deformable soft bodies that leverage jointed internal skeletons to move efficiently across land. Simultaneous de novo generation of external and internal structures is achieved by (i) modeling 3D endoskeletal body plans as integrated collections of elastic and rigid cells that directly attach to form soft tissues anchored to compound rigid bodies; (ii) encoding these discrete mechanical subsystems into a continuous yet coherent latent embedding; (iii) optimizing the sensorimotor coordination of each decoded design using model-free reinforcement learning; and (iv) navigating this smooth yet highly non-convex latent manifold using evolutionary strategies. This yields an endless stream of novel species of ``higher robots'' that, like all higher animals, harness the mechanical advantages of both elastic tissues and skeletal levers for terrestrial travel. It also provides a plug-and-play experimental platform for benchmarking evolutionary design and representation learning algorithms in complex hierarchical embodied systems.
****
Haorui Wang,Marta Skreta,Cher Tian Ser,Wenhao Gao,Lingkai Kong,Felix Strieth-Kalthoff,Chenru Duan,Yuchen Zhuang,Yue Yu,Yanqiao Zhu,Yuanqi Du,Alan Aspuru-Guzik,Kirill Neklyudov,Chao Zhang
Efficient Evolutionary Search Over Chemical Space with Large Language Models
Molecular discovery, when formulated as an optimization problem, presents significant computational challenges because optimization objectives can be non-differentiable. Evolutionary Algorithms (EAs), often used to optimize black-box objectives in molecular discovery, traverse chemical space by performing random mutations and crossovers, leading to a large number of expensive objective evaluations. In this work, we ameliorate this shortcoming by incorporating chemistry-aware Large Language Models (LLMs) into EAs. Namely, we redesign crossover and mutation operations in EAs using LLMs trained on large corpora of chemical information. We perform extensive empirical studies on both commercial and open-source models on multiple tasks involving property optimization, molecular rediscovery, and structure-based drug design, demonstrating that the joint usage of LLMs with EAs yields superior performance over all baseline models across single- and multi-objective settings. We demonstrate that our algorithm improves both the quality of the final solution and convergence speed, thereby reducing the number of required objective evaluations.
****
Wenlong Deng,Yize Zhao,Vala Vakilian,Minghui Chen,Xiaoxiao Li,Christos Thrampoulidis
DARE the Extreme: Revisiting Delta-Parameter Pruning For Fine-Tuned Models
Storing open-source fine-tuned models separately introduces redundancy and increases response times in applications utilizing multiple models. Delta-parameter pruning (DPP), particularly the random drop and rescale (DARE) method proposed by Yu et al., addresses this by pruning the majority of delta parameters—the differences between fine-tuned and pre-trained model weights—while typically maintaining minimal performance loss. However, DARE fails when either the pruning rate or the magnitude of the delta parameters is large. We highlight two key reasons for this failure: (1) an excessively large rescaling factor as pruning rates increase, and (2) high mean and variance in the delta parameters. To push DARE’s limits, we introduce DAREx (DARE the eXtreme), which features two algorithmic improvements: (1) DAREx-q, a rescaling factor modification that significantly boosts performance at high pruning rates (e.g., > 30% on COLA and SST2 for encoder models, with even greater gains in decoder models), and (2) DAREx-L2, which combines DARE with AdamR, an in-training method that applies appropriate delta regularization before DPP. We also demonstrate that DAREx-q can be seamlessly combined with vanilla parameter-efficient fine-tuning techniques like LoRA and can facilitate structural DPP. Additionally, we revisit the application of importance-based pruning techniques within DPP, demonstrating that they outperform random-based m

ethods when delta parameters are large. Through this comprehensive study, we dev
elop a pipeline for selecting the most appropriate DPP method under various prac
tical scenarios.
****

Alexey Bochkovskiy,Amaël Delaunoy,Hugo Germain,Marcel Santos,Yichao Zhou,Stephan
 Richter,Vladlen Koltun
Depth Pro: Sharp Monocular Metric Depth in Less Than a Second
We present a foundation model for zero-shot metric monocular depth estimation. O
ur model, Depth Pro, synthesizes high-resolution depth maps with unparalleled sh
arpness and high-frequency details. The predictions are metric, with absolute sc
ale, without relying on the availability of metadata such as camera intrinsics.
And the model is fast, producing a 2.25-megapixel depth map in 0.3 seconds on a
standard GPU. These characteristics are enabled by a number of technical contrib
utions, including an efficient multi-scale vision transformer for dense predicti
on, a training protocol that combines real and synthetic datasets to achieve hig
h metric accuracy alongside fine boundary tracing, dedicated evaluation metrics
for boundary accuracy in estimated depth maps, and state-of-the-art focal length
 estimation from a single image. Extensive experiments analyze specific design c
hoices and demonstrate that Depth Pro outperforms prior work along multiple dime
nsions. We release code & weights at https://github.com/apple/ml-depth-pro
****

Chenxi Zheng,Yihong Lin,Bangzhen Liu,Xuemiao Xu,Yongwei Nie,Shengfeng He
RecDreamer: Consistent Text-to-3D Generation via Uniform Score Distillation
Current text-to-3D generation methods based on score distillation often suffer f
rom geometric inconsistencies, leading to repeated patterns across different pos
es of 3D assets. This issue, known as the Multi-Face Janus problem, arises becau
se existing methods struggle to maintain consistency across varying poses and ar
e biased toward a canonical pose. While recent work has improved pose control an
d approximation, these efforts are still limited by this inherent bias, which sk
ews the guidance during generation.
To address this, we propose a solution called RecDreamer, which reshapes the und
erlying data distribution to achieve more consistent pose representation. The co
re idea behind our method is to rectify the prior distribution, ensuring that po
se variation is uniformly distributed rather than biased toward a canonical form
. By modifying the prescribed distribution through an auxiliary function, we can
 reconstruct the density of the distribution to ensure compliance with specific
marginal constraints. In particular, we ensure that the marginal distribution of
 poses follows a uniform distribution, thereby eliminating the biases introduced
 by the prior knowledge.
We incorporate this rectified data distribution into existing score distillation
 algorithms, a process we refer to as uniform score distillation. To efficiently
 compute the posterior distribution required for the auxiliary function, RecDrea
mer introduces a training-free classifier that estimates pose categories in a pl
ug-and-play manner. Additionally, we utilize various approximation techniques fo
r noisy states, significantly improving system performance.
Our experimental results demonstrate that RecDreamer effectively mitigates the M
ulti-Face Janus problem, leading to more consistent 3D asset generation across d
ifferent poses.
****

Xingyi Zhou,Anurag Arnab,Chen Sun,Cordelia Schmid
Dense Video Object Captioning from Disjoint Supervision
We propose a new task and model for dense video object captioning -- detecting,
tracking and captioning trajectories of objects in a video. This task unifies sp
atial and temporal localization in video, whilst also requiring fine-grained vis
ual understanding that is best described by natural language. We propose a unifi
ed model, and demonstrate how our end-to-end approach is more accurate and tempo
rally coherent than a multi-stage pipeline combining state-of-the-art detection,
 tracking, and captioning models. Moreover, we propose a training strategy based
 on a mixture of disjoint tasks, which allows us to leverage diverse, large-scal
e datasets which supervise different parts of our model. Although each pretraini

ng task only provides weak supervision, they are complementary and, when combine d, result in noteworthy zero-shot ability and serve as strong initialization for additional finetuning to further improve accuracy. We carefully design new metr ics capturing all components of our task, and show how we can repurpose existing video grounding datasets (e.g. VidSTG and VLN) for our new task. We show that o ur model improves upon a number of strong baselines for this new task. Furthermo re, we can apply our model to the task of spatial grounding, outperforming prior state-of-the-art on VidSTG and VLN, without explicitly training for it. Our cod e is available at https://github.com/google-research/scenic.
****

Taoran Fang,Tianhong Gao,Chunping Wang,YihaoShang,Wei Chow,Lei CHEN,Yang Yang
KAA: Kolmogorov-Arnold Attention for Enhancing Attentive Graph Neural Networks
Graph neural networks (GNNs) with attention mechanisms, often referred to as att entive GNNs, have emerged as a prominent paradigm in advanced GNN models in rece nt years. However, our understanding of the critical process of scoring neighbor nodes remains limited, leading to the underperformance of many existing attenti ve GNNs. In this paper, we unify the scoring functions of current attentive GNNs and propose Kolmogorov-Arnold Attention (KAA), which integrates the Kolmogorov-Arnold Network (KAN) architecture into the scoring process. KAA enhances the per formance of scoring functions across the board and can be applied to nearly all existing attentive GNNs. To compare the expressive power of KAA with other scori ng functions, we introduce Maximum Ranking Distance (MRD) to quantitatively esti mate their upper bounds in ranking errors for node importance. Our analysis reve als that, under limited parameters and constraints on width and depth, both line ar transformation-based and MLP-based scoring functions exhibit finite expressiv e power. In contrast, our proposed KAA, even with a single-layer KAN parameteriz ed by zero-order B-spline functions, demonstrates nearly infinite expressive pow er. Extensive experiments on both node-level and graph-level tasks using various backbone models show that KAA-enhanced scoring functions consistently outperfor m their original counterparts, achieving performance improvements of over 20% in some cases.
****

Runqi Lin,Bo Han,Fengwang Li,Tongliang Liu
Understanding and Enhancing the Transferability of Jailbreaking Attacks
Jailbreaking attacks can effectively manipulate open-source large language model s (LLMs) to produce harmful responses. However, these attacks exhibit limited tr ansferability, failing to disrupt proprietary LLMs consistently. To reliably ide ntify vulnerabilities in proprietary LLMs, this work investigates the transferab ility of jailbreaking attacks by analysing their impact on the model's intent pe rception. By incorporating adversarial sequences, these attacks can redirect the source LLM's focus away from malicious-intent tokens in the original input, the reby obstructing the model's intent recognition and eliciting harmful responses. Nevertheless, these adversarial sequences fail to mislead the target LLM's inte nt perception, allowing the target LLM to refocus on malicious-intent tokens and abstain from responding. Our analysis further reveals the inherent $\textit{dis tributional dependency}$ within the generated adversarial sequences, whose effec tiveness stems from overfitting the source LLM's parameters, resulting in limite d transferability to target LLMs. To this end, we propose the Perceived-importan ce Flatten (PiF) method, which uniformly disperses the model's focus across neut ral-intent tokens in the original input, thus obscuring malicious-intent tokens without relying on overfitted adversarial sequences. Extensive experiments demon strate that PiF provides an effective and efficient red-teaming evaluation for p roprietary LLMs.
****

Baolong Bi,Shenghua Liu,Yiwei Wang,Lingrui Mei,Junfeng Fang,Hongcheng Gao,Shiyu Ni,Xueqi Cheng
Is Factuality Enhancement a Free Lunch For LLMs? Better Factuality Can Lead to W orse Context-Faithfulness
As the modern tools of choice for text understanding and generation, large langu age models (LLMs) are expected to accurately output answers by leveraging the in

put context.
This requires LLMs to possess both context-faithfulness and factual accuracy.
While extensive efforts aim to reduce hallucinations through factuality enhancem ent methods, they also pose risks of hindering context-faithfulness, as factuali ty enhancement can lead LLMs to become overly confident in their parametric know ledge, causing them to overlook the relevant input context.
In this work, we argue that current factuality enhancement methods can significa ntly undermine the context-faithfulness of LLMs.
We first revisit the current factuality enhancement methods and evaluate their e ffectiveness in enhancing factual accuracy.
Next, we evaluate their performance on knowledge editing tasks to assess the pot ential impact on context-faithfulness.
The experimental results reveal that while these methods may yield inconsistent improvements in factual accuracy, they also cause a more severe decline in conte xt-faithfulness, with the largest decrease reaching a striking 69.7\%.
To explain these declines, we analyze the hidden states and logit distributions for the tokens representing new knowledge and parametric knowledge respectively, highlighting the limitations of current approaches.
Our finding highlights the complex trade-offs inherent in enhancing LLMs.
Therefore, we recommend that more research on LLMs' factuality enhancement make efforts to reduce the sacrifice of context-faithfulness.
****
Jiachen Lei,Julius Berner,Jiongxiao Wang,Zhongzhu Chen,Chaowei Xiao,Zhongjie Ba, Kui Ren,Jun Zhu,Anima Anandkumar
Robust Representation Consistency Model via Contrastive Denoising
Robustness is essential for deep neural networks, especially in security-sensiti ve applications. To this end, randomized smoothing provides theoretical guarante es for certifying robustness against adversarial perturbations. Recently, diffus ion models have been successfully employed for randomized smoothing to purify no ise-perturbed samples before making predictions with a standard classifier. Whil e these methods excel at small perturbation radii, they struggle with larger per turbations and incur a significant computational overhead during inference compa red to classical methods. To address this, we reformulate the generative modelin g task along the diffusion trajectories in pixel space as a discriminative task in the latent space. Specifically, we use instance discrimination to achieve con sistent representations along the trajectories by aligning temporally adjacent p oints. After fine-tuning based on the learned representations, our model enables implicit denoising-then-classification via a single prediction, substantially r educing inference costs. We conduct extensive experiments on various datasets an d achieve state-of-the-art performance with minimal computation budget during in ference. For example, our method outperforms the certified accuracy of diffusion -based methods on ImageNet across all perturbation radii by 5.3\% on average, wi th up to 11.6\% at larger radii, while reducing inference costs by 85x on averag e.
****
Yiding Jiang,Allan Zhou,Zhili Feng,Sadhika Malladi,J Zico Kolter
Adaptive Data Optimization: Dynamic Sample Selection with Scaling Laws
The composition of pretraining data is a key determinant of foundation models' p erformance, but there is no standard guideline for allocating a limited computat ional budget across different data sources. Most current approaches either rely on extensive experiments with smaller models or dynamic data adjustments that al so require proxy models, both of which significantly increase the workflow compl exity and computational overhead. In this paper, we introduce Adaptive Data Opti mization (ADO), an algorithm that optimizes data distributions in an online fash ion, concurrent with model training. Unlike existing techniques, ADO does not re quire external knowledge, proxy models, or modifications to the model update. In stead, ADO uses per-domain scaling laws to estimate the learning potential of ea ch domain during training and adjusts the data mixture accordingly, making it mo re scalable and easier to integrate. Experiments demonstrate that ADO can achiev e comparable or better performance than prior methods while maintaining computat

ional efficiency across different computation scales, offering a practical solut
ion for dynamically adjusting data distribution without sacrificing flexibility
or increasing costs. Beyond its practical benefits, ADO also provides a new pers
pective on data collection strategies via scaling laws.
****

Jingyun Xue,Hongfa Wang,Qi Tian,Yue Ma,Andong Wang,Zhiyuan Zhao,Shaobo Min,Wenzh
e Zhao,Kaihao Zhang,Heung-Yeung Shum,Wei Liu,Mengyang LIU,Wenhan Luo
Towards Multiple Character Image Animation Through Enhancing Implicit Decoupling
Controllable character image animation has a wide range of applications. Althoug
h existing studies have consistently improved performance, challenges persist in
 the field of character image animation, particularly concerning stability in co
mplex backgrounds and tasks involving multiple characters. To address these chal
lenges, we propose a novel multi-condition guided framework for character image
animation, employing several well-designed input modules to enhance the implicit
 decoupling capability of the model. First, the optical flow guider calculates t
he background optical flow map as guidance information, which enables the model
to implicitly learn to decouple the background motion into background constants
and background momentum during training, and generate a stable background by set
ting zero background momentum during inference. Second, the depth order guider c
alculates the order map of the characters, which transforms the depth informatio
n into the positional information of multiple characters. This facilitates the i
mplicit learning of decoupling different characters, especially in accurately se
parating the occluded body parts of multiple characters. Third, the reference po
se map is input to enhance the ability to decouple character texture and pose in
formation in the reference image. Furthermore, to fill the gap of fair evaluatio
n of multi-character image animation, we propose a new benchmark comprising abou
t 4,000 frames. Extensive qualitative and quantitative evaluations demonstrate t
hat our method excels in generating high-quality character animations, especiall
y in scenarios of complex backgrounds and multiple characters.
****

Cong Lu,Shengran Hu,Jeff Clune
Intelligent Go-Explore: Standing on the Shoulders of Giant Foundation Models
Go-Explore is a powerful family of algorithms designed to solve hard-exploration
 problems built on the principle of archiving discovered states, and iteratively
 returning to and exploring from the most promising states. This approach has le
d to superhuman performance across a wide variety of challenging problems includ
ing Atari games and robotic control, but requires manually designing heuristics
to guide exploration (i.e., determine which states to save and explore from, and
 what actions to consider next), which is time-consuming and infeasible in gener
al. To resolve this, we propose Intelligent Go-Explore (IGE) which greatly exten
ds the scope of the original Go-Explore by replacing these handcrafted heuristic
s with the intelligence and internalized human notions of interestingness captur
ed by giant pretrained foundation models (FMs). This provides IGE with a human-l
ike ability to instinctively identify how interesting or promising any new state
 is (e.g., discovering new objects, locations, or behaviors), even in complex en
vironments where heuristics are hard to define. Moreover, IGE offers the excitin
g opportunity to recognize and capitalize on serendipitous discoveries---states
encountered during exploration that are valuable in terms of exploration, yet wh
ere what makes them interesting was not anticipated by the human user. We evalua
te our algorithm on a diverse range of language and vision-based tasks that requ
ire search and exploration. Across these tasks, IGE strongly exceeds classic rei
nforcement learning and graph search baselines, and also succeeds where prior st
ate-of-the-art FM agents like Reflexion completely fail. Overall, Intelligent Go
-Explore combines the tremendous strengths of FMs and the powerful Go-Explore al
gorithm, opening up a new frontier of research into creating more generally capa
ble agents with impressive exploration capabilities. All our code is open-source
d at: https://github.com/conglu1997/intelligent-go-explore.
****

Xueyao Zhang,Xiaohui Zhang,Kainan Peng,Zhenyu Tang,Vimal Manohar,Yingru Liu,Jeff
 Hwang,Dangna Li,Yuhao Wang,Julian Chan,Yuan Huang,Zhizheng Wu,Mingbo Ma

Vevo: Controllable Zero-Shot Voice Imitation with Self-Supervised Disentanglemen
t
The imitation of voice, targeted on specific speech attributes such as timbre an
d speaking style, is crucial in speech generation. However, existing methods rel
y heavily on annotated data, and struggle with effectively disentangling timbre
and style, leading to challenges in achieving controllable generation, especiall
y in zero-shot scenarios. To address these issues, we propose Vevo, a versatile
zero-shot voice imitation framework with controllable timbre and style. Vevo ope
rates in two core stages: (1) Content-Style Modeling: Given either text or speec
h's content tokens as input, we utilize an autoregressive transformer to generat
e the content-style tokens, which is prompted by a style reference; (2) Acoustic
 Modeling: Given the content-style tokens as input, we employ a flow-matching tr
ansformer to produce acoustic representations, which is prompted by a timbre ref
erence. To obtain the content and content-style tokens of speech, we design a fu
lly self-supervised approach that progressively decouples the timbre, style, and
 linguistic content of speech. Specifically, we adopt VQ-VAE as the tokenizer fo
r the continuous hidden features of HuBERT. We treat the vocabulary size of the
VQ-VAE codebook as the information bottleneck, and adjust it carefully to obtain
 the disentangled speech representations. Solely self-supervised trained on 60K
hours of audiobook speech data, without any fine-tuning on style-specific corpor
a, Vevo matches or surpasses existing methods in accent and emotion conversion t
asks. Additionally, Vevo’s effectiveness in zero-shot voice conversion and text-
to-speech tasks further demonstrates its strong generalization and versatility.
Audio samples are available at https://versavoice.github.io/.
****

Guangsheng Bao,Yanbin Zhao,Juncai He,Yue Zhang
Glimpse: Enabling White-Box Methods to Use Proprietary Models for Zero-Shot LLM-
Generated Text Detection
Advanced large language models (LLMs) can generate text almost indistinguishable
 from human-written text, highlighting the importance of LLM-generated text dete
ction. However, current zero-shot techniques face challenges as white-box method
s are restricted to use weaker open-source LLMs, and black-box methods are limit
ed by partial observation from stronger proprietary LLMs. It seems impossible to
 enable white-box methods to use proprietary models because API-level access to
the models neither provides full predictive distributions nor inner embeddings.
To traverse the divide, we propose **Glimpse**, a probability distribution estim
ation approach, predicting the full distributions from partial observations. Des
pite the simplicity of Glimpse, we successfully extend white-box methods like En
tropy, Rank, Log-Rank, and Fast-DetectGPT to latest proprietary models. Experime
nts show that Glimpse with Fast-DetectGPT and GPT-3.5 achieves an average AUROC
of about 0.95 in five latest source models, improving the score by 51\% relative
 to the remaining space of the open source baseline. It demonstrates that the la
test LLMs can effectively detect their own outputs, suggesting that advanced LLM
s may be the best shield against themselves. We release our code and data at htt
ps://github.com/baoguangsheng/glimpse.
****

Armand Foucault,Francois Malgouyres,Franck Mamalet
HADAMRNN: BINARY AND SPARSE TERNARY ORTHOGONAL RNNS
Binary and sparse ternary weights in neural networks enable faster computations
and lighter representations, facilitating their use on edge devices with limited
 computational power. Meanwhile, vanilla RNNs are highly sensitive to changes in
 their recurrent weights, making the binarization and ternarization of these wei
ghts inherently challenging. To date, no method has successfully achieved binari
zation
or ternarization of vanilla RNN weights. We present a new approach leveraging th
e properties of Hadamard matrices to parameterize a subset of binary and sparse
ternary orthogonal matrices. This method enables the training of orthogonal RNNs
 (ORNNs) with binary and sparse ternary recurrent weights, effectively creating
a specific class of binary and sparse ternary vanilla RNNs. The resulting ORNNs,
 called HadamRNN and Block-HadamRNN, are evaluated on benchmarks such as the cop

y task, permuted and sequential MNIST tasks, and IMDB dataset. Despite binarizat
ion or sparse ternarization, these RNNs maintain performance levels comparable t
o state-of-the-art full-precision models, highlighting the effectiveness of our
approach. Notably, our approach is the first solution with binary recurrent weig
hts capable of tackling the copy task over 1000 timesteps.
****

Jiatao Gu,Yuyang Wang,Yizhe Zhang,Qihang Zhang,Dinghuai Zhang,Navdeep Jaitly,Jos
hua M. Susskind,Shuangfei Zhai
Denoising Autoregressive Transformers for Scalable Text-to-Image Generation
Diffusion models have become the dominant approach for visual generation. They a
re trained by denoising a Markovian process which gradually adds noise to the in
put. We argue that the Markovian property limits the model’s ability to fully ut
ilize the generation trajectory, leading to inefficiencies during training and i
nference. In this paper, we propose DART, a transformer-based model that unifies
 autoregressive (AR) and diffusion within a non-Markovian framework. DART itera
tively denoises image patches spatially and spectrally using an AR model that ha
s the same architecture as standard language models. DART does not rely on image
 quantization, which enables more effective image modeling while maintaining fle
xibility. Furthermore, DART seamlessly trains with both text and image data in a
 unified model. Our approach demonstrates competitive performance on class-condi
tioned and text-to-image generation tasks, offering a scalable, efficient altern
ative to traditional diffusion models. Through this unified framework, DART sets
 a new benchmark for scalable, high-quality image synthesis.
****

Smail Ait Bouhsain,Rachid Alami,Thierry Simeon
Learning Geometric Reasoning Networks For Robot Task And Motion Planning
Task and Motion Planning (TAMP) is a computationally challenging robotics proble
m due to the tight coupling of discrete symbolic planning and continuous geometr
ic planning of robot motions. In particular, planning manipulation tasks in comp
lex 3D environments leads to a large number of costly geometric planner queries
to verify the feasibility of considered actions and plan their motions. To addre
ss this issue, we propose Geometric Reasoning Networks (GRN), a graph neural net
work (GNN)-based model for action and grasp feasibility prediction, designed to
significantly reduce the dependency on the geometric planner. Moreover, we intro
duce two key interpretability mechanisms: inverse kinematics (IK) feasibility pr
ediction and grasp obstruction (GO) estimation. These modules not only improve f
easibility predictions accuracy, but also explain why certain actions or grasps
are infeasible, thus allowing a more efficient search for a feasible solution. T
hrough extensive experimental results, we show that our model outperforms state-
of-the-art methods, while maintaining generalizability to more complex environme
nts, diverse object shapes, multi-robot settings, and real-world robots.
****

Xueyi Liu,Jianibieke Adalibieke,Qianwei Han,Yuzhe Qin,Li Yi
DexTrack: Towards Generalizable Neural Tracking Control for Dexterous Manipulati
on from Human References
We address the challenge of developing a generalizable neural tracking controlle
r for dexterous manipulation from human references. This controller aims to mana
ge a dexterous robot hand to manipulate diverse objects for various purposes def
ined by kinematic human-object interactions. Developing such a controller is com
plicated by the intricate contact dynamics of dexterous manipulation and the nee
d for adaptivity, generalizability, and robustness. Current reinforcement learni
ng and trajectory optimization methods often fall short due to their dependence
on task-specific rewards or precise system models. We introduce an approach that
 curates large-scale successful robot tracking demonstrations, comprising pairs
of human references and robot actions, to train a neural controller. Utilizing a
 data flywheel, we iteratively enhance the controller's performance, as well as
the number and quality of successful tracking demonstrations. We exploit availab
le tracking demonstrations and carefully integrate reinforcement learning and im
itation learning to boost the controller's performance in dynamic environments.
At the same time, to obtain high-quality tracking demonstrations, we individuall

y optimize per-trajectory tracking by leveraging the learned tracking controller in a homotopy optimization method. The homotopy optimization, mimicking chain-of-thought, aids in solving challenging trajectory tracking problems to increase demonstration diversity. We showcase our success by training a generalizable neural controller and evaluating it in both simulation and real world. Our method achieves over a 10% improvement in success rates compared to leading baselines. The project website with animated results is available at [DexTrack](https://meowuu7.github.io/DexTrack/).
****
Akio Hayakawa,Masato Ishii,Takashi Shibuya,Yuki Mitsufuji
MMDisCo: Multi-Modal Discriminator-Guided Cooperative Diffusion for Joint Audio and Video Generation
This study aims to construct an audio-video generative model with minimal computational cost by leveraging pre-trained single-modal generative models for audio and video.
To achieve this, we propose a novel method that guides single-modal models to cooperatively generate well-aligned samples across modalities.
Specifically, given two pre-trained base diffusion models, we train a lightweight joint guidance module to adjust scores separately estimated by the base models to match the score of joint distribution over audio and video.
We show that this guidance can be computed using the gradient of the optimal discriminator, which distinguishes real audio-video pairs from fake ones independently generated by the base models.
Based on this analysis, we construct a joint guidance module by training this discriminator.
Additionally, we adopt a loss function to stabilize the discriminator's gradient and make it work as a noise estimator, as in standard diffusion models.
Empirical evaluations on several benchmark datasets demonstrate that our method improves both single-modal fidelity and multimodal alignment with relatively few parameters.
The code is available at: https://github.com/SonyResearch/MMDisCo.
****
Boyang Zheng,Chumeng Liang,Xiaoyu Wu
Targeted Attack Improves Protection against Unauthorized Diffusion Customization
Diffusion models build a new milestone for image generation yet raising public concerns, for they can be fine-tuned on unauthorized images for customization. Protection based on adversarial attacks rises to encounter this unauthorized diffusion customization, by adding protective watermarks to images and poisoning diffusion models. However, current protection, leveraging untargeted attacks, does not appear to be effective enough. In this paper, we propose a simple yet effective improvement for the protection against unauthorized diffusion customization by introducing targeted attacks. We show that by carefully selecting the target, targeted attacks significantly outperform untargeted attacks in poisoning diffusion models and degrading the customization image quality. Extensive experiments validate the superiority of our method on two mainstream customization methods of diffusion models, compared to existing protections. To explain the surprising success of targeted attacks, we delve into the mechanism of attack-based protections and propose a hypothesis based on our observation, which enhances the comprehension of attack-based protections. To the best of our knowledge, we are the first to both reveal the vulnerability of diffusion models to targeted attacks and leverage targeted attacks to enhance protection against unauthorized diffusion customization.
****
Lisa Dunlap,Krishna Mandal,Trevor Darrell,Jacob Steinhardt,Joseph E. Gonzalez
VibeCheck: Discover and Quantify Qualitative Differences in Large Language Models
Large language models (LLMs) often exhibit subtle yet distinctive characteristics in their outputs that users intuitively recognize, but struggle to quantify. These "vibes" -- such as tone, formatting, or writing style -- influence user pre

ferences, yet traditional evaluations focus primarily on the singular vibe of correctness.
We introduce $\textbf{VibeCheck}$, a system for automatically comparing a pair of LLMs by discovering identifying traits of a model ("vibes") that are well-defined, differentiating, and user-aligned. VibeCheck iteratively discovers vibes from model outputs and then utilizes a panel of LLM judges to quantitatively measure the utility of each vibe.
We validate that the vibes generated by VibeCheck align with those found in human discovery and run VibeCheck on pairwise preference data from real-world user conversations with Llama-3-70b vs GPT-4. VibeCheck reveals that Llama has a friendly, funny, and somewhat controversial vibe. These vibes predict model identity with 80% accuracy and human preference with 61% accuracy. Lastly, we run VibeCheck on a variety of models and tasks, including summarization, math, and captioning to provide insight into differences in model behavior. VibeCheck discovers vibes like Command X prefers to add concrete intros and conclusions when summarizing in comparison to TNGL, Llama-405b often overexplains its thought process on math problems compared to GPT-4o, and GPT-4 prefers to focus on the mood and emotions of the scene when captioning compared to Gemini-1.5-Flash.
****
Shubham Ugare,Rohan Gumaste,Tarun Suresh,Gagandeep Singh,Sasa Misailovic
IterGen: Iterative Semantic-aware Structured LLM Generation with Backtracking
Large Language Models (LLMs) are widely used for tasks such as natural language and code generation, but their outputs often suffer from issues like hallucination, toxicity, and incorrect results. Current libraries for structured LLM generation rely on left-to-right decoding without support for backtracking, limiting the ability to correct or refine outputs mid-generation.

To address this, we introduce IterGen, a user-friendly library for iterative, grammar-guided LLM generation that enables users to move both forward and backward within the generated output based on grammar symbols.
By leveraging a symbol-to-position mapping and maintaining the key-value (KV) cache state, IterGen ensures efficient and structured generation while allowing for corrections during the process. We demonstrate IterGen's effectiveness in two important applications: reducing privacy leakage in LLM outputs, improving the accuracy of LLM-generated SQL and Vega-Lite queries.

Our code and additional resources are available at https://structuredllm.com.
****
Brendan Leigh Ross,Hamidreza Kamkari,Tongzi Wu,Rasa Hosseinzadeh,Zhaoyan Liu,George Stein,Jesse C. Cresswell,Gabriel Loaiza-Ganem
A Geometric Framework for Understanding Memorization in Generative Models
As deep generative models have progressed, recent work has shown them to be capable of memorizing and reproducing training datapoints when deployed. These findings call into question the usability of generative models, especially in light of the legal and privacy risks brought about by memorization. To better understand this phenomenon, we propose the *manifold memorization hypothesis* (MMH), a geometric framework which leverages the manifold hypothesis into a clear language in which to reason about memorization. We propose to analyze memorization in terms of the relationship between the dimensionalities of (i) the ground truth data manifold and (ii) the manifold learned by the model. This framework provides a formal standard for "how memorized" a datapoint is and systematically categorizes memorized data into two types: memorization driven by overfitting and memorization driven by the underlying data distribution. By analyzing prior work in the context of the MMH, we explain and unify assorted observations in the literature. We empirically validate the MMH using synthetic data and image datasets up to the scale of Stable Diffusion, developing new tools for detecting and preventing generation of memorized samples in the process.
****
Pengyang Ling,Jiazi Bu,Pan Zhang,Xiaoyi Dong,Yuhang Zang,Tong Wu,Huaian Chen,Jiaqi Wang,Yi Jin

MotionClone: Training-Free Motion Cloning for Controllable Video Generation
Motion-based controllable video generation offers the potential for creating cap
tivating visual content. Existing methods typically necessitate model training t
o encode particular motion cues or incorporate fine-tuning to inject certain mot
ion patterns, resulting in limited flexibility and generalization. In this work,
 we propose MotionClone, a training-free framework that enables motion cloning f
rom reference videos to versatile motion-controlled video generation, including
text-to-video and image-to-video. Based on the observation that the dominant com
ponents in temporal-attention maps drive motion synthesis, while the rest mainly
 capture noisy or very subtle motions, MotionClone utilizes sparse temporal atte
ntion weights as motion representations for motion guidance, facilitating divers
e motion transfer across varying scenarios. Meanwhile, MotionClone allows for th
e direct extraction of motion representation through a single denoising step, by
passing the cumbersome inversion processes and thus promoting both efficiency an
d flexibility.
Extensive experiments demonstrate that MotionClone exhibits proficiency in both
global camera motion and local object motion, with notable superiority in terms
of motion fidelity, textual alignment, and temporal consistency.
****
Armand Comas,Aditya Chattopadhyay,Feliu Formosa,Changyu Liu,Octavia Camps,Rene V
idal
InCoDe: Interpretable Compressed Descriptions For Image Generation
Generative models have been successfully applied in diverse domains, from natura
l language processing to image synthesis. However, despite this success, a key c
hallenge that remains is the ability to control the semantic content of the scen
e being generated. We argue that adequate control of the generation process requ
ires a data representation that allows users to access and efficiently manipulat
e the semantic factors shaping the data distribution. This work advocates for th
e adoption of succinct, informative, and interpretable representations, quantifi
ed using information-theoretic principles. Through extensive experiments, we dem
onstrate the efficacy of our proposed framework both qualitatively and quantitat
ively. Our work contributes to the ongoing quest to enhance both controllability
 and interpretability in the generation process. Code available at github.com/Ar
mandCom/InCoDe.
****
Weronika Ormaniec,Scott Sussex,Lars Lorch,Bernhard Schölkopf,Andreas Krause
Standardizing Structural Causal Models
Synthetic datasets generated by structural causal models (SCMs) are commonly use
d for benchmarking causal structure learning algorithms. However, the variances
and pairwise correlations in SCM data tend to increase along the causal ordering
. Several popular algorithms exploit these artifacts, possibly leading to conclu
sions that do not generalize to real-world settings. Existing metrics like $\ope
ratorname{Var}$-sortability and $\operatorname{R^2}$-sortability quantify these
patterns, but they do not provide tools to remedy them. To address this, we prop
ose internally-standardized structural causal models (iSCMs), a modification of
SCMs that introduces a standardization operation at each variable during the gen
erative process. By construction, iSCMs are not Var-sortable. W
e also find empirical evidence that they are mostly not $\operatorname{R^2}$-sor
table for commonly-used graph families. Moreover, contrary to the post-hoc stand
ardization of data generated by standard SCMs, we prove that linear iSCMs are le
ss identifiable from prior knowledge on the weights and do not collapse to deter
ministic relationships in large systems, which may make iSCMs a useful model in
causal inference beyond the benchmarking problem studied here. Our code is publi
cly available at: https://github.com/werkaaa/iscm.
****
Tristan Luca Saidi,Abigail Hickok,Andrew J. Blumberg
Recovering Manifold Structure Using Ollivier Ricci Curvature
We introduce ORC-ManL, a new algorithm to prune spurious edges from nearest neig
hbor graphs using a criterion based on Ollivier-Ricci curvature and estimated me
tric distortion. Our motivation comes from manifold learning: we show that when

the data generating the nearest-neighbor graph consists of noisy samples from a low-dimensional manifold, edges that shortcut through the ambient space have more negative Ollivier-Ricci curvature than edges that lie along the data manifold. We demonstrate that our method outperforms alternative pruning methods and that it significantly improves performance on many downstream geometric data analysis tasks that use nearest neighbor graphs as input. Specifically, we evaluate on manifold learning, persistent homology, dimension estimation, and others. We also show that ORC-ManL can be used to improve clustering and manifold learning of single-cell RNA sequencing data. Finally, we provide empirical convergence experiments that support our theoretical findings.
****
Neil Rathi,Johannes Mehrer,Badr AlKhamissi,Taha Osama A Binhuraib,Nicholas Blauch,Martin Schrimpf
TopoLM: brain-like spatio-functional organization in a topographic language model
Neurons in the brain are spatially organized such that neighbors on tissue often exhibit similar response profiles. In the human language system, experimental studies have observed clusters for syntactic and semantic categories, but the mechanisms underlying this functional organization remain unclear. Here, building on work from the vision literature, we develop TopoLM, a transformer language model with an explicit two-dimensional spatial representation of model units. By combining a next-token prediction objective with a spatial smoothness loss, representations in this model assemble into clusters that correspond to semantically interpretable groupings of text and closely match the functional organization in the brain's language system. TopoLM successfully predicts the emergence of a spatially organized cortical language system as well as the organization of functional clusters selective for fine-grained linguistic features empirically observed in human cortex. Our results suggest that the functional organization of the human language system is driven by a unified spatial objective, and provide a functionally and spatially aligned model of language processing in the brain.Neurons in the brain are spatially organized such that neighbors on tissue often exhibit similar response profiles. In the human language system, experimental studies have observed clusters for syntactic and semantic categories, but the mechanisms underlying this functional organization remain unclear. Here, building on work from the vision literature, we develop TopoLM, a transformer language model with an explicit two-dimensional spatial representation of model units. By combining a next-token prediction objective with a spatial smoothness loss, representations in this model assemble into clusters that correspond to semantically interpretable groupings of text and closely match the functional organization in the brain's language system. TopoLM successfully predicts the emergence of a spatially organized cortical language system as well as the organization of functional clusters selective for fine-grained linguistic features empirically observed in human cortex. Our results suggest that the functional organization of the human language system is driven by a unified spatial objective, and provide a functionally and spatially aligned model of language processing in the brain.
****
Alexandra Butoi,Ghazal Khalighinejad,Anej Svete,Josef Valvoda,Ryan Cotterell,Brian DuSell
Training Neural Networks as Recognizers of Formal Languages
Characterizing the computational power of neural network architectures in terms of formal language theory remains a crucial line of research, as it describes lower and upper bounds on the reasoning capabilities of modern AI. However, when empirically testing these bounds, existing work often leaves a discrepancy between experiments and the formal claims they are meant to support. The problem is that formal language theory pertains specifically to recognizers: machines that receive a string as input and classify whether it belongs to a language. On the other hand, it is common instead to evaluate language models on proxy tasks, e.g., language modeling or sequence-to-sequence transduction, that are similar in only an informal sense to the underlying theory. We correct this mismatch by training and evaluating neural networks directly as binary classifiers of strings, usi

ng a general method that can be applied to a wide variety of languages. As part of this, we extend an algorithm recently proposed by Snæbjarnarson et al. (2025) for efficient length-controlled sampling of strings from regular languages. We provide results on a variety of languages across the Chomsky hierarchy for three neural architectures: a simple RNN, an LSTM, and a causally-masked transformer. We find that the RNN and LSTM often outperform the transformer, and that auxiliary training objectives such as language modeling can help, although no single objective uniformly improves performance across languages and architectures. Our contributions will facilitate theoretically sound empirical testing of language recognition claims in future work. We have released our datasets as a benchmark called FLaRe (Formal Language Recognition), along with our code.
****
Dhruv Agarwal,Manoj Ghuhan Arivazhagan,Rajarshi Das,Sandesh Swamy,Sopan Khosla,Rashmi Gangadharaiah
Searching for Optimal Solutions with LLMs via Bayesian Optimization
Scaling test-time compute to search for optimal solutions is an important step towards building generally-capable language models that can reason. Recent work, however, shows that tasks of varying complexity require distinct search strategies to solve optimally, thus making it challenging to design a one-size-fits-all approach. Prior solutions either attempt to predict task difficulty to select the optimal search strategy, often infeasible in practice, or use a static, pre-defined strategy, e.g., repeated parallel sampling or greedy sequential search, which is sub-optimal. In this work, we argue for an alternative view using the probabilistic framework of Bayesian optimization (BO), where the search strategy is adapted dynamically based on the evolving uncertainty estimates of solutions as search progresses. To this end, we introduce Bayesian-OPRO (BOPRO)––a generalization of a recent method for in-context optimization, which iteratively samples from new proposal distributions by modifying the prompt to the LLM with a subset of its previous generations selected to explore or exploit different parts of the search space. We evaluate our method on word search, molecule optimization, and a joint hypothesis+program search task using a 1-D version of the challenging Abstraction and Reasoning Corpus (1D-ARC). Our results show that BOPRO outperforms all baselines in word search (≥10 points) and molecule optimization (higher quality and 17% fewer invalid molecules), but trails a best-k prompting strategy in program search. Our analysis reveals that despite the ability to balance exploration and exploitation using BOPRO, failure is likely due to the inability of code representation models in distinguishing sequences with low edit-distances.
****
Grace Zhang,Ayush Jain,Injune Hwang,Shao-Hua Sun,Joseph J Lim
QMP: Q-switch Mixture of Policies for Multi-Task Behavior Sharing
Multi-task reinforcement learning (MTRL) aims to learn several tasks simultaneously for better sample efficiency than learning them separately. Traditional methods achieve this by sharing parameters or relabeling data between tasks. In this work, we introduce a new framework for sharing behavioral policies across tasks, which can be used in addition to existing MTRL methods. The key idea is to improve each task's off-policy data collection by employing behaviors from other task policies. Selectively sharing helpful behaviors acquired in one task to collect training data for another task can lead to higher-quality trajectories, leading to more sample-efficient MTRL. Thus, we introduce a simple and principled framework called Q-switch mixture of policies (QMP) that selectively shares behavior between different task policies by using the task's Q-function to evaluate and select useful shareable behaviors. We theoretically analyze how QMP improves the sample efficiency of the underlying RL algorithm. Our experiments show that QMP's behavioral policy sharing provides complementary gains over many popular MTRL algorithms and outperforms alternative ways to share behaviors in various manipulation, locomotion, and navigation environments. Videos are available at https://qmp-mtrl.github.io/.
****
Brian Formento,Chuan-Sheng Foo,See-Kiong Ng
Confidence Elicitation: A New Attack Vector for Large Language Models

A fundamental issue in deep learning has been adversarial robustness. As these s
ystems have scaled, such issues have persisted. Currently, large language models
 (LLMs) with billions of parameters suffer from adversarial attacks just like th
eir earlier, smaller counterparts. However, the threat models have changed. Prev
iously, having gray-box access, where input embeddings or output logits/probabil
ities were visible to the user, might have been reasonable. However, with the in
troduction of closed-source models, no information about the model is available
apart from the generated output. This means that current black-box attacks can o
nly utilize the final prediction to detect if an attack is successful. In this w
ork, we investigate and demonstrate the potential of attack guidance, akin to us
ing output probabilities, while having only black-box access in a classification
 setting. This is achieved through the ability to elicit confidence from the mod
el. We empirically show that the elicited confidence is calibrated and not hallu
cinated for current LLMs. By minimizing the elicited confidence, we can therefor
e increase the likelihood of misclassification. Our new proposed paradigm demons
trates promising state-of-the-art results on three datasets across two models (
LLaMA-3-8B-Instruct and Mistral-7B-Instruct-V0.3) when comparing our technique t
o existing hard-label black-box attack methods that introduce word-level substit
utions. The code is publicly available at GitHub: Confidence_Elicitation_Attacks
.
****
Zhuowei Chen,Qiannan Zhang,Shichao Pei
Injecting Universal Jailbreak Backdoors into LLMs in Minutes
Jailbreak backdoor attacks on LLMs have garnered attention for their effectivene
ss and stealth. However, existing methods rely on the crafting of poisoned datas
ets and the time-consuming process of fine-tuning. In this work, we propose Jail
breakEdit, a novel jailbreak backdoor injection method that exploits model editi
ng techniques to inject a universal jailbreak backdoor into safety-aligned LLMs
with minimal intervention *in minutes*. JailbreakEdit integrates a multi-node ta
rget estimation to estimate the jailbreak space, thus creating shortcuts from th
e backdoor to this estimated jailbreak space that induce jailbreak actions. Our
attack effectively shifts the models' attention by attaching strong semantics to
 the backdoor, enabling it to bypass internal safety mechanisms. Experimental re
sults show that JailbreakEdit achieves a high jailbreak success rate on jailbrea
k prompts while preserving generation quality, and safe performance on normal qu
eries. Our findings underscore the effectiveness, stealthiness, and explainabili
ty of JailbreakEdit, emphasizing the need for more advanced defense mechanisms i
n LLMs.
****
Dawit Mureja Argaw,Xian Liu,Qinsheng Zhang,Joon Son Chung,Ming-Yu Liu,Fitsum Red
a
High-Quality Joint Image and Video Tokenization with Causal VAE
Generative modeling has seen significant advancements in image and video synthes
is. However, the curse of dimensionality remains a significant obstacle, especia
lly for video generation, given its inherently complex and high-dimensional natu
re. Many existing works rely on low-dimensional latent spaces from pretrained im
age autoencoders. However, this approach overlooks temporal redundancy in videos
 and often leads to temporally incoherent decoding. To address this issue, we pr
opose a video compression network that reduces the dimensionality of visual data
 both spatially and temporally. Our model, based on a variational autoencoder, e
mploys causal 3D convolution to handle images and videos jointly. The key contri
butions of our work include a scale-agnostic encoder for preserving video fideli
ty, a novel spatio-temporal down/upsampling block for robust long-sequence model
ing, and a flow regularization loss for accurate motion decoding.
Our approach outperforms competitors in video quality and compression rates acro
ss various datasets. Experimental analyses also highlight its potential as a rob
ust autoencoder for video generation training.
****
Rui Qiao,Zhaoxuan Wu,Jingtan Wang,Pang Wei Koh,Bryan Kian Hsiang Low
Group-robust Sample Reweighting for Subpopulation Shifts via Influence Functions

Machine learning models often have uneven performance among subpopulations
(a.k.a., groups) in the data distributions. This poses a significant challenge f
or the
models to generalize when the proportions of the groups shift during deployment.
To improve robustness to such shifts, existing approaches have developed strateg
ies
that train models or perform hyperparameter tuning using the group-labeled data
to minimize the worst-case loss over groups. However, a non-trivial amount of
high-quality labels is often required to obtain noticeable improvements. Given
the costliness of the labels, we propose to adopt a different paradigm to enhanc
e
group label efficiency: utilizing the group-labeled data as a target set to opti
mize
the weights of other group-unlabeled data. We introduce Group-robust Sample
Reweighting (GSR), a two-stage approach that first learns the representations fr
om
group-unlabeled data, and then tinkers the model by iteratively retraining its l
ast
layer on the reweighted data using influence functions. Our GSR is theoretically
sound, practically lightweight, and effective in improving the robustness to sub
-
population shifts. In particular, GSR outperforms the previous state-of-the-art
approaches that require the same amount or even more group labels. Our code is
available at https://github.com/qiaoruiyt/GSR.
****
Sedigheh Eslami,Gerard de Melo
Mitigate the Gap: Improving Cross-Modal Alignment in CLIP
Contrastive Language--Image Pre-training (CLIP) has manifested remarkable improv
ements in zero-shot classification and cross-modal vision-language tasks. Yet, f
rom a geometrical point of view, the CLIP embedding space has been found to have
 a pronounced modality gap. This gap renders the embedding space overly sparse a
nd disconnected, with different modalities being densely distributed in distinct
 subregions of the hypersphere. In this work, we propose AlignCLIP, in order to
improve the alignment between text and image embeddings, and thereby reduce the
modality gap. AlignCLIP increases the cross-modal alignment, and yields gains ac
ross several zero-shot and fine-tuning downstream evaluations by sharing the lea
rnable parameters between the modality encoders and a semantically-regularized s
eparation objective function on the uni-modal embeddings. The source code and mo
del checkpoints for reproducing our experiments are available at https://github.
com/sarahESL/AlignCLIP.
****
Lena Krieger,Anna Beer,Pernille Matthews,Anneka Myrup Thiesson,Ira Assent
FairDen: Fair Density-Based Clustering
Fairness in data mining tasks like clustering has recently become an increasingl
y important aspect.
However, few clustering algorithms exist that focus on fair groupings of data wi
th sensitive attributes.
Including fairness in the clustering objective is especially hard for density-ba
sed clustering, as it does not directly optimize a closed form objective like ce
ntroid-based or spectral methods.

This paper introduces FairDen, the first fair, density-based clustering algorith
m.
We capture the dataset's density-connectivity structure in a similarity matrix t
hat we manipulate to encourage a balanced clustering.
In contrast to state-of-the-art, FairDen inherently handles categorical attribut
es, noise, and data with several sensitive attributes or groups.
We show that FairDen finds meaningful and fair clusters in extensive experiments
.
****

Konstantin Hess,Stefan Feuerriegel
Stabilized Neural Prediction of Potential Outcomes in Continuous Time
Patient trajectories from electronic health records are widely used to estimate conditional average potential outcomes (CAPOs) of treatments over time, which then allows to personalize care. Yet, existing neural methods for this purpose have a key limitation: while some adjust for time-varying confounding, these methods assume that the time series are recorded in discrete time. In other words, they are constrained to settings where measurements and treatments are conducted at fixed time steps, even though this is unrealistic in medical practice. In this work, we aim to estimate CAPOs in continuous time. The latter is of direct practical relevance because it allows for modeling patient trajectories where measurements and treatments take place at arbitrary, irregular timestamps. We thus propose a new method called stabilized continuous time inverse propensity network (SCIP-Net). For this, we further derive stabilized inverse propensity weights for robust estimation of the CAPOs. To the best of our knowledge, our SCIP-Net is the first neural method that performs proper adjustments for time-varying confounding in continuous time.
****
Seth Aycock,David Stap,Di Wu,Christof Monz,Khalil Sima'an
Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?
Extremely low-resource (XLR) languages lack substantial corpora for training NLP models, motivating the use of all available resources such as dictionaries and grammar books. Machine Translation from One Book (Tanzer et al., 2024) suggests that prompting long-context LLMs with one grammar book enables English–Kalamang translation, an XLR language unseen by LLMs—a noteworthy case of linguistics helping an NLP task. We investigate the source of this translation ability, finding almost all improvements stem from the book’s parallel examples rather than its grammatical explanations. We find similar results for Nepali and Guarani, seen low-resource languages, and we achieve performance comparable to an LLM with a grammar book by simply fine-tuning an encoder-decoder translation model. We then investigate where grammar books help by testing two linguistic tasks, grammaticality judgment and gloss prediction, and we explore what kind of grammatical knowledge helps by introducing a typological feature prompt that achieves leading results on these more relevant tasks. We thus emphasise the importance of task-appropriate data for XLR languages: parallel examples for translation, and grammatical data for linguistic tasks. As we find no evidence that long-context LLMs can make effective use of grammatical explanations for XLR translation, we conclude data collection for multilingual XLR tasks such as translation is best focused on parallel data over linguistic description.
****
Xi Wang,Taketomo Isazawa,Liana Mikaelyan,James Hensman
KBLaM: Knowledge Base augmented Language Model
In this paper, we propose Knowledge Base augmented Language Model (KBLAM), a new method for augmenting Large Language Models (LLMs) with external knowledge. KBLAM works with a knowledge base (KB) constructed from a corpus of documents, transforming each piece of knowledge in the KB into continuous key-value vector pairs via pre-trained sentence encoders with linear adapters and
integrating them into pre-trained LLMs via a specialized rectangular attention mechanism. Unlike Retrieval-Augmented Generation, KBLAM eliminates external retrieval modules, and unlike in-context learning, its computational overhead scales linearly with KB size rather than quadratically. Our approach enables integrating a large KB of more than 10K triples into an 8B pre-trained LLM of only 8K context window on one single A100 80GB GPU and allows for dynamic updates without model fine-tuning or retraining. Experiments demonstrate KBLAM’s effectiveness in various tasks, including question-answering and open-ended reasoning, while providing interpretable insights into its use of the augmented knowledge. Code and datasets are available at https://github.com/microsoft/KBLaM/
****
Alireza Mousavi-Hosseini,Adel Javanmard,Murat A Erdogdu

Robust Feature Learning for Multi-Index Models in High Dimensions
Recently, there have been numerous studies on feature learning with neural netwo
rks, specifically on learning single- and multi-index models where the target is
 a function of a low-dimensional projection of the input. Prior works have shown
 that in high dimensions, the majority of the compute and data resources are spe
nt on recovering the low-dimensional projection; once this subspace is recovered
, the remainder of the target can be learned independently of the ambient dimens
ion. However, implications of feature learning in adversarial settings remain un
explored. In this work, we take the first steps towards understanding adversaria
lly robust feature learning with neural networks. Specifically, we prove that th
e hidden directions of a multi-index model offer a Bayes optimal low-dimensional
 projection for robustness against ℓ_2-bounded adversarial perturbations un
der the squared loss, assuming that the multi-index coordinates are statisticall
y independent from the rest of the coordinates. Therefore, robust learning can b
e achieved by first performing standard feature learning, then robustly tuning a
 linear readout layer on top of the standard representations. In particular, we
show that adversarially robust learning is just as easy as standard learning. Sp
ecifically, the additional number of samples needed to robustly learn multi-inde
x models when compared to standard learning, does not depend on dimensionality.
****
Qichao Shentu,Beibu Li,Kai Zhao,Yang Shu,Zhongwen Rao,Lujia Pan,Bin Yang,Chenjua
n Guo
Towards a General Time Series Anomaly Detector with Adaptive Bottlenecks and Dua
l Adversarial Decoders
Time series anomaly detection plays a vital role in a wide range of applications
. Existing methods require training one specific model for each dataset, which e
xhibits limited generalization capability across different target datasets, hind
ering anomaly detection performance in various scenarios with scarce training da
ta. Aiming at this problem, we propose constructing a general time series anomal
y detection model, which is pre-trained on extensive multi-domain datasets and c
an subsequently apply to a multitude of downstream scenarios. The significant di
vergence of time series data across different domains presents two primary chall
enges in building such a general model: (1) meeting the diverse requirements of
appropriate information bottlenecks tailored to different datasets in one unifie
d model, and (2) enabling distinguishment between multiple normal and abnormal p
atterns, both are crucial for effective anomaly detection in various target scen
arios. To tackle these two challenges, we propose a General time series anomaly
Detector with Adaptive Bottlenecks and Dual Adversarial Decoders (DADA), which e
nables flexible selection of bottlenecks based on different data and explicitly
enhances clear differentiation between normal and abnormal series. We conduct ex
tensive experiments on nine target datasets from different domains. After pre-tr
aining on multi-domain data, DADA, serving as a zero-shot anomaly detector for t
hese datasets, still achieves competitive or even superior results compared to t
hose models tailored to each specific dataset.
****
Yarden As,Bhavya Sukhija,Lenart Treven,Carmelo Sferrazza,Stelian Coros,Andreas K
rause
ActSafe: Active Exploration with Safety Constraints for Reinforcement Learning
Reinforcement learning (RL) is ubiquitous in the development of modern AI system
s. However, state-of-the-art RL agents require extensive, and potentially
unsafe, interactions with their environments to learn effectively. These limitat
ions
confine RL agents to simulated environments, hindering their ability to learn
directly in real-world settings. In this work, we present ActSafe, a novel
model-based RL algorithm for safe and efficient exploration. ActSafe learns
a well-calibrated probabilistic model of the system and plans optimistically
w.r.t. the epistemic uncertainty about the unknown dynamics, while enforcing
pessimism w.r.t. the safety constraints. Under regularity assumptions on the
constraints and dynamics, we show that ActSafe guarantees safety during
learning while also obtaining a near-optimal policy in finite time. In addition,

we
propose a practical variant of ActSafe that builds on latest model-based RL adva
ncements and enables safe exploration even in high-dimensional settings such
as visual control. We empirically show that ActSafe obtains state-of-the-art
performance in difficult exploration tasks on standard safe deep RL benchmarks
while ensuring safety during learning.
****
Bryan Chan,Xinyi Chen,András György,Dale Schuurmans
Toward Understanding In-context vs. In-weight Learning
It has recently been demonstrated empirically that in-context learning emerges i
n transformers when certain distributional properties are present in the trainin
g data, but this ability can also diminish upon further training. We provide a n
ew theoretical understanding of these phenomena by identifying simplified distri
butional properties that give rise to the emergence and eventual disappearance o
f in-context learning. We do so by first analyzing a simplified model that uses
a gating mechanism to choose between an in-weight and an in-context predictor. T
hrough a combination of a generalization error and regret analysis we identify c
onditions where in-context and in-weight learning emerge. These theoretical find
ings are then corroborated experimentally by comparing the behaviour of a full t
ransformer on the simplified distributions to that of the stylized model, demons
trating aligned results. We then extend the study to a full large language model
, showing how fine-tuning on various collections of natural language prompts can
 elicit similar in-context and in-weight learning behaviour.
****
Maksym Andriushchenko,Nicolas Flammarion
Does Refusal Training in LLMs Generalize to the Past Tense?
Refusal training is widely used to prevent LLMs from generating harmful, undesir
able, or illegal outputs. We reveal a curious generalization gap in the current
refusal training approaches: simply reformulating a harmful request in the past
tense (e.g., *"How to make a Molotov cocktail?"* to *"How did people make a Molo
tov cocktail?"*) is often sufficient to jailbreak many state-of-the-art LLMs. We
 systematically evaluate this method on Llama-3 8B, Claude-3.5 Sonnet, GPT-3.5 T
urbo, Gemma-2 9B, Phi-3-Mini, GPT-4o-mini, GPT-4o, o1-mini, o1-preview, and R2D2
 models using GPT-3.5 Turbo as a reformulation model. For example, the success r
ate of this simple attack on GPT-4o increases from 1\% using direct requests to
88\% using 20 past-tense reformulation attempts on harmful requests from Jailbre
akBench with GPT-4 as a jailbreak judge. Interestingly, we also find that reform
ulations in the future tense are less effective, suggesting that refusal guardra
ils tend to consider past historical questions more benign than hypothetical fut
ure questions. Moreover, our experiments on fine-tuning GPT-3.5 Turbo show that
defending against past reformulations is feasible when past tense examples are e
xplicitly included in the fine-tuning data. Overall, our findings highlight that
 the widely used alignment techniques---such as SFT, RLHF, and adversarial train
ing---employed to align the studied models can be brittle and do not always gene
ralize as intended. We provide code and jailbreak artifacts at https://github.co
m/tml-epfl/llm-past-tense.
****
Rui Xu,Chao Chen,Yue Sun,Parvathinathan Venkitasubramaniam,Sihong Xie
Wasserstein-Regularized Conformal Prediction under General Distribution Shift
Conformal prediction yields a prediction set with guaranteed $1-\alpha$ coverage
 of the true target under the i.i.d. assumption,
which can fail and lead to a gap between $1-\alpha$ and the actual coverage. Pri
or studies bound the gap using total variation distance, which cannot identify t
he gap changes under distribution shift at different α, thus serving as a
 weak indicator of prediction set validity. Besides, existing methods are mostly
 limited to covariate shifts, while general joint distribution shifts are more c
ommon in practice but less researched. In response, we first propose a Wasserste
in distance-based upper bound of the coverage gap and analyze the bound using pr
obability measure pushforwards between the shifted joint data and conformal scor
e distributions, enabling a separation of the effect of covariate and concept sh

ifts over the coverage gap. We exploit the separation to design algorithms based on importance weighting and regularized representation learning (WR-CP) to reduce the Wasserstein bound with a finite-sample error bound. WR-CP achieves a controllable balance between conformal prediction accuracy and efficiency. Experiments on six datasets prove that WR-CP can reduce coverage gaps to 3.2% across different confidence levels and outputs prediction sets 37% smaller than the worst-case approach on average.
****
Yibo Zhang,Lihong Wang,Changqing Zou,Tieru Wu,Rui Ma
Diff3DS: Generating View-Consistent 3D Sketch via Differentiable Curve Rendering
3D sketches are widely used for visually representing the 3D shape and structure of objects or scenes. However, the creation of 3D sketch often requires users to possess professional artistic skills. Existing research efforts primarily focus on enhancing the ability of interactive sketch generation in 3D virtual systems. In this work, we propose Diff3DS, a novel differentiable rendering framework for generating view-consistent 3D sketch by optimizing 3D parametric curves under various supervisions. Specifically, we perform perspective projection to render the 3D rational Bézier curves into 2D curves, which are subsequently converted to a 2D raster image via our customized differentiable rasterizer. Our framework bridges the domains of 3D sketch and raster image, achieving end-to-end optimization of 3D sketch through gradients computed in the 2D image domain. Our Diff3DS can enable a series of novel 3D sketch generation tasks, including text-to-3D sketch and image-to-3D sketch, supported by the popular distillation-based supervision, such as Score Distillation Sampling (SDS). Extensive experiments have yielded promising results and demonstrated the potential of our framework. Project: https://yiboz2001.github.io/Diff3DS/
****
Zhanpeng Zhou,Mingze Wang,Yuchen Mao,Bingrui Li,Junchi Yan
Sharpness-Aware Minimization Efficiently Selects Flatter Minima Late In Training
Sharpness-Aware Minimization (SAM) has substantially improved the generalization of neural networks under various settings. Despite the success, its effectiveness remains poorly understood. In this work, we discover an intriguing phenomenon in the training dynamics of SAM, shedding light on understanding its implicit bias towards flatter minima over Stochastic Gradient Descent (SGD). Specifically, we find that *SAM efficiently selects flatter minima late in training*. Remarkably, even a few epochs of SAM applied at the end of training yield nearly the same generalization and solution sharpness as full SAM training. Subsequently, we delve deeper into the underlying mechanism behind this phenomenon. Theoretically, we identify two phases in the learning dynamics after applying SAM late in training: i) SAM first escapes the minimum found by SGD exponentially fast; and ii) then rapidly converges to a flatter minimum within the same valley. Furthermore, we empirically investigate the role of SAM during the early training phase. We conjecture that the optimization method chosen in the late phase is more crucial in shaping the final solution's properties. Based on this viewpoint, we extend our findings from SAM to Adversarial Training.
****
Yu Wang,Xinshuang Liu,Xiusi Chen,Sean O'Brien,Junda Wu,Julian McAuley
Self-Updatable Large Language Models by Integrating Context into Model Parameters
Despite significant advancements in large language models (LLMs), the rapid and frequent integration of small-scale experiences, such as interactions with surrounding objects, remains a substantial challenge. Two critical factors in assimilating these experiences are (1) **Efficacy**: the ability to accurately remember recent events; (2) **Retention**: the capacity to recall long-past experiences. Current methods either embed experiences within model parameters using continual learning, model editing, or knowledge distillation techniques, which often struggle with rapid updates and complex interactions, or rely on external storage to achieve long-term retention, thereby increasing storage requirements. In this paper, we propose **SELF-PARAM** (Self-Updatable Large Language Models with Parameter Integration). SELF-PARAM requires no extra parameters while ensuring nea

r-optimal efficacy and long-term retention. Our method employs a training object ive that minimizes the Kullback-Leibler (KL) divergence between the predictions of an original model (with access to contextual information) and a target model (without such access). By generating diverse question-answer pairs related to th e knowledge and minimizing the KL divergence across this dataset, we update the target model to internalize the knowledge seamlessly within its parameters. Eval uations on question-answering and conversational recommendation tasks demonstrat e that SELF-PARAM significantly outperforms existing methods, even when accounti ng for non-zero storage requirements. This advancement paves the way for more ef ficient and scalable integration of experiences in large language models by embe dding knowledge directly into model parameters.
****

Kehua Feng,Keyan Ding,Jing Yu,Yiwen Qu,Zhiwen Chen,chengfei lv,Gang Yu,Qiang Zha ng,Huajun Chen
SaMer: A Scenario-aware Multi-dimensional Evaluator for Large Language Models
Evaluating the response quality of large language models (LLMs) for open-ended q uestions poses a significant challenge, especially given the subjectivity and mu lti-dimensionality of "quality" in natural language generation. Existing LLM eva luators often neglect that different scenarios require distinct evaluation crite ria. In this work, we propose **SaMer**, a scenario-aware multi-dimensional eval uator designed to provide both overall and fine-grained assessments of LLM-gener ated responses. Unlike fixed-dimension evaluation approaches, SaMer adapts to di fferent scenarios by automatically identifying and prioritizing relevant evaluat ion dimensions tailored to the given query. To achieve this, we construct a larg e-scale fine-grained preference dataset spanning multiple real-world scenarios, each with distinct evaluation dimensions. We then leverage a text embedding mode l combined with three specialized heads to predict the appropriate evaluation di mensions and corresponding scores, as well as the respective weights that contri bute to the overall score. The resulting model offers fine-grained and interpret able evaluations and shows robust adaptability across diverse scenarios. Extensi ve experiments on eight single rating and pairwise comparison datasets demonstra te that SaMer outperforms existing baselines in a variety of evaluation tasks, s howcasing its robustness, versatility, and generalizability.
****

Chengxin Wang,Yiran Zhao,Shaofeng Cai,Gary Tan
Investigating Pattern Neurons in Urban Time Series Forecasting
Urban time series forecasting is crucial for smart city development and is key t o sustainable urban management. Although urban time series models (UTSMs) are ef fective in general forecasting, they often overlook low-frequency events, such a s holidays and extreme weather, leading to degraded performance in practical app lications. In this paper, we first investigate how UTSMs handle these infrequent patterns from a neural perspective. Based on our findings, we propose $\textbf{ P}$attern $\textbf{N}$euron guided $\textbf{Train}$ing ($\texttt{PN-Train}$), a novel training method that features (i) a $\textit{perturbation-based detector}$ to identify neurons responsible for low-frequency patterns in UTSMs, and (ii) a $\textit{fine-tuning mechanism}$ that enhances these neurons without compromisi ng representation learning on high-frequency patterns. Empirical results demonst rate that $\texttt{PN-Train}$ considerably improves forecasting accuracy for low -frequency events while maintaining high performance for high-frequency events. The code is available at https://github.com/cwang-nus/PN-Train.
****

Zichen Wang,Yaokun Ji,Jianing Tian,Shuangjia Zheng
Retrieval Augmented Diffusion Model for Structure-informed Antibody Design and O ptimization
Antibodies are essential proteins responsible for immune responses in organisms, capable of specifically recognizing antigen molecules of pathogens. Recent adva nces in generative models have significantly enhanced rational antibody design. However, existing methods mainly create antibodies from scratch without template constraints, leading to model optimization challenges and unnatural sequences. To address these issues, we propose a retrieval-augmented diffusion framework, t

ermed RADAb, for efficient antibody design. Our method leverages a set of struct
ural homologous motifs that align with query structural constraints to guide the
 generative model in inversely optimizing antibodies according to desired design
 criteria. Specifically, we introduce a structure-informed retrieval mechanism t
hat integrates these exemplar motifs with the input backbone through a novel dua
l-branch denoising module, utilizing both structural and evolutionary informatio
n. Additionally, we develop a conditional diffusion model that iteratively refin
es the optimization process by incorporating both global context and local evolu
tionary conditions. Our approach is agnostic to the choice of generative models.
 Empirical experiments demonstrate that our method achieves state-of-the-art per
formance in multiple antibody inverse folding and optimization tasks, offering a
 new perspective on biomolecular generative models.
****
Yang Liu,Chuanchen Luo,Zhongkai Mao,Junran Peng,Zhaoxiang Zhang
CityGaussianV2: Efficient and Geometrically Accurate Reconstruction for Large-Sc
ale Scenes
Recently, 3D Gaussian Splatting (3DGS) has revolutionized radiance field reconst
ruction, manifesting efficient and high-fidelity novel view synthesis. However,
accurately representing surfaces, especially in large and complex scenarios, rem
ains a significant challenge due to the unstructured nature of 3DGS. In this pap
er, we present CityGaussianV2, a novel approach for large-scale scene reconstruc
tion that addresses critical challenges related to geometric accuracy and effici
ency. Building on the favorable generalization capabilities of 2D Gaussian Splat
ting (2DGS), we address its convergence and scalability issues. Specifically, we
 implement a decomposed-gradient-based densification and depth regression techni
que to eliminate blurry artifacts and accelerate convergence. To scale up, we in
troduce an elongation filter that mitigates Gaussian count explosion caused by 2
DGS degeneration. Furthermore, we optimize the CityGaussian pipeline for paralle
l training, achieving up to 10$\times$ compression, at least 25\% savings in tra
ining time, and a 50\% decrease in memory usage. We also established standard ge
ometry benchmarks under large-scale scenes. Experimental results demonstrate tha
t our method strikes a promising balance between visual quality, geometric accur
acy, as well as storage and training costs.
****
Xiang Yue,Yueqi Song,Akari Asai,Seungone Kim,Jean de Dieu Nyandwi,Simran Khanuja
,Anjali Kantharuban,Lintang Sutawika,Sathyanarayanan Ramamoorthy,Graham Neubig
Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages
Despite recent advances in multimodal large language models (MLLMs), their devel
opment has predominantly focused on English- and western-centric datasets and ta
sks, leaving most of the world's languages and diverse cultural contexts underre
presented.
This paper introduces PANGEA, a multilingual multimodal LLM trained on PANGEAINS
, a diverse 6M instruction dataset spanning 39 languages. PANGEAINS features: 1)
 high-quality English instructions, 2) carefully machine-translated instructions
, and 3) culturally relevant multimodal tasks to ensure cross-cultural coverage.

To rigorously assess models' capabilities, we introduce PANGEABENCH, a holistic
evaluation suite encompassing 14 datasets covering 47 languages.
Results show that PANGEA significantly outperforms existing open-source models i
n multilingual settings and diverse cultural contexts. Ablation studies further
reveal the importance of English data proportions, language popularity, and the
number of multimodal training samples on overall performance. We fully open-sou
rce our data, code, and trained checkpoints, to facilitate the development of in
clusive and robust multilingual MLLMs, promoting equity and accessibility across
 a broader linguistic and cultural spectrum.
****
Yue Wu,Zhiqing Sun,Huizhuo Yuan,Kaixuan Ji,Yiming Yang,Quanquan Gu
Self-Play Preference Optimization for Language Model Alignment
Standard reinforcement learning from human feedback (RLHF) approaches relying on
 parametric models like the Bradley-Terry model fall short in capturing the intr

ansitivity and irrationality in human preferences. Recent advancements suggest t
hat directly working with preference probabilities can yield a more accurate ref
lection of human preferences, enabling more flexible and accurate language model
 alignment. In this paper, we propose a self-play-based method for language mode
l alignment, which treats the problem as a constant-sum two-player game aimed at
 identifying the Nash equilibrium policy. Our approach, dubbed *Self-Play Prefer
ence Optimization* (SPPO), utilizes iterative policy updates to provably approxi
mate the Nash equilibrium.
Additionally, we propose a new SPPO objective which is both strongly motivated b
y theory and is simple and effective in practice.
In our experiments, using only 60k prompts (without responses) from the UltraFee
dback dataset and without any prompt augmentation, by leveraging a pre-trained p
reference model PairRM with only 0.4B parameters, SPPO can obtain a model from f
ine-tuning Mistral-7B-Instruct-v0.2 that achieves the state-of-the-art length-co
ntrolled win-rate of 28.53\% against GPT-4-Turbo on AlpacaEval 2.0. It also outp
erforms the (iterative) DPO and IPO on MT-Bench, Arena-Hard, and the Open LLM Le
aderboard.
Starting from a stronger base model Llama-3-8B-Instruct, we are able to achieve
a length-controlled win rate of 38.77\%.
Notably, the strong performance of SPPO is achieved without additional external
supervision (e.g., responses, preferences, etc.) from GPT-4 or other stronger la
nguage models.
****

Wenxi Lv,Qinliang Su,Wenchao Xu
One-for-All Few-Shot Anomaly Detection via Instance-Induced Prompt Learning
Anomaly detection methods under the 'one-for-all' paradigm aim to develop a unif
ied model capable of detecting anomalies across multiple classes. However, these
 approaches typically require a large number of normal samples for model trainin
g, which may not always be feasible in practice. Few-shot anomaly detection meth
ods can address scenarios with limited data but often require a tailored model f
or each class, struggling within the 'one-for-one' paradigm. In this paper, we f
irst proposed the one-for-all few-shot anomaly detection method with the assista
nce of vision-language model. Different from previous CLIP-based methods learnin
g fix prompts for each class, our method learn a class-shared prompt generator t
o adaptively generate suitable prompt for each instance. The prompt generator is
 trained by aligning the prompts with the visual space and utilizing guidance fr
om general textual descriptions of normality and abnormality. Furthermore, we ad
dress the mismatch problem of the memory bank within one-for-all paradigm. Exten
sive experimental results on MVTec and VisA demonstrate the superiority of our m
ethod in few-shot anomaly detection task under the one-for-all paradigm.
****

Yongqi Ding,Lin Zuo,Mengmeng Jing,Pei He,Hanpu Deng
Rethinking Spiking Neural Networks from an Ensemble Learning Perspective
Spiking neural networks (SNNs) exhibit superior energy efficiency but suffer fro
m limited performance. In this paper, we consider SNNs as ensembles of temporal
subnetworks that share architectures and weights, and highlight a crucial issue
that affects their performance: excessive differences in initial states (neurona
l membrane potentials) across timesteps lead to unstable subnetwork outputs, res
ulting in degraded performance. To mitigate this, we promote the consistency of
the initial membrane potential distribution and output through membrane potentia
l smoothing and temporally adjacent subnetwork guidance, respectively, to improv
e overall stability and performance. Moreover, membrane potential smoothing faci
litates forward propagation of information and backward propagation of gradients
, mitigating the notorious temporal gradient vanishing problem. Our method requi
res only minimal modification of the spiking neurons without adapting the networ
k structure, making our method generalizable and showing consistent performance
gains in 1D speech, 2D object, and 3D point cloud recognition tasks. In particul
ar, on the challenging CIFAR10-DVS dataset, we achieved 83.20\% accuracy with on
ly four timesteps. This provides valuable insights into unleashing the potential
 of SNNs.

****
Hyogon Ryu,NaHyeon Park,Hyunjung Shim
DGQ: Distribution-Aware Group Quantization for Text-to-Image Diffusion Models
Despite the widespread use of text-to-image diffusion models across various task
s, their computational and memory demands limit practical applications.
To mitigate this issue, quantization of diffusion models has been explored. It r
educes memory usage and computational costs by compressing weights and activatio
ns into lower-bit formats.
However, existing methods often struggle to preserve both image quality and text
-image alignment, particularly in lower-bit($<$ 8bits) quantization.
In this paper, we analyze the challenges associated with quantizing text-to-imag
e diffusion models from a distributional perspective. Our analysis reveals that
activation outliers play a crucial role in determining image quality.
Additionally, we identify distinctive patterns in cross-attention scores, which
significantly affects text-image alignment.
To address these challenges, we propose Distribution-aware Group Quantization (D
GQ), a method that identifies and adaptively handles pixel-wise and channel-wise
 outliers to preserve image quality. Furthermore, DGQ applies prompt-specific lo
garithmic quantization scales to maintain text-image alignment.
Our method demonstrates remarkable performance on datasets such as MS-COCO and P
artiPrompts. We are the first to successfully achieve low-bit quantization of te
xt-to-image diffusion models without requiring additional fine-tuning of weight
quantization parameters. Code is available at \link{https://github.com/ugonfor/D
GQ}.
****
Deqing Fu,Tong Xiao,Rui Wang,Wang Zhu,Pengchuan Zhang,Guan Pang,Robin Jia,Lawren
ce Chen
TLDR: Token-Level Detective Reward Model for Large Vision Language Models
Although reward models have been successful in improving multimodal large langua
ge models, the reward models themselves remain brutal and contain minimal inform
ation. Notably, existing reward models only mimic human annotations by assigning
 only one feedback to any text, no matter how long the text is. In the realm of
multimodal language models, where models are required to process both images and
 texts, a naive reward model may learn implicit biases toward texts and become l
ess grounded in images. In this paper, we propose a **T**oken-**L**evel **D**ete
ctive **R**eward Model (**TLDR**) to provide fine-grained annotations to each te
xt token. We first introduce a perturbation-based method to generate synthetic h
ard negatives and their token-level labels to train TLDR models. Then we show th
e rich usefulness of TLDR models both in assisting off-the-shelf models to self-
correct their generations, and in serving as a hallucination evaluation tool. We
 show that TLDR automatically trains a token-level likelihood optimization, and
can improve the base model's performance significantly. Finally, we show that TL
DR models can significantly speed up human annotation by 3 times to acquire a br
oader range of high-quality vision language data.
****
Wenjing Yan,Kai Zhang,Xiaolu Wang,Xuanyu Cao
Problem-Parameter-Free Federated Learning
Federated learning (FL) has garnered significant attention from academia and ind
ustry in recent years due to its advantages in data privacy, scalability, and co
mmunication efficiency. However, current FL algorithms face a critical limitatio
n: their performance heavily depends on meticulously tuned hyperparameters, part
icularly the learning rate or stepsize. This manual tuning process is challengin
g in federated settings due to data heterogeneity and limited accessibility of l
ocal datasets. Consequently, the reliance on problem-specific parameters hinders
 the widespread adoption of FL and potentially compromises its performance in dy
namic or diverse environments. To address this issue, we introduce PAdaMFed, a n
ovel algorithm for nonconvex FL that carefully combines adaptive stepsize and mo
mentum techniques. PAdaMFed offers two key advantages: 1) it operates autonomous
ly without relying on problem-specific parameters; and 2) it manages data hetero
geneity and partial participation without requiring heterogeneity bounds. Despit

e these benefits, PAdaMFed provides several strong theoretical guarantees: 1) It achieves state-of-the-art convergence rates with a sample complexity of $\mathcal{O}(\epsilon^{-4})$ and communication complexity of $\mathcal{O}(\epsilon^{-3})$ to obtain an accuracy of $||\nabla f\left(\boldsymbol{\theta}\right)|| \leq \epsilon$, even using constant learning rates; 2) these complexities can be improved to the best-known $\mathcal{O}(\epsilon^{-3})$ for sampling and $\mathcal{O}(\epsilon^{-2})$ for communication when incorporating variance reduction; 3) it exhibits linear speedup with respect to the number of local update steps and participating clients at each global round. These attributes make PAdaMFed highly scalable and adaptable for various real-world FL applications. Extensive empirical evidence on both image classification and sentiment analysis tasks validates the efficacy of our approaches.
****

Jie Wang,Tingfa Xu,Lihe Ding,Xinjie Zhang,Long Bai,Jianan Li
PvNeXt: Rethinking Network Design and Temporal Motion for Point Cloud Video Recognition
Point cloud video perception has become an essential task for the realm of 3D vision. Current 4D representation learning techniques typically engage in iterative processing coupled with dense query operations. Although effective in capturing temporal features, this approach leads to substantial computational redundancy. In this work, we propose a framework, named as PvNeXt, for effective yet efficient point cloud video recognition, via personalized one-shot query operation. Specially, PvNeXt consists of two key modules, the Motion Imitator and the Single-Step Motion Encoder. The former module, the Motion Imitator, is designed to capture the temporal dynamics inherent in sequences of point clouds, thus generating the virtual motion corresponding to each frame. The Single-Step Motion Encoder performs a one-step query operation, associating point cloud of each frame with its corresponding virtual motion frame, thereby extracting motion cues from point cloud sequences and capturing temporal dynamics across the entire sequence. Through the integration of these two modules, {PvNeXt} enables personalized one-shot queries for each frame, effectively eliminating the need for frame-specific looping and intensive query processes. Extensive experiments on multiple benchmarks demonstrate the effectiveness of our method.
****

Vijay Lingam,Behrooz Omidvar Tehrani,Sujay Sanghavi,Gaurav Gupta,Sayan Ghosh,Linbo Liu,Jun Huan,Anoop Deoras
Enhancing Language Model Agents using Diversity of Thoughts
A popular approach to building agents using Language Models (LMs) involves iteratively prompting the LM, reflecting on its outputs, and updating the input prompts until the desired task is achieved. However, our analysis reveals two key shortcomings in the existing methods: (i) limited exploration of the decision space due to repetitive reflections, which result in redundant inputs, and (ii) an inability to leverage insights from previously solved tasks. To address these issues, we introduce DoT (Diversity of Thoughts), a novel framework that a) explicitly reduces redundant reflections to enhance decision-space exploration, and b) incorporates a task-agnostic memory component to enable knowledge retrieval from previously solved tasks—unlike current approaches that operate in isolation for each task. Through extensive experiments on a suite of programming benchmarks (HumanEval, MBPP, and LeetCodeHardGym) using a variety of LMs, DoT demonstrates up to a $\textbf{10}$% improvement in Pass@1 while maintaining cost-effectiveness. Furthermore, DoT is modular by design. For instance, when the diverse reflection module of DoT is integrated with existing methods like Tree of Thoughts (ToT), we observe a significant $\textbf{13}$% improvement on Game of 24 (one of the main benchmarks of ToT), highlighting the broad applicability and impact of our contributions across various reasoning tasks.
****

Chenghua Huang,Zhizhen Fan,Lu Wang,Fangkai Yang,Pu Zhao,Zeqi Lin,Qingwei Lin,Dongmei Zhang,Saravan Rajmohan,Qi Zhang
SELF-EVOLVED REWARD LEARNING FOR LLMS
Reinforcement Learning from Human Feedback (RLHF) is a crucial technique for ali

gning language models with human preferences and is a key factor in the success of modern conversational models like GPT-4, ChatGPT, and Llama 2. A significant challenge in employing RLHF lies in training a reliable RM, which relies on high-quality labels. Typically, these labels are provided by human experts or a stronger AI, both of which can be costly and introduce bias that may affect the language model's responses. As models improve, human input may become less effective in enhancing their performance. This paper explores the potential of using the RM itself to generate additional training data for a more robust RM. Our experiments demonstrate that reinforcement learning from self-feedback outperforms baseline approaches.
We conducted extensive experiments with our approach on multiple datasets, such as HH-RLHF and UltraFeedback, and models including Mistral and Llama 3, comparing it against various baselines. Our results indicate that, even with a limited amount of human-labeled data, learning from self-feedback can robustly enhance the performance of the RM, thereby improving the capabilities of large language models.
****
Cristian Rodriguez-Opazo,Ehsan Abbasnejad,Damien Teney,Hamed Damirchi,Edison Marrese-Taylor,Anton van den Hengel
Synergy and Diversity in CLIP: Enhancing Performance Through Adaptive Backbone Ensembling
Contrastive Language-Image Pretraining (CLIP) stands out as a prominent method for image representation learning. Various architectures, from vision transformers~(ViTs) to convolutional networks (ResNets) have been trained with CLIP to serve as general solutions to diverse vision tasks.
This paper explores the differences across various CLIP-trained vision backbones.
Despite using the same data and training objective, we find that these architectures have notably different representations,
different classification performance across datasets, and different robustness properties to certain types of image perturbations.
Our findings indicate a remarkable possible synergy across backbones
by leveraging their respective strengths.
In principle, classification accuracy could be improved by over 40 percentage with an informed selection of the optimal backbone per test example.
Using this insight, we develop a straightforward yet powerful approach to adaptively ensemble multiple backbones.
The approach uses as few as one labeled example per class
to tune the adaptive combination of backbones.
On a large collection of datasets, the method achieves a remarkable increase in accuracy of up to 39.1\% over the best single backbone, well beyond traditional ensembles.
****
Thomas Robert,Mher Safaryan,Ionut-Vlad Modoranu,Dan Alistarh
LDAdam: Adaptive Optimization from Low-Dimensional Gradient Statistics
We introduce LDAdam, a memory-efficient optimizer for training large models, that performs adaptive optimization steps within lower dimensional subspaces, while consistently exploring the full parameter space during training. This strategy keeps the optimizer's memory footprint to a fraction of the model size. LDAdam relies on a new projection-aware update rule for the optimizer states that allows for transitioning between subspaces, i.e., estimation of the statistics of the projected gradients. To mitigate the errors due to low-rank projection, LDAdam integrates a new generalized error feedback mechanism, which explicitly accounts for both gradient and optimizer state compression. We prove the convergence of LDAdam under standard assumptions, and provide empirical evidence that LDAdam allows for efficient fine-tuning and pre-training of language models.
****
Zhixin Lai,Keqiang Sun,Fu-Yun Wang,Dhritiman Sagar,Erli Ding
InstantPortrait: One-Step Portrait Editing via Diffusion Multi-Objective Distillation

Real-time instruction-based portrait image editing is crucial in various applica
tions, including filters, augmented reality, and video communications, etc. Howe
ver, real-time portrait editing presents three significant challenges: identity
preservation, fidelity to editing instructions, and fast model inference. Given
that these aspects often present a trade-off, concurrently addressing them poses
 an even greater challenge. While diffusion-based image editing methods have sho
wn promising capabilities in personalized image editing in recent years, they la
ck a dedicated focus on portrait editing and thus suffer from the aforementioned
 problems as well. To address the gap, this paper introduces an Instant-Portrait
 Network (IPNet), the first one-step diffusion-based model for portrait editing.
 We train the network in two stages. We first employ an annealing identity loss
to train an Identity Enhancement Network (IDE-Net), to ensure robust identity pr
eservation. We then train the IPNet using a novel diffusion Multi-Objective Dist
illation approach that integrates adversarial loss, identity distillation loss,
and a novel Facial-Style Enhancing loss. The Diffusion Multi-Objective Distillat
ion approach efficiently reduces inference steps, ensures identity consistency,
and enhances the precision of instruction-based editing. Extensive comparison wi
th prior models demonstrates IPNet as a superior model in terms of identity pres
ervation, text fidelity, and inference speed.
****

Kunhao Zheng,Juliette Decugis,Jonas Gehring,Taco Cohen,benjamin negrevergne,Gabr
iel Synnaeve
What Makes Large Language Models Reason in (Multi-Turn) Code Generation?
Prompting techniques such as chain-of-thought have established themselves as a p
opular vehicle for improving the outputs of large language models (LLMs). For co
de generation, however, their exact mechanics and efficacy are under-explored us
ing unified metrics and benchmarks. We thus investigate the effects of a wide ra
nge of prompting strategies with a focus on automatic re-prompting over multiple
 turns and computational requirements. After systematically decomposing reasonin
g, instruction, and execution feedback prompts, we conduct an extensive grid sea
rch on the competitive programming benchmarks CodeContests and TACO for multiple
 LLM families and sizes (Llama 3.0 and 3.1, 8B, 70B, 405B, and GPT-4o). Our stud
y reveals strategies that consistently improve performance across all models wit
h small and large sampling budgets. We then show how finetuning with such an opt
imal configuration allows models to internalize the induced reasoning process an
d obtain improvements in performance and scalability for multi-turn code generat
ion.
****

Beomsu Kim,Jaemin Kim,Jeongsol Kim,Jong Chul Ye
Generalized Consistency Trajectory Models for Image Manipulation
Diffusion-based generative models excel in unconditional generation, as well as
on applied tasks such as image editing and restoration. The success of diffusion
 models lies in the iterative nature of diffusion: diffusion breaks down the com
plex process of mapping noise to data into a sequence of simple denoising tasks.
 Moreover, we are able to exert fine-grained control over the generation process
 by injecting guidance terms into each denoising step. However, the iterative pr
ocess is also computationally intensive, often taking from tens up to thousands
of function evaluations. Although consistency trajectory models (CTMs) enable tr
aversal between any time points along the probability flow ODE (PFODE) and score
 inference with a single function evaluation, CTMs only allow translation from G
aussian noise to data. Thus, this work aims to unlock the full potential of CTMs
 by proposing generalized CTMs (GCTMs), which translate between arbitrary distri
butions via ODEs. We discuss the design space of GCTMs and demonstrate their eff
icacy in various image manipulation tasks such as image-to-image translation, re
storation, and editing. Code is available at https://github.com/1202kbs/GCTM.
****

Michael Crawshaw,Mingrui Liu
Complexity Lower Bounds of Adaptive Gradient Algorithms for Non-convex Stochasti
c Optimization under Relaxed Smoothness
Recent results in non-convex stochastic optimization demonstrate the convergence

of popular adaptive algorithms (e.g., AdaGrad) under the (L_0, L_1)-smoothness condition, but the rate of convergence is a higher-order polynomial in terms of problem parameters like the smoothness constants. The complexity guaranteed by such algorithms to find an ϵ-stationary point may be significantly larger than the optimal complexity of $\Theta \left(\Delta L \sigma^2 \epsilon^{-4} \right)$ achieved by SGD in the L-smooth setting, where Δ is the initial optimality gap, σ^2 is the variance of stochastic gradient. However, it is currently not known whether these higher-order dependencies can be tightened. To answer this question, we investigate complexity lower bounds for several adaptive optimization algorithms in the (L_0, L_1)-smooth setting, with a focus on the dependence in terms of problem parameters Δ, L_0, L_1. We provide complexity bounds for three variations of AdaGrad, which show at least a quadratic dependence on problem parameters Δ, L_0, L_1. Notably, we show that the decorrelated variant of AdaGrad-Norm requires at least $\Omega \left(\Delta^2 L_1^2 \sigma^2 \epsilon^{-4} \right)$ stochastic gradient queries to find an ϵ-stationary point. We also provide a lower bound for SGD with a broad class of adaptive stepsizes. Our results show that, for certain adaptive algorithms, the (L_0, L_1)-smooth setting is fundamentally more difficult than the standard smooth setting, in terms of the initial optimality gap and the smoothness constants.
****
Muhammad Anwar Ma'sum,Mahardhika Pratama,Lin Liu,H Habibullah,Ryszard Kowalczyk
Federated Few-Shot Class-Incremental Learning
This study proposes a challenging yet practical Federated Few-Shot Class-Incremental Learning (FFSCIL) problem, where clients only hold very few samples for new classes. We develop a novel Unified Optimized Prototype Prompt (UOPP) model to simultaneously handle catastrophic forgetting, over-fitting, and prototype bias in FFSCIL. UOPP utilizes task-wise prompt learning to mitigate task interference and over-fitting, unified static-dynamic prototypes to achieve a stability-plasticity balance, and adaptive dual heads for enhanced inferences. Dynamic prototypes represent new classes in the current few-shot task and are rectified to deal with prototype bias. Our comprehensive experimental results show that UOPP significantly outperforms state-of-the-art (SOTA) methods on three datasets with improvements up to 76% on average accuracy and 90% on harmonic mean accuracy respectively. Our extensive analysis shows UOPP robustness in various numbers of local clients and global rounds, low communication costs, and moderate running time. The source code of UOPP is publicly available at https://github.com/anwarmaxsum/FFSCIL.
****
Jason Ramapuram,Federico Danieli,Eeshan Gunesh Dhekane,Floris Weers,Dan Busbridge,Pierre Ablin,Tatiana Likhomanenko,Jagrit Digani,Zijin Gu,Amitis Shidani,Russell Webb
Theory, Analysis, and Best Practices for Sigmoid Self-Attention
Attention is a key part of the transformer architecture. It is a sequence-to-sequence mapping that transforms each sequence element into a weighted sum of values. The weights are typically obtained as the softmax of dot products between keys and queries. Recent work has explored alternatives to softmax attention in transformers, such as ReLU and sigmoid activations. In this work, we revisit sigmoid attention and conduct an in-depth theoretical and empirical analysis. Theoretically, we prove that transformers with sigmoid attention are universal function approximators and benefit from improved regularity compared to softmax attention. Through detailed empirical analysis, we identify stabilization of large initial attention norms during the early stages of training as a crucial factor for the successful training of models with sigmoid attention, outperforming prior attempts. We also introduce FLASHSIGMOID, a hardware-aware and memory-efficient implementation of sigmoid attention yielding a 17% inference kernel speed-up over FLASHATTENTION2 on H100 GPUs. Experiments across language, vision, and speech show that properly normalized sigmoid attention matches the strong performance of softmax attention on a wide range of domains and scales, which previous attempts at sigmoid attention were unable to fully achieve. Our work unifies prior art and

establishes best practices for sigmoid attention as a drop-in softmax replaceme
nt in transformers.
****

Jaron Maene,Vincent Derkinderen,Pedro Zuidberg Dos Martires
KLay: Accelerating Arithmetic Circuits for Neurosymbolic AI
A popular approach to neurosymbolic AI involves mapping logic formulas to arithm
etic circuits (computation graphs consisting of sums and products) and passing t
he outputs of a neural network through these circuits. This approach enforces sy
mbolic constraints onto a neural network in a principled and end-to-end differen
tiable way. Unfortunately, arithmetic circuits are challenging to run on modern
tensor accelerators as they exhibit a high degree of irregular sparsity. To addr
ess this limitation, we introduce knowledge layers (KLay), a new data structure
to represent arithmetic circuits that can be efficiently parallelized on GPUs. M
oreover, we contribute two algorithms used in the translation of traditional cir
cuit representations to KLay and a further algorithm that exploits parallelizati
on opportunities during circuit evaluations. We empirically show that KLay achie
ves speedups of multiple orders of magnitude over the state of the art, thereby
paving the way towards scaling neurosymbolic AI to larger real-world application
s.
****

Yoad Tewel,Rinon Gal,Dvir Samuel,Yuval Atzmon,Lior Wolf,Gal Chechik
Add-it: Training-Free Object Insertion in Images With Pretrained Diffusion Model
s
Adding Object into images based on text instructions is a challenging task in se
mantic image editing, requiring a balance between preserving the original scene
and seamlessly integrating the new object in a fitting location. Despite extensi
ve efforts, existing models often struggle with this balance, particularly with
finding a natural location for adding an object in complex scenes. We introduce
Add-it, a training-free approach that extends diffusion models' attention mechan
isms to incorporate information from three key sources: the scene image, the tex
t prompt, and the generated image itself. Our weighted extended-attention mechan
ism maintains structural consistency and fine details while ensuring natural obj
ect placement. Without task-specific fine-tuning, Add-it achieves state-of-the-a
rt results on both real and generated image insertion benchmarks, including our
newly constructed "Additing Affordance Benchmark" for evaluating object placemen
t plausibility, outperforming supervised methods. Human evaluations show that Ad
d-it is preferred in over 80% of cases, and it also demonstrates improvements in
various automated metrics.
****

Shengyu Feng,Xiang Kong,Shuang Ma,Aonan Zhang,Dong Yin,Chong Wang,Ruoming Pang,Y
iming Yang
Step-by-Step Reasoning for Math Problems via Twisted Sequential Monte Carlo
Augmenting the multi-step reasoning abilities of Large Language Models (LLMs) ha
s been a persistent challenge. Recently, verification has shown promise in impro
ving solution consistency by evaluating generated outputs. However, current veri
fication approaches suffer from sampling inefficiencies, requiring a large numbe
r of samples to achieve satisfactory performance. Additionally, training an effe
ctive verifier often depends on extensive process supervision, which is costly t
o acquire. In this paper, we address these limitations by introducing a novel ve
rification method based on Twisted Sequential Monte Carlo (TSMC). TSMC sequentia
lly refines its sampling effort to focus exploration on promising candidates, re
sulting in more efficient generation of high-quality solutions. We apply TSMC to
LLMs by estimating the expected future rewards at partial solutions. This appro
ach results in a more straightforward training target that eliminates the need f
or step-wise human annotations. We empirically demonstrate the advantages of our
method across multiple math benchmarks, and also validate our theoretical analy
sis of both our approach and existing verification methods.
****

Prakash Chandra Chhipa,Gautam Vashishtha,Settur Jithamanyu,Rajkumar Saini,Mubara
k Shah,Marcus Liwicki

ASTrA: Adversarial Self-supervised Training with Adaptive-Attacks
Existing self-supervised adversarial training (self-AT) methods rely on hand-crafted adversarial attack strategies for PGD attacks, which fail to adapt to the evolving learning dynamics of the model and do not account for instance-specific characteristics of images. This results in sub-optimal adversarial robustness and limits the alignment between clean and adversarial data distributions. To address this, we propose $\textit{ASTrA}$ ($\textbf{A}$dversarial $\textbf{S}$elf-supervised $\textbf{Tr}$aining with $\textbf{A}$daptive-Attacks), a novel framework introducing a learnable, self-supervised attack strategy network that autonomously discovers optimal attack parameters through exploration-exploitation in a single training episode. ASTrA leverages a reward mechanism based on contrastive loss, optimized with REINFORCE, enabling adaptive attack strategies without labeled data or additional hyperparameters. We further introduce a mixed contrastive objective to align the distribution of clean and adversarial examples in representation space. ASTrA achieves state-of-the-art results on CIFAR10, CIFAR100, and STL10 while integrating seamlessly as a plug-and-play module for other self-AT methods. ASTrA shows scalability to larger datasets, demonstrates strong semi-supervised performance, and is resilient to robust overfitting, backed by explainability analysis on optimal attack strategies. Project page for source code and other details at https://prakashchhipa.github.io/projects/ASTrA.
****
Xingrun Xing,Boyan Gao,Zheng Liu,David A. Clifton,Shitao Xiao,Wanpeng Zhang,Li Du,Zheng Zhang,Guoqi Li,Jiajun Zhang
SpikeLLM: Scaling up Spiking Neural Network to Large Language Models via Saliency-based Spiking
Recent advancements in large language models (LLMs) with billions of parameters have improved performance in various applications, but their inference processes demand significant energy and computational resources. In contrast, the human brain, with approximately 86 billion neurons, is much more energy-efficient than LLMs with similar parameters. Inspired by this, we redesign 7$\sim$70 billion parameter LLMs using bio-plausible spiking mechanisms, emulating the efficient behavior of the human brain. We propose the first spiking large language model, SpikeLLM. Coupled with the proposed model, two essential approaches are proposed to improve spike training efficiency: Generalized Integrate-and-Fire (GIF) neurons to compress spike length from T to $\frac{T}{L} \log_2 L$ bits, and an Optimal Brain Spiking framework to divide outlier channels and allocate different T for GIF neurons, which further compresses spike length to approximate log_2T bits. The necessity of spike-driven LLM is proved by comparison with quantized LLMs with similar operations. In the OmniQuant pipeline, SpikeLLM reduces 11.01\% WikiText2 perplexity and improves 2.55\% accuracy of common scene reasoning on a LLAMA-7B W4A4 model. In the GPTQ pipeline, SpikeLLM achieves direct additive in linear layers, significantly exceeding PB-LLMs. Our code is publicly available at https://github.com/Xingrun-Xing2/SpikeLLM.
****
Haotian Wu,Gongpu Chen,Deniz Gunduz
Actions Speak Louder Than Words: Rate-Reward Trade-off in Markov Decision Processes
The impact of communication on decision-making systems has been extensively studied under the assumption of dedicated communication channels. We instead consider communicating through actions, where the message is embedded into the actions of an agent which interacts with the environment in a Markov decision process (MDP) framework. We conceptualize the MDP environment as a finite-state channel (FSC), where the actions of the agent serve as the channel input, while the states of the MDP observed by another agent (i.e., receiver) serve as the channel output. Here, we treat the environment as a communication channel over which the agent communicates through its actions, while at the same time, trying to maximize its reward. We first characterize the optimal information theoretic trade-off between the average reward and the rate of reliable communication in the infinite-horizon regime. Then, we propose a novel framework to design a joint control/coding policy, termed Act2Comm, which seamlessly embeds messages into actions. From

a communication perspective, Act2Comm functions as a learning-based channel coding scheme for non-differentiable FSCs under input-output constraints. From a control standpoint, Act2Comm learns an MDP policy that incorporates communication capabilities, though at the cost of some control performance. Overall, Act2Comm effectively balances the dual objectives of control and communication in this environment. Experimental results validate Act2Comm's capability to enable reliable communication while maintaining a certain level of control performance.
****
Shuhong Zheng,Zhipeng Bao,Ruoyu Zhao,Martial Hebert,Yu-Xiong Wang
Diff-2-in-1: Bridging Generation and Dense Perception with Diffusion Models
Beyond high-fidelity image synthesis, diffusion models have recently exhibited promising results in dense visual perception tasks. However, most existing work treats diffusion models as a standalone component for perception tasks, employing them either solely for off-the-shelf data augmentation or as mere feature extractors. In contrast to these isolated and thus sub-optimal efforts, we introduce a integrated, versatile, diffusion-based framework, Diff-2-in-1, that can simultaneously handle both multi-modal data generation and dense visual perception, through a unique exploitation of the diffusion-denoising process. Within this framework, we further enhance discriminative visual perception via multi-modal generation, by utilizing the denoising network to create multi-modal data that mirror the distribution of the original training set. Importantly, Diff-2-in-1 optimizes the utilization of the created diverse and faithful data by leveraging a novel self-improving learning mechanism. Comprehensive experimental evaluations validate the effectiveness of our framework, showcasing consistent performance improvements across various discriminative backbones and high-quality multi-modal data generation characterized by both realism and usefulness. Our project website is available at https://zsh2000.github.io/diff-2-in-1.github.io/.
****
Sangyun Lee,Yilun Xu,Tomas Geffner,Giulia Fanti,Karsten Kreis,Arash Vahdat,Weili Nie
Truncated Consistency Models
Consistency models have recently been introduced to accelerate the generation speed of diffusion models by directly predicting the solution (data) of the probability flow ODE (PF ODE) from initial noise.
However, the training of consistency models requires learning to map all intermediate points along PF ODE trajectories to their corresponding endpoints. This task is much more challenging than the ultimate objective of one-step generation, which only concerns the PF ODE's noise-to-data mapping.
We empirically find that this training paradigm limits the one-step generation performance of consistency models.
To address this issue, we generalize consistency training to the truncated time range, which allows the model to ignore denoising tasks at earlier time steps and focus its capacity on generation.
We propose a new parameterization of the consistency function and a two-stage training procedure that prevent the truncated-time training from collapsing to a trivial solution.
Experiments on CIFAR-10 and ImageNet 64×64 datasets show that our method achieves better one-step and two-step FIDs than the state-of-the-art consistency models such as iCT-deep,
using more than 2$\times$ smaller networks.
****
Clémentine Carla Juliette Dominé,Nicolas Anguita,Alexandra Maria Proca,Lukas Braun,Daniel Kunin,Pedro A. M. Mediano,Andrew M Saxe
From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks
Biological and artificial neural networks develop internal representations that enable them to perform complex tasks. In artificial networks, the effectiveness of these models relies on their ability to build task specific representation, a process influenced by interactions among datasets, architectures, initialization strategies, and optimization algorithms. Prior studies highlight that different initializations can place networks in either a lazy regime, where representati

ons remain static, or a rich/feature learning regime, where representations evolve dynamically. Here, we examine how initialization influences learning dynamics in deep linear neural networks, deriving exact solutions for lambda-balanced initializations-defined by the relative scale of weights across layers. These solutions capture the evolution of representations and the Neural Tangent Kernel across the spectrum from the rich to the lazy regimes. Our findings deepen the theoretical understanding of the impact of weight initialization on learning regimes, with implications for continual learning, reversal learning, and transfer learning, relevant to both neuroscience and practical applications.
****
Mingyang Liu,Gabriele Farina,Asuman E. Ozdaglar
A Policy-Gradient Approach to Solving Imperfect-Information Games with Best-Iterate Convergence
Policy gradient methods have become a staple of any single-agent reinforcement learning toolbox, due to their combination of desirable properties: iterate convergence, efficient use of stochastic trajectory feedback, and theoretically-sound avoidance of importance sampling corrections. In multi-agent imperfect-information settings (extensive-form games), however, it is still unknown whether the same desiderata can be guaranteed while retaining theoretical guarantees. Instead, sound methods for extensive-form games rely on approximating \emph{counterfactual} values (as opposed to Q values), which are incompatible with policy gradient methodologies. In this paper, we investigate whether policy gradient can be safely used in two-player zero-sum imperfect-information extensive-form games (EFGs). We establish positive results, showing for the first time that a policy gradient method leads to provable best-iterate convergence to a regularized Nash equilibrium in self-play.
****
Shuyang Jiang,Yusheng Liao,Ya Zhang,Yanfeng Wang,Yu Wang
Fine-tuning with Reserved Majority for Noise Reduction
Parameter-efficient fine-tuning (PEFT) has revolutionized supervised fine-tuning, where LoRA and its variants gain the most popularity due to their low training costs and zero inference latency.
However, LoRA tuning not only injects knowledgeable features but also noisy hallucination during fine-tuning, which hinders the utilization of tunable parameters with the increasing LoRA rank.
In this work, we first investigate in-depth the redundancies among LoRA parameters with substantial empirical studies.
Aiming to resemble the learning capacity of high ranks from the findings, we set up a new fine-tuning framework, \textbf{P}arameter-\textbf{Re}dundant \textbf{Fine-\textbf{T}uning (\preft), which follows the vanilla LoRA tuning process but is required to reduce redundancies before merging LoRA parameters back to pre-trained models.
Based on this framework, we propose \textbf{No}ise reduction with \textbf{R}eserved \textbf{M}ajority~(\norm), which decomposes the LoRA parameters into majority parts and redundant parts with random singular value decomposition.
The major components are determined by the proposed \search method, specifically employing subspace similarity to confirm the parameter groups that share the highest similarity with the base weight.
By employing \norm, we enhance both the learning capacity and benefits from larger ranks, which consistently outperforms both LoRA and other \preft-based methods on various downstream tasks, such as general instruction tuning, math reasoning and code generation.
Code is available at \url{https://github.com/pixas/NoRM}.
****
Zhiteng Li,Xianglong Yan,Tianao Zhang,Haotong Qin,Dong Xie,Jiang Tian,zhongchao shi,Linghe Kong,Yulun Zhang,Xiaokang Yang
ARB-LLM: Alternating Refined Binarizations for Large Language Models
Large Language Models (LLMs) have greatly pushed forward advancements in natural language processing, yet their high memory and computational demands hinder practical deployment. Binarization, as an effective compression technique, can shri

nk model weights to just 1 bit, significantly reducing the high demands on compu
tation and memory. However, current binarization methods struggle to narrow the
distribution gap between binarized and full-precision weights, while also overlo
oking the column deviation in LLM weight distribution. To tackle these issues, w
e propose ARB-LLM, a novel 1-bit post-training quantization (PTQ) technique tail
ored for LLMs. To narrow the distribution shift between binarized and full-preci
sion weights, we first design an alternating refined binarization (ARB) algorith
m to progressively update the binarization parameters, which significantly reduc
es the quantization error. Moreover, considering the pivot role of calibration d
ata and the column deviation in LLM weights, we further extend ARB to ARB-X and
ARB-RC. In addition, we refine the weight partition strategy with column-group b
itmap (CGB), which further enhance performance. Equipping ARB-X and ARB-RC with
CGB, we obtain ARB-LLM$_{\text{X}}$ and ARB-LLM$ _{\text{RC}} $ respectively, wh
ich significantly outperform state-of-the-art (SOTA) binarization methods for LL
Ms.
As a binary PTQ method, our ARB-LLM$ _{\text{RC}} $ is the first to surpass FP16
 models of the same size. Code: https://github.com/ZHITENGLI/ARB-LLM.
****
Yongchang Hao,Mengyao Zhai,Hossein Hajimirsadeghi,Sepidehsadat Hosseini,Frederic
k Tung
Radar: Fast Long-Context Decoding for Any Transformer
Transformer models have demonstrated exceptional performance across a wide range
 of applications. Though forming the foundation of Transformer models, the dot-p
roduct attention does not scale well to long-context data since its time require
ment grows quadratically with context length. In this work, we propose Radar, a
training-free approach that accelerates inference by dynamically searching for t
he most important context tokens. For any pre-trained Transformer, Radar can red
uce the decoding time complexity without training or heuristically evicting toke
ns. Moreover, we provide theoretical justification for our approach, demonstrati
ng that Radar can reliably identify the most important tokens with high probabil
ity. We conduct extensive comparisons with the previous methods on a wide range
of tasks. The results demonstrate that Radar achieves the state-of-the-art perfo
rmance across different architectures with reduced time complexity, offering a p
ractical solution for efficient long-context processing of Transformers. The cod
e is publicly available at https://github.com/BorealisAI/radar-decoding.
****
Eugene Choi,Arash Ahmadian,Matthieu Geist,Olivier Pietquin,Mohammad Gheshlaghi A
zar
Self-Improving Robust Preference Optimization
Online and offline $\mathtt{RLHF}$ methods, such as $\mathtt{PPO}$ and $\mathtt{
DPO}$, have been highly successful in aligning AI with human preferences. Despit
e their success, however, these methods suffer from fundamental limitations: $\m
athbf{(a)}$ Models trained with $\mathtt{RLHF}$ can learn from mistakes or negat
ive examples through RL mechanism or contrastive loss during training. However,
at inference time, they lack an innate self-improvement mechanism for error corr
ections. $\mathbf{(b)}$ The optimal solution of existing methods is highly task-
dependent, making it difficult for them to generalize to new tasks. To address t
hese challenges, we propose Self-Improving Robust Preference Optimization ($\mat
htt{SRPO}$), a practical and mathematically principled offline $\mathtt{RLHF}$ f
ramework. The key idea behind $\mathtt{SRPO}$ is to cast the problem of learning
 from human preferences as a self-improvement process, mathematically formulated
 as a min-max objective that jointly optimizes a self-improvement policy and a g
enerative policy in an adversarial fashion. Crucially, the solution for this opt
imization problem is independent of the training task, which makes it robust to
its changes. We then show that this objective can be reformulated as a non-adver
sarial offline loss, which can be efficiently optimized using standard supervise
d learning techniques at scale. To demonstrate $\mathtt{SRPO}$’s effectiveness,
we evaluate it using AI Win-Rate (WR) against human (GOLD) completions. When tes
ted on the XSum dataset, $\mathtt{SRPO}$ outperforms $\mathtt{DPO}$ by a margin
of $\mathbf{15}$% after 5 self-revisions, achieving an impressive $\mathbf{90}

$% WR. Moreover, on the challenging Arena-Hard prompts, $\mathtt{SRPO}$ outperfo
rms both $\mathtt{DPO}$ and $\mathtt{IPO}$ (by $\mathbf{4}$% without revision an
d $\mathbf{6}$% after a single revision), reaching a $\mathbf{56}$% WR against a
gainst $\mathtt{Llama-3.1-8B-Instruct}$.
****
Konstantinos Karatapanis,Vasilis Kontonis,Christos Tzamos
Oracle efficient truncated statistics
We study the problem of learning from truncated samples: instead of observing
samples from some underlying population $p^\ast$, we observe only the examples t
hat fall in some survival set $S \subset \mathbb{R}^d$ whose probability mass (m
easured with respect to $p^\ast$) is at least α. Assuming membership ora
cle access to the truncation set S, prior works obtained algorithms for the ca
se where $p^\ast$ is Gaussian or more generally an exponential family with stron
gly convex likelihood --- albeit with a super-polynomial
dependency on the (inverse) survival mass $1/\alpha$
both in terms of runtime and in number of oracle calls to the set S. In this
work we design a new learning method with runtime and query complexity polynomia
l in $1/\alpha$.
Our result significantly improves over the prior works
by focusing on efficiently solving the underlying optimization problem using a g
eneral
purpose optimization algorithm with minimal assumptions.
****
Byung Hyun Lee,Sungjin Lim,Seunggyu Lee,Dong Un Kang,Se Young Chun
Concept Pinpoint Eraser for Text-to-image Diffusion Models via Residual Attentio
n Gate
Remarkable progress in text-to-image diffusion models has brought a major concer
n about potentially generating images on inappropriate or trademarked concepts.
Concept erasing has been investigated with the goals of deleting target concepts
 in diffusion models while preserving other concepts with minimal distortion. To
 achieve these goals, recent concept erasing methods usually fine-tune the cross
-attention layers of diffusion models. In this work, we first show that merely u
pdating the cross-attention layers in diffusion models, which is mathematically
equivalent to adding linear modules to weights, may not be able to preserve dive
rse remaining concepts. Then, we propose a novel framework, dubbed Concept Pinpo
int Eraser (CPE), by adding nonlinear Residual Attention Gates (ResAGs) that sel
ectively erase (or cut) target concepts while safeguarding remaining concepts fr
om broad distributions by employing an attention anchoring loss to prevent the f
orgetting. Moreover, we adversarially train CPE with ResAG and learnable text em
beddings in an iterative manner to maximize erasing performance and enhance robu
stness against adversarial attacks. Extensive experiments on the erasure of cele
brities, artistic styles, and explicit contents demonstrated that the proposed C
PE outperforms prior arts by keeping diverse remaining concepts while deleting t
he target concepts with robustness against attack prompts. Code is available at
https://github.com/Hyun1A/CPE.
****
Hyuk Park,Grani A. Hanasusanto,Yingying Li
Robust System Identification: Finite-sample Guarantees and Connection to Regular
ization
We consider the problem of learning nonlinear dynamical systems from a single sa
mple trajectory. While the least squares estimate (LSE) is commonly used for thi
s task, it suffers from poor identification errors when the sample size is small
 or the model fails to capture the system's true dynamics. To overcome these lim
itations, we propose a robust LSE framework, which incorporates robust optimizat
ion techniques, and prove that it is equivalent to regularizing LSE using genera
l Schatten p-norms. We provide non-asymptotic performance guarantees for linea
r systems, achieving an error rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$, and
 show that it avoids the curse of dimensionality, unlike state-of-the-art Wasser
stein robust optimization models. Empirical results demonstrate substantial impr
ovements in real-world system identification and online control tasks, outperfor

ming existing methods.
****
Saumya Gupta,Dimitris Samaras,Chao Chen
TopoDiffusionNet: A Topology-aware Diffusion Model
Diffusion models excel at creating visually impressive images but often struggle to generate images with a specified topology. The Betti number, which represents the number of structures in an image, is a fundamental measure in topology. Yet, diffusion models fail to satisfy even this basic constraint. This limitation restricts their utility in applications requiring exact control, like robotics and environmental modeling. To address this, we propose TopoDiffusionNet (TDN), a novel approach that enforces diffusion models to maintain the desired topology. We leverage tools from topological data analysis, particularly persistent homology, to extract the topological structures within an image. We then design a topology-based objective function to guide the denoising process, preserving intended structures while suppressing noisy ones. Our experiments across four datasets demonstrate significant improvements in topological accuracy. TDN is the first to integrate topology with diffusion models, opening new avenues of research in this area.
****
Zijia Zhao,Haoyu Lu,Yuqi Huo,Yifan Du,Tongtian Yue,Longteng Guo,Bingning Wang,weipeng chen,Jing Liu
Needle In A Video Haystack: A Scalable Synthetic Evaluator for Video MLLMs
Video understanding is a crucial next step for multimodal large language models (MLLMs).
Various benchmarks are introduced for better evaluating the MLLMs.
Nevertheless, current video benchmarks are still inefficient for evaluating video models during iterative development due to the high cost of constructing datasets and the difficulty in isolating specific skills.
In this paper, we propose VideoNIAH (Video Needle in A Haystack), a benchmark construction framework through synthetic video generation.
VideoNIAH decouples video content from their query-responses by inserting unrelated visual 'needles' into original videos.
The framework automates the generation of query-response pairs using predefined rules, minimizing manual labor. The queries focus on specific aspects of video understanding, enabling more skill-specific evaluations. The separation between video content and the queries also allow for increased video variety and evaluations across different lengths.
Utilizing VideoNIAH, we compile a video benchmark, VNBench, which includes tasks such as retrieval, ordering, and counting to evaluate three key aspects of video understanding: temporal perception, chronological ordering, and spatio-temporal coherence. We conduct a comprehensive evaluation of both proprietary and open-source models, uncovering significant differences in their video understanding capabilities across various tasks. Additionally, we perform an in-depth analysis of the test results and model configurations. Based on these findings, we provide some advice for improving video MLLM training, offering valuable insights to guide future research and model development.
****
Hanling Tian,Yuhang Liu,Mingzhen He,Zhengbao He,Zhehao Huang,Ruikai Yang,Xiaolin Huang
Simulating Training Dynamics to Reconstruct Training Data from Deep Neural Networks
Whether deep neural networks (DNNs) memorize the training data is a fundamental open question in understanding deep learning. A direct way to verify the memorization of DNNs is to reconstruct training data from DNNs’ parameters. Since parameters are gradually determined by data throughout training, characterizing training dynamics is important for reconstruction. Pioneering works rely on the linear training dynamics of shallow NNs with large widths, but cannot be extended to more practical DNNs which have non-linear dynamics. We propose Simulation of training Dynamics (SimuDy) to reconstruct training data from DNNs. Specifically, we simulate the training dynamics by training the model from the initial parameter

s with a dummy dataset, then optimize this dummy dataset so that the simulated d
ynamics reach the same final parameters as the true dynamics. By incorporating d
ummy parameters in the simulated dynamics, SimuDy effectively describes non-line
ar training dynamics. Experiments demonstrate that SimuDy significantly outperfo
rms previous approaches when handling non-linear training dynamics, and for the
first time, most training samples can be reconstructed from a trained ResNet’s p
arameters.
****
Zongyue Qin,Ziniu Hu,Zifan He,Neha Prakriya,Jason Cong,Yizhou Sun
Optimized Multi-Token Joint Decoding With Auxiliary Model for LLM Inference
Large language models (LLMs) have achieved remarkable success across diverse tas
ks, yet their inference processes are hindered by substantial time and energy de
mands due to single-token generation at each decoding step. While previous metho
ds such as speculative decoding mitigate these inefficiencies by producing multi
ple tokens per step, each token is still generated by its single-token distribut
ion,
thereby enhancing speed without improving effectiveness. In contrast, our work s
imultaneously enhances inference speed and improves the output effectiveness. We
 consider multi-token joint decoding (MTJD), which generates multiple tokens fro
m their joint distribution at each iteration, theoretically reducing perplexity
and enhancing task performance. However, MTJD suffers from the high cost of samp
ling from the joint distribution of multiple tokens. Inspired by speculative dec
oding, we introduce multi-token assisted decoding (MTAD), a novel framework desi
gned to accelerate MTJD. MTAD leverages a smaller auxiliary model to approximate
 the joint distribution of a larger model, incorporating a verification mechanis
m that not only ensures the accuracy of this approximation, but also improves th
e
decoding efficiency over conventional speculative decoding. Theoretically, we de
monstrate that MTAD closely approximates exact MTJD with bounded error. Empirica
l evaluations using Llama-2 and OPT models ranging from 13B to 70B parameters ac
ross various tasks reveal that MTAD reduces perplexity by 21.2% and improves dow
nstream performance compared to standard single-token sampling.
Furthermore, MTAD achieves a 1.42× speed-up and consumes 1.54× less energy than
conventional speculative decoding methods. These results highlight MTAD’s abilit
y to make multi-token joint decoding both effective and efficient, promoting mor
e sustainable and high-performance deployment of LLMs.
****
Gaurav Sahu,Abhay Puri,Juan A. Rodriguez,Amirhossein Abaskohi,Mohammad Chegini,A
lexandre Drouin,Perouz Taslakian,Valentina Zantedeschi,Alexandre Lacoste,David V
azquez,Nicolas Chapados,Christopher Pal,Sai Rajeswar,Issam H. Laradji
InsightBench: Evaluating Business Analytics Agents Through Multi-Step Insight Ge
neration
Data analytics is essential for extracting valuable insights from data that can
assist organizations in making effective decisions. We introduce InsightBench, a
 benchmark dataset with three key features. First, it consists of 100 datasets r
epresenting diverse business use cases such as finance and incident management,
each accompanied by a carefully curated set of insights planted in the datasets.
 Second, unlike existing benchmarks focusing on answering single queries, Insigh
tBench evaluates agents based on their ability to perform end-to-end data analyt
ics, including formulating questions, interpreting answers, and generating a sum
mary of insights and actionable steps. Third, we conducted comprehensive quality
 assurance to ensure that each dataset in the benchmark had clear goals and incl
uded relevant and meaningful questions and analysis. Furthermore, we implement a
 two-way evaluation mechanism using LLaMA-3 as an effective, open-source evaluat
or to assess agents’ ability to extract insights. We also propose AgentPoirot, o
ur baseline data analysis agent capable of performing end-to-end data analytics.
 Our evaluation on InsightBench shows that AgentPoirot outperforms existing appr
oaches (such as Pandas Agent) that focus on resolving single queries. We also co
mpare the performance of open- and closed-source LLMs and various evaluation str
ategies. Overall, this benchmark serves as a testbed to motivate further develop

ment in comprehensive automated data analytics and can be accessed here: https://github.com/ServiceNow/insight-bench.
****
Jingyang Zhang,Jingwei Sun,Eric Yeats,Yang Ouyang,Martin Kuo,Jianyi Zhang,Hao Frank Yang,Hai Li
Min-K%++: Improved Baseline for Pre-Training Data Detection from Large Language Models
The problem of pre-training data detection for large language models (LLMs) has received growing attention due to its implications in critical issues like copyright violation and test data contamination. Despite improved performance, existing methods (including the state-of-the-art, Min-K%) are mostly developed upon simple heuristics and lack solid, reasonable foundations. In this work, we propose a novel and theoretically motivated methodology for pre-training data detection, named Min-K%++. Specifically, we present a key insight that training samples tend to be local maxima of the modeled distribution along each input dimension through maximum likelihood training, which in turn allow us to insightfully translate the problem into identification of local maxima. Then, we design our method accordingly that works under the discrete distribution modeled by LLMs, whose core idea is to determine whether the input forms a mode or has relatively high probability under the conditional categorical distribution. Empirically, the proposed method achieves new SOTA performance across multiple settings (evaluated with 5 families of 10 models and 2 benchmarks). On the WikiMIA benchmark, Min-K%++ outperforms the runner-up by 6.2% to 10.5% in detection AUROC averaged over five models. On the more challenging MIMIR benchmark, it consistently improves upon reference-free methods while performing on par with reference-based method that requires an extra reference model.
****
Bin Xie,Yingfei Liu,Tiancai Wang,Jiale Cao,Xiangyu Zhang
Glad: A Streaming Scene Generator for Autonomous Driving
The generation and simulation of diverse real-world scenes have significant application value in the field of autonomous driving, especially for the corner cases. Recently, researchers have explored employing neural radiance fields or diffusion models to generate novel views or synthetic data under driving scenes. However, these approaches suffer from unseen scenes or restricted video length, thus lacking sufficient adaptability for data generation and simulation. To address these issues, we propose a simple yet effective framework, named Glad, to generate video data in a frame-by-frame style. To ensure the temporal consistency of synthetic video, we introduce a latent variable propagation module, which views the latent features of previous frame as noise prior and injects it into the latent features of current frame. In addition, we design a streaming data sampler to orderly sample the original image in a video clip at continuous iterations. Given the reference frame, our Glad can be viewed as a streaming simulator by generating the videos for specific scenes. Extensive experiments are performed on the widely-used nuScenes dataset. Experimental results demonstrate that our proposed Glad achieves promising performance, serving as a strong baseline for online video generation. We will release the source code and models publicly.
****
Hannah Lawrence,Vasco Portilheiro,Yan Zhang,Sékou-Oumar Kaba
Improving Equivariant Networks with Probabilistic Symmetry Breaking
Equivariance encodes known symmetries into neural networks, often enhancing generalization. However, equivariant networks cannot *break* symmetries: the output of an equivariant network must, by definition, have at least the same self-symmetries as its input. This poses an important problem, both (1) for prediction tasks on domains where self-symmetries are common, and (2) for generative models, which must break symmetries in order to reconstruct from highly symmetric latent spaces. This fundamental limitation can in fact be addressed by considering *equivariant conditional distributions*, instead of equivariant functions. We therefore present novel theoretical results that establish necessary and sufficient conditions for representing such distributions. Concretely, this representation provides a practical framework for breaking symmetries in any equivariant network

via randomized canonicalization. Our method, SymPE (Symmetry-breaking Positional Encodings), admits a simple interpretation in terms of positional encodings. This approach expands the representational power of equivariant networks while retaining the inductive bias of symmetry, which we justify through generalization bounds. Experimental results demonstrate that SymPE significantly improves performance of group-equivariant and graph neural networks across diffusion models for graphs, graph autoencoders, and lattice spin system modeling.
****
Seunghun Lee,Jinyoung Park,Jaewon Chu,Minseo Yoon,Hyunwoo J. Kim
Latent Bayesian Optimization via Autoregressive Normalizing Flows
Bayesian Optimization (BO) has been recognized for its effectiveness in optimizing expensive and complex objective functions.
Recent advancements in Latent Bayesian Optimization (LBO) have shown promise by integrating generative models such as variational autoencoders (VAEs) to manage the complexity of high-dimensional and structured data spaces.
However, existing LBO approaches often suffer from the value discrepancy problem, which arises from the reconstruction gap between input and latent spaces.
This value discrepancy problem propagates errors throughout the optimization process, leading to suboptimal outcomes.
To address this issue, we propose a Normalizing Flow-based Bayesian Optimization (NF-BO), which utilizes normalizing flow as a generative model to establish one-to-one encoding function from the input space to the latent space, along with its left-inverse decoding function, eliminating the reconstruction gap. Specifically, we introduce SeqFlow, an autoregressive normalizing flow for sequence data. In addition, we develop a new candidate sampling strategy that dynamically adjusts the exploration probability for each token based on its importance.
Through extensive experiments, our NF-BO method demonstrates superior performance in molecule generation tasks, significantly outperforming both traditional and recent LBO approaches.
****
Muhammad Gohar Javed,chuan guo,Li cheng,Xingyu Li
InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling
Generating realistic 3D human-human interactions from textual descriptions remains a challenging task. Existing approaches, typically based on diffusion models, often produce results lacking realism and fidelity. In this work, we introduce *InterMask*, a novel framework for generating human interactions using collaborative masked modeling in discrete space. InterMask first employs a VQ-VAE to transform each motion sequence into a 2D discrete motion token map. Unlike traditional 1D VQ token maps, it better preserves fine-grained spatio-temporal details and promotes *spatial awareness* within each token. Building on this representation, InterMask utilizes a generative masked modeling framework to collaboratively model the tokens of two interacting individuals. This is achieved by employing a transformer architecture specifically designed to capture complex spatio-temporal inter-dependencies. During training, it randomly masks the motion tokens of both individuals and learns to predict them. For inference, starting from fully masked sequences, it progressively fills in the tokens for both individuals. With its enhanced motion representation, dedicated architecture, and effective learning strategy, InterMask achieves state-of-the-art results, producing high-fidelity and diverse human interactions. It outperforms previous methods, achieving an FID of 5.154 (vs 5.535 of in2IN) on the InterHuman dataset and 0.399 (vs 5.207 of InterGen) on the InterX dataset. Additionally, InterMask seamlessly supports reaction generation without the need for model redesign or fine-tuning.
****
Yuta Natsubori,Masataka Ushiku,Yuta Saito
Cross-Domain Off-Policy Evaluation and Learning for Contextual Bandits
Off-Policy Evaluation and Learning (OPE/L) in contextual bandits is rapidly gaining popularity in real systems because new policies can be evaluated and learned securely using only historical logged data. However, existing methods in OPE/L cannot handle many challenging but prevalent scenarios such as few-shot data, deterministic logging policies, and new actions. In many applications, such as per

sonalized medicine, content recommendations, education, and advertising, we need to evaluate and learn new policies in the presence of these challenges. Existing methods cannot evaluate and optimize effectively in these situations due to the notorious variance issue or limited exploration in the logged data. To enable OPE/L even under these unsolved challenges, we propose a new problem setup of Cross-Domain OPE/L, where we have access not only to the logged data from the target domain in which the new policy will be implemented but also to logged datasets collected from other domains. This novel formulation is widely applicable because we can often use historical data not only from the target hospital, country, device, or user segment but also from other hospitals, countries, devices, or segments. We develop a new estimator and policy gradient method to solve OPE/L by leveraging both target and source datasets, resulting in substantially enhanced OPE/L in the previously unsolved situations in our empirical evaluations.
****

Tianyu Xie,Harry Richman,Jiansi Gao,Frederick A Matsen,Cheng Zhang
PhyloVAE: Unsupervised Learning of Phylogenetic Trees via Variational Autoencoders
Learning informative representations of phylogenetic tree structures is essential for analyzing evolutionary relationships. Classical distance-based methods have been widely used to project phylogenetic trees into Euclidean space, but they are often sensitive to the choice of distance metric and may lack sufficient resolution. In this paper, we introduce *phylogenetic variational autoencoders* (PhyloVAEs), an unsupervised learning framework designed for representation learning and generative modeling of tree topologies. Leveraging an efficient encoding mechanism inspired by autoregressive tree topology generation, we develop a deep latent-variable generative model that facilitates fast, parallelized topology generation. PhyloVAE combines this generative model with a collaborative inference model based on learnable topological features, allowing for high-resolution representations of phylogenetic tree samples. Extensive experiments demonstrate PhyloVAE's robust representation learning capabilities and fast generation of phylogenetic tree topologies.
****

Raphael T. Husistein,Markus Reiher,Marco Eckhoff
NEAR: A Training-Free Pre-Estimator of Machine Learning Model Performance
Artificial neural networks have been shown to be state-of-the-art machine learning models in a wide variety of applications, including natural language processing and image recognition. However, building a performant neural network is a laborious task and requires substantial computing power. Neural Architecture Search (NAS) addresses this issue by an automatic selection of the optimal network from a set of potential candidates. While many NAS methods still require training of (some) neural networks, zero-cost proxies promise to identify the optimal network without training. In this work, we propose the zero-cost proxy Network Expressivity by Activation Rank (NEAR). It is based on the effective rank of the pre- and post-activation matrix, i.e., the values of a neural network layer before and after applying its activation function. We demonstrate the cutting-edge correlation between this network score and the model accuracy on NAS-Bench-101 and NATS-Bench-SSS/TSS. In addition, we present a simple approach to estimate the optimal layer sizes in multi-layer perceptrons. Furthermore, we show that this score can be utilized to select hyperparameters such as the activation function and the neural network weight initialization scheme.
****

Hao He,Yinghao Xu,Yuwei Guo,Gordon Wetzstein,Bo Dai,Hongsheng Li,Ceyuan Yang
CameraCtrl: Enabling Camera Control for Video Diffusion Models
Controllability plays a crucial role in video generation, as it allows users to create and edit content more precisely. Existing models, however, lack control of camera pose that serves as a cinematic language to express deeper narrative nuances. To alleviate this issue, we introduce \method, enabling accurate camera pose control for video diffusion models. Our approach explores effective camera trajectory parameterization along with a plug-and-play camera pose control module that is trained on top of a video diffusion model, leaving other modules of the

base model untouched. Moreover, a comprehensive study on the effect of various training datasets is conducted, suggesting that videos with diverse camera distributions and similar appearance to the base model indeed enhance controllability and generalization. Experimental results demonstrate the effectiveness of \method in achieving precise camera control with different video generation models, marking a step forward in the pursuit of dynamic and customized video storytelling from textual and camera pose inputs.
****
Zhanwei Zhang,Shizhao Sun,Wenxiao Wang,Deng Cai,Jiang Bian
FlexCAD: Unified and Versatile Controllable CAD Generation with Fine-tuned Large Language Models
Recently, there is a growing interest in creating computer-aided design (CAD) models based on user intent, known as controllable CAD generation. Existing work offers limited controllability and needs separate models for different types of control, reducing efficiency and practicality. To achieve controllable generation across all CAD construction hierarchies, such as sketch-extrusion, extrusion, sketch, face, loop and curve, we propose FlexCAD, a unified model by fine-tuning large language models (LLMs). First, to enhance comprehension by LLMs, we represent a CAD model as a structured text by abstracting each hierarchy as a sequence of text tokens. Second, to address various controllable generation tasks in a unified model, we introduce a hierarchy-aware masking strategy. Specifically, during training, we mask a hierarchy-aware field in the CAD text with a mask token. This field, composed of a sequence of tokens, can be set flexibly to represent various hierarchies. Subsequently, we ask LLMs to predict this masked field. During inference, the user intent is converted into a CAD text with a mask token replacing the part the user wants to modify, which is then fed into FlexCAD to generate new CAD models.
Comprehensive experiments on public dataset demonstrate the effectiveness of FlexCAD in both generation quality and controllability.
****
Stanislav Fort
Scaling Laws for Adversarial Attacks on Language Model Activations and Tokens
We explore a class of adversarial attacks targeting the activations of language models to derive upper-bound scaling laws on their attack susceptibility. By manipulating a relatively small subset of model activations, a, we demonstrate the ability to control the exact prediction of a significant number (in some cases up to 1000) of subsequent tokens t. We empirically verify a scaling law where the maximum number of target tokens predicted, t_max, depends linearly on the number of tokens a whose activations the attacker controls as $t_\mathrm{max} = \kappa a$. We find that the number of bits the attacker controls on the input to exert a single bit of control on the output (a property we call \textit{attack resistance χ}) is remarkably stable between ≈ 16 and ≈ 25 over orders of magnitude of model sizes and between model families. Compared to attacks directly on input tokens, attacks on activations are predictably much stronger, however, we identify a surprising regularity where one bit of input steered either via activations or via tokens is able to exert a surprisingly similar amount of control over the model predictions. This gives support for the hypothesis that adversarial attacks are a consequence of dimensionality mismatch between the input and output spaces. A practical implication of the ease of attacking language model activations instead of tokens is for multi-modal and selected retrieval models. By using language models as a controllable test-bed to study adversarial attacks, we explored input-output dimension regimes that are inaccessible in computer vision and greatly extended the empirical support for the dimensionality theory of adversarial attacks.
****
Xuandong Zhao,Lei Li,Yu-Xiang Wang
Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs
In this paper, we propose a new decoding method called Permute-and-Flip (PF) decoder. It enjoys stability properties similar to the standard sampling decoder, but is provably up to 2x better in its quality-stability tradeoff than sampling a

nd never worse than any other decoder. We also design a cryptographic watermarki
ng scheme analogous to Aaronson (2023)'s Gumbel watermark, but naturally tailore
d for PF decoder. The watermarking scheme does not change the distribution to sa
mple, while allowing arbitrarily low false positive rate and high recall wheneve
r the generated text has high entropy. Our experiments show that the PF decoder
(and its watermarked counterpart) significantly outperform(s) naive sampling (an
d its Gumbel watermarked counterpart) in terms of perplexity, while retaining th
e same stability (and detectability), hence making it a promising new approach f
or LLM decoding. The code is available at https://github.com/XuandongZhao/pf-dec
oding
****

Ali Hasan,Haoming Yang,Yuting Ng,Vahid Tarokh
Elliptic Loss Regularization
Regularizing neural networks is important for anticipating model behavior in reg
ions of the data space that are not well represented. In this work, we propose a
 regularization technique for enforcing a level of smoothness in the mapping bet
ween the input space and the loss. We specify the level of regularity by requiri
ng that the loss of the network satisfies an elliptic operator over the data dom
ain. To do this, we modify the usual empirical risk minimization objective such
that we instead minimize a new objective that satisfies an elliptic operator ove
r points within the domain. This allows us to use existing theory on elliptic op
erators to anticipate the behavior of the error for points outside the training
set. We propose a tractable computational method that approximates the behavior
of the elliptic operator while being computationally efficient. Finally, we anal
yze the properties of the proposed regularization to understand the performance
on common problems of distribution shift and group imbalance. Numerical experime
nts empirically confirm the promise of the proposed regularization technique.
****

Kanan Gupta,Stephan Wojtowytsch
Nesterov acceleration in benignly non-convex landscapes
While momentum-based optimization algorithms are commonly used in the notoriousl
y non-convex optimization problems of deep learning, their analysis has historic
ally been restricted to the convex and strongly convex setting. In this article,
 we partially close this gap between theory and practice and demonstrate that vi
rtually identical guarantees can be obtained in optimization problems with a `be
nign' non-convexity. We show that these weaker geometric assumptions are well ju
stified in overparametrized deep learning, at least locally. Variations of this
result are obtained for a continuous time model of Nesterov's accelerated gradie
nt descent algorithm (NAG), the classical discrete time version of NAG, and vers
ions of NAG with stochastic gradient estimates with purely additive noise and wi
th noise that exhibits both additive and multiplicative scaling.
****

Tianxu Li,Kun Zhu
Toward Efficient Multi-Agent Exploration With Trajectory Entropy Maximization
Recent works have increasingly focused on learning decentralized policies for ag
ents as a solution to the scalability challenges in Multi-Agent Reinforcement Le
arning (MARL), where agents typically share the parameters of a policy network t
o make action decisions. However, this parameter sharing can impede efficient ex
ploration, as it may lead to similar behaviors among agents. Different from prev
ious mutual information-based methods that promote multi-agent diversity, we int
roduce a novel multi-agent exploration method called Trajectory Entropy Explorat
ion (TEE). Our method employs a particle-based entropy estimator to maximize the
 entropy of different agents' trajectories in a contrastive trajectory represent
ation space, resulting in diverse trajectories and efficient exploration. This e
ntropy estimator avoids challenging density modeling and scales effectively in h
igh-dimensional multi-agent settings. We integrate our method with MARL algorith
ms by deploying an intrinsic reward for each agent to encourage entropy maximiza
tion. To validate the effectiveness of our method, we test our method in challen
ging multi-agent tasks from several MARL benchmarks. The results demonstrate tha
t our method consistently outperforms existing state-of-the-art methods.

****
Lehan Wang,Haonan Wang,Honglong Yang,Jiaji Mao,Zehong Yang,Jun Shen,Xiaomeng Li
Interpretable Bilingual Multimodal Large Language Model for Diverse Biomedical T
asks
Several medical Multimodal Large Languange Models (MLLMs) have been developed to
 address tasks involving visual images with textual instructions across various
medical modalities, achieving impressive results.
Most current medical generalist models are region-agnostic, treating the entire
image as a holistic representation. However, they struggle to identify which spe
cific regions they are focusing on when generating a sentence.
To mimic the behavior of doctors, who typically begin by reviewing the entire im
age before concentrating on specific regions for a thorough evaluation, we aim t
o enhance the capability of medical MLLMs in understanding anatomical regions wi
thin entire medical scans.
To achieve it, we first formulate \textbf{Region-Centric tasks} and construct a
\textbf{large-scale dataset, MedRegInstruct,} to incorporate regional informatio
n into training. Combining our collected dataset with other medical multimodal c
orpora for training, we propose a \textbf{Region-Aware medical MLLM, MedRegA}, w
hich is the first bilingual generalist medical AI system to simultaneously handl
e image-level and region-level medical vision-language tasks across a broad rang
e of modalities. Our MedRegA not only enables three region-centric tasks, but al
so achieves the best performance for visual question answering, report generatio
n and medical image classification over 8 modalities, showcasing significant ver
satility. Experiments demonstrate that our model can not only accomplish powerfu
l performance across various medical vision-language tasks in bilingual settings
, but also recognize and detect structures in multimodal medical scans, boosting
 the interpretability and user interactivity of medical MLLMs. The codes and mod
el will be made publicly available.
****
Aya Abdelsalam Ismail,Tuomas Oikarinen,Amy Wang,Julius Adebayo,Samuel Don Stanto
n,Hector Corrada Bravo,Kyunghyun Cho,Nathan C. Frey
Concept Bottleneck Language Models For Protein Design
We introduce Concept Bottleneck Protein Language Models (CB-pLM), a generative m
asked language model with a layer where each neuron corresponds to an interpreta
ble concept. Our architecture offers three key benefits: i) Control: We can inte
rvene on concept values to precisely control the properties of generated protein
s, achieving a 3$\times$ larger change in desired concept values compared to bas
elines. ii) Interpretability: A linear mapping between concept values and predic
ted tokens allows transparent analysis of the model's decision-making process. i
ii) Debugging: This transparency facilitates easy debugging of trained models. O
ur models achieve pre-training perplexity and downstream task performance compar
able to traditional masked protein language models, demonstrating that interpret
ability does not compromise performance. While adaptable to any language model,
we focus on masked protein language models due to their importance in drug disco
very and the ability to validate our model's capabilities through real-world exp
eriments and expert knowledge. We scale our CB-pLM from 24 million to 3 billion
parameters, making them the largest Concept Bottleneck Models trained and the fi
rst capable of generative language modeling.
****
Yang Liu,Zinan Zheng,Jiashun Cheng,Fugee Tsung,Deli Zhao,Yu Rong,Jia Li
CirT: Global Subseasonal-to-Seasonal Forecasting with Geometry-inspired Transfor
mer
Accurate Subseasonal-to-Seasonal (S2S) climate forecasting is pivotal for decisi
on-making including agriculture planning and disaster preparedness but is known
to be challenging due to its chaotic nature. Although recent data-driven models
have shown promising results, their performance is limited by inadequate conside
ration of geometric inductive biases. Usually, they treat the spherical weather
data as planar images, resulting in an inaccurate representation of locations an
d spatial relations. In this work, we propose the geometric-inspired Circular Tr
ansformer (CirT) to model the cyclic characteristic of the graticule, consisting

of two key designs: (1) Decomposing the weather data by latitude into circular patches that serve as input tokens to the Transformer; (2) Leveraging Fourier transform in self-attention to capture the global information and model the spatial periodicity. Extensive experiments on the Earth Reanalysis 5 (ERA5) reanalysis dataset demonstrate our model yields a significant improvement over the advanced data-driven models, including PanguWeather and GraphCast, as well as skillful ECMWF systems. Additionally, we empirically show the effectiveness of our model designs and high-quality prediction over spatial and temporal dimensions.
****

Terry Yue Zhuo,Vu Minh Chien,Jenny Chim,Han Hu,Wenhao Yu,Ratnadira Widyasari,Imam Nur Bani Yusuf,Haolan Zhan,Junda He,Indraneil Paul,Simon Brunner,Chen GONG,James Hoang,Armel Randy Zebaze,Xiaoheng Hong,Wen-Ding Li,Jean Kaddour,Ming Xu,Zhihan Zhang,Prateek Yadav,Naman Jain,Alex Gu,Zhoujun Cheng,Jiawei Liu,Qian Liu,Zijian Wang,Binyuan Hui,Niklas Muennighoff,David Lo,Daniel Fried,Xiaoning Du,Harm de Vries,Leandro Von Werra

BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions

Task automation has been greatly empowered by the recent advances in Large Language Models (LLMs) via Python code, where the tasks range from software engineering development to general-purpose reasoning. While current benchmarks have shown that LLMs can solve tasks using programs like human developers, the majority of their evaluations are limited to short and self-contained algorithmic tasks or standalone function calls. Solving challenging and practical tasks requires the capability of utilizing **diverse function calls as tools** to efficiently implement functionalities like data analysis and web development. In addition, using multiple tools to solve a task needs compositional reasoning by accurately understanding **complex instructions**. Fulfilling both of these characteristics can pose a great challenge for LLMs. To assess how well LLMs can solve challenging and practical tasks via programs, we introduce BigCodeBench, a benchmark that challenges LLMs to invoke multiple function calls as tools from 139 libraries and 7 domains for 1,140 fine-grained tasks. To evaluate LLMs rigorously, each task encompasses 5.6 test cases with an average branch coverage of 99%. In addition, we propose a natural-language-oriented variant of BigCodeBench, BigCodeBench-Instruct, that automatically transforms the original docstrings into short instructions containing only essential information. Our extensive evaluation of 60 LLMs shows that **LLMs are not yet capable of following complex instructions to use function calls precisely, with scores up to 60%, significantly lower than the human performance of 97%**. The results underscore the need for further advancements in this area.
****

Junjie Li,Yang Liu,Weiqing Liu,Shikai Fang,Lewen Wang,Chang Xu,Jiang Bian

MarS: a Financial Market Simulation Engine Powered by Generative Foundation Model

Generative models aim to simulate realistic effects of various actions across different contexts, from text generation to visual effects. Despite significant efforts to build real-world simulators, the application of generative models to virtual worlds, like financial markets, remains under-explored. In financial markets, generative models can simulate complex market effects of participants with various behaviors, enabling interaction under different market conditions, and training strategies without financial risk. This simulation relies on the finest structured data in financial market like orders thus building the finest realistic simulation. We propose Large Market Model (LMM), an order-level generative foundation model, for financial market simulation, akin to language modeling in the digital world. Our financial Market Simulation engine (MarS), powered by LMM, addresses the domain-specific need for realistic, interactive and controllable order generation. Key observations include LMM's strong scalability across data size and model complexity, and MarS's robust and practicable realism in controlled generation with market impact. We showcase MarS as a forecast tool, detection system, analysis platform, and agent training environment, thus demonstrating MarS's ``paradigm shift'' potential for a variety of financial applications. We rel

ease the code of MarS at https://github.com/microsoft/MarS/.
****
Michelle D Zhao,Henny Admoni,Reid Simmons,Aaditya Ramdas,Andrea Bajcsy
Conformalized Interactive Imitation Learning: Handling Expert Shift and Intermittent Feedback
In interactive imitation learning (IL), uncertainty quantification offers a way for the learner (i.e. robot) to contend with distribution shifts encountered during deployment by actively seeking additional feedback from an expert (i.e. human) online. Prior works use mechanisms like ensemble disagreement or Monte Carlo dropout to quantify when black-box IL policies are uncertain; however, these approaches can lead to overconfident estimates when faced with deployment-time distribution shifts. Instead, we contend that we need uncertainty quantification algorithms that can leverage the expert human feedback received during deployment time to adapt the robot's uncertainty online. To tackle this, we draw upon online conformal prediction, a distribution-free method for constructing prediction intervals online given a stream of ground-truth labels. Human labels, however, are intermittent in the interactive IL setting. Thus, from the conformal prediction side, we introduce a novel uncertainty quantification algorithm called intermittent quantile tracking (IQT) that leverages a probabilistic model of intermittent labels, maintains asymptotic coverage guarantees, and empirically achieves desired coverage levels. From the interactive IL side, we develop ConformalDAgger, a new approach wherein the robot uses prediction intervals calibrated by IQT as a reliable measure of deployment-time uncertainty to actively query for more expert feedback. We compare ConformalDAgger to prior uncertainty-aware DAgger methods in scenarios where the distribution shift is (and isn't) present because of changes in the expert's policy. We find that in simulated and hardware deployments on a 7DOF robotic manipulator, ConformalDAgger detects high uncertainty when the expert shifts and increases the number of interventions compared to baselines, allowing the robot to more quickly learn the new behavior.
****
Jinsu Yoo,Zhenyang Feng,Tai-Yu Pan,Yihong Sun,Cheng Perng Phoo,Xiangyu Chen,Mark Campbell,Kilian Q Weinberger,Bharath Hariharan,Wei-Lun Chao
Learning 3D Perception from Others' Predictions
Accurate 3D object detection in real-world environments requires a huge amount of annotated data with high quality. Acquiring such data is tedious and expensive, and often needs repeated effort when a new sensor is adopted or when the detector is deployed in a new environment. We investigate a new scenario to construct 3D object detectors: *learning from the predictions of a nearby unit that is equipped with an accurate detector.* For example, when a self-driving car enters a new area, it may learn from other traffic participants whose detectors have been optimized for that area. This setting is label-efficient, sensor-agnostic, and communication-efficient: nearby units only need to share the predictions with the ego agent (e.g., car). Naively using the received predictions as ground-truths to train the detector for the ego car, however, leads to inferior performance. We systematically study the problem and identify viewpoint mismatches and mislocalization (due to synchronization and GPS errors) as the main causes, which unavoidably result in false positives, false negatives, and inaccurate pseudo labels. We propose a distance-based curriculum, first learning from closer units with similar viewpoints and subsequently improving the quality of other units' predictions via self-training. We further demonstrate that an effective pseudo label refinement module can be trained with a handful of annotated data, largely reducing the data quantity necessary to train an object detector. We validate our approach on the recently released real-world collaborative driving dataset, using reference cars' predictions as pseudo labels for the ego car. Extensive experiments including several scenarios (e.g., different sensors, detectors, and domains) demonstrate the effectiveness of our approach toward label-efficient learning of 3D perception from other units' predictions.
****
Shang Wang
W-PCA Based Gradient-Free Proxy for Efficient Search of Lightweight Language Mod

els
The demand for efficient natural language processing (NLP) systems has led to th
e development of lightweight language models. Previous work in this area has pri
marily focused on manual design or training-based neural architecture search (NA
S) methods. Recently, zero-shot NAS methods have been proposed for evaluating la
nguage models without the need for training. However, prevailing approaches to z
ero-shot NAS often face challenges such as biased evaluation metrics and computa
tional inefficiencies.
In this paper, we introduce weight-weighted PCA (W-PCA), a novel zero-shot NAS m
ethod specifically tailored for lightweight language models. Our approach utiliz
es two evaluation proxies: the parameter count and the number of principal compo
nents with cumulative contribution exceeding η in the feed-forward neural (
FFN) layer. Additionally, by eliminating the need for gradient computations, we
 optimize the evaluation time, thus enhancing the efficiency of designing and ev
aluating lightweight language models.
We conduct a comparative analysis on the GLUE and SQuAD datasets to evaluate our
 approach. The results demonstrate that our method significantly reduces trainin
g time compared to one-shot NAS methods and achieves higher scores in the testin
g phase compared to previous state-of-the-art training-based methods. Furthermor
e, we perform ranking evaluations on a dataset sampled from the FlexiBERT search
 space. Our approach exhibits superior ranking correlation and further reduces s
olving time compared to other zero-shot NAS methods that require gradient comput
ation.
****
Zeju Qiu,Weiyang Liu,Haiwen Feng,Zhen Liu,Tim Z. Xiao,Katherine M. Collins,Joshu
a B. Tenenbaum,Adrian Weller,Michael J. Black,Bernhard Schölkopf
Can Large Language Models Understand Symbolic Graphics Programs?
Against the backdrop of enthusiasm for large language models (LLMs), there is a
growing need to scientifically assess their capabilities and shortcomings. This
is nontrivial in part because it is difficult to find tasks which the models hav
e not encountered during training. Utilizing symbolic graphics programs, we prop
ose a domain well-suited to test multiple spatial-semantic reasoning skills of L
LMs. Popular in computer graphics, these programs procedurally generate visual d
ata. While LLMs exhibit impressive skills in general program synthesis and analy
sis, symbolic graphics programs offer a new layer of evaluation: they allow us t
o test an LLM's ability to answer semantic questions about the images or 3D geom
etries without a vision encoder. To semantically understand the symbolic program
s, LLMs would need to possess the ability to "imagine" and reason how the corres
ponding graphics content would look with only the symbolic description of the lo
cal curvatures and strokes. We use this task to evaluate LLMs by creating a larg
e benchmark for the semantic visual understanding of symbolic graphics programs,
 built procedurally with minimal human effort. Particular emphasis is placed on
transformations of images that leave the image level semantics invariant while i
ntroducing significant changes to the underlying program. We evaluate commercial
 and open-source LLMs on our benchmark to assess their ability to reason about v
isual output of programs, finding that LLMs considered stronger at reasoning gen
erally perform better. Lastly, we introduce a novel method to improve this abili
ty -- Symbolic Instruction Tuning (SIT), in which the LLM is finetuned with pre-
collected instruction data on symbolic graphics programs. Interestingly, we find
 that SIT not only improves LLM's understanding on symbolic programs, but it als
o improves general reasoning ability on various other benchmarks.
****
Xinyu Zhang,Daolang Huang,Samuel Kaski,Julien Martinelli
PABBO: Preferential Amortized Black-Box Optimization
Preferential Bayesian Optimization (PBO) is a sample-efficient method to learn l
atent user utilities from preferential feedback over a pair of designs. It relie
s on a statistical surrogate model for the latent function, usually a Gaussian p
rocess, and an acquisition strategy to select the next candidate pair to get use
r feedback on. Due to the non-conjugacy of the associated likelihood, every PBO
step requires a significant amount of computations with various approximate infe

rence techniques. This computational overhead is incompatible with the way human s interact with computers, hindering the use of PBO in real-world cases. Buildin g on the recent advances of amortized BO, we propose to circumvent this issue by fully amortizing PBO, meta-learning both the surrogate and the acquisition func tion. Our method comprises a novel transformer neural process architecture, trai ned using reinforcement learning and tailored auxiliary losses.
On a benchmark composed of synthetic and real-world datasets, our method is seve ral orders of magnitude faster than the usual Gaussian process-based strategies and often outperforms them in accuracy.
****

Tinghao Xie,Xiangyu Qi,Yi Zeng,Yangsibo Huang,Udari Madhushani Sehwag,Kaixuan Hu ang,Luxi He,Boyi Wei,Dacheng Li,Ying Sheng,Ruoxi Jia,Bo Li,Kai Li,Danqi Chen,Pet er Henderson,Prateek Mittal
SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal
Evaluating aligned large language models' (LLMs) ability to recognize and reject unsafe user requests is crucial for safe, policy-compliant deployments. Existin g evaluation efforts, however, face three limitations that we address with **SOR RY-Bench**, our proposed benchmark. **First**, existing methods often use coarse -grained taxonomies of unsafe topics, and are over-representing some fine-graine d topics. For example, among the ten existing datasets that we evaluated, tests for refusals of self-harm instructions are over 3x less represented than tests f or fraudulent activities. SORRY-Bench improves on this by using a fine-grained t axonomy of 44 potentially unsafe topics, and 440 class-balanced unsafe instructi ons, compiled through human-in-the-loop methods. **Second**, evaluations often o verlook the linguistic formatting of prompts, like different languages, dialects , and more --- which are only implicitly considered in many evaluations. We supp lement SORRY-bench with 20 diverse linguistic augmentations to systematically ex amine these effects. **Third**, existing evaluations rely on large LLMs (e.g., G PT-4) for evaluation, which can be computationally expensive. We investigate des ign choices for creating a fast, accurate automated safety evaluator. By collect ing 7K+ human annotations and conducting a meta-evaluation of diverse LLM-as-a-j udge designs, we show that fine-tuned 7B LLMs can achieve accuracy comparable to GPT-4 scale LLMs, with lower computational cost. Putting these together, we eva luate over 50 proprietary and open-weight LLMs on SORRY-Bench, analyzing their d istinctive safety refusal behaviors. We hope our effort provides a building bloc k for systematic evaluations of LLMs' safety refusal capabilities, in a balanced , granular, and efficient manner. Benchmark demo, data, code, and models are ava ilable through https://sorry-bench.github.io.
****

Guangkun Nie,Gongzheng Tang,Shenda Hong
Dist Loss: Enhancing Regression in Few-Shot Region through Distribution Distance Constraint
Imbalanced data distributions are prevalent in real-world scenarios, presenting significant challenges in both classification and regression tasks. This imbalan ce often causes deep learning models to overfit in regions with abundant data (m anyshot regions) while underperforming in regions with sparse data (few-shot reg ions). Such characteristics limit the applicability of deep learning models acro ss various domains, notably in healthcare, where rare cases often carry greater clinical significance. While recent studies have highlighted the benefits of inc orporating distributional information in imbalanced classification tasks, simila r strategies have been largely unexplored in imbalanced regression. To address t his gap, we propose Dist Loss, a novel loss function that integrates distributio nal information into model training by jointly optimizing the distribution dista nce between model predictions and target labels, alongside sample-wise predictio n errors. This dual-objective approach encourages the model to balance its predi ctions across different label regions, leading to significant improvements in ac curacy in fewshot regions. We conduct extensive experiments across three dataset s spanning computer vision and healthcare: IMDB-WIKI-DIR, AgeDB-DIR, and ECG-KDI R. The results demonstrate that Dist Loss effectively mitigates the impact of im balanced data distributions, achieving state-of-the-art performance in few-shot

regions. Furthermore, Dist Loss is easy to integrate and complements existing methods. To facilitate further research, we provide our implementation at https://github.com/Ngk03/DIR-Dist-Loss.
****
Ankur Singha,Elia Cellini,Kim Andrea Nicoli,Karl Jansen,Stefan Kühn,Shinichi Nakajima
Multilevel Generative Samplers for Investigating Critical Phenomena
Investigating critical phenomena or phase transitions is of high interest in physics and chemistry, for which Monte Carlo (MC) simulations, a crucial tool for numerically analyzing macroscopic properties of given systems, are often hindered by an emerging divergence of correlation length---known as scale invariance at criticality (SIC) in the renormalization group theory. SIC causes the system to behave the same at any length scale, from which many existing sampling methods suffer: long-range correlations cause critical slowing down in Markov chain Monte Carlo (MCMC), and require intractably large receptive fields for generative samplers. In this paper, we propose a Renormalization-informed Generative Critical Sampler (RiGCS)---a novel sampler specialized for near-critical systems, where SIC is leveraged as an advantage rather than a nuisance. Specifically, RiGCS builds on MultiLevel Monte Carlo (MLMC) with Heat Bath (HB) algorithms, which perform ancestral sampling from low-resolution to high-resolution lattice configurations with site wise-independent conditional HB sampling. Although MLMC-HB is highly efficient under exact SIC, it suffers from a low acceptance rate under slight SIC violation. Notably, SIC violation always occurs in finite-size systems, and may induce long-range and higher-order interactions in the renormalized distributions, which are not considered by independent HB samplers. RiGCS enhances MLMC-HB by replacing a part of the conditional HB sampler with generative models that capture those residual interactions and improve the sampling efficiency. Our experiments show that the effective sample size of RiGCS is a few orders of magnitude higher than state-of-the-art generative model baselines in sampling configurations for 128×128 two-dimensional Ising systems. SIC also allows us to adopt a specialized sequential training protocol with model transfer, which significantly accelerates training.
****
Yawei Li,David Rügamer,Bernd Bischl,Mina Rezaei
Calibrating LLMs with Information-Theoretic Evidential Deep Learning
Fine-tuned large language models (LLMs) often exhibit overconfidence, particularly when trained on small datasets, resulting in poor calibration and inaccurate uncertainty estimates.
Evidential Deep Learning (EDL), an uncertainty-aware approach, enables uncertainty estimation in a single forward pass, making it a promising method for calibrating fine-tuned LLMs. However, despite its computational efficiency, EDL is prone to overfitting, as its training objective can result in overly concentrated probability distributions.
To mitigate this, we propose regularizing EDL by incorporating an information bottleneck (IB). Our approach IB-EDL suppresses spurious information in the evidence generated by the model and encourages truly predictive information to influence both the predictions and uncertainty estimates. Extensive experiments across various fine-tuned LLMs and tasks demonstrate that IB-EDL outperforms both existing EDL and non-EDL approaches.
By improving the trustworthiness of LLMs, IB-EDL facilitates their broader adoption in domains requiring high levels of confidence calibration.
****
Zulfikar Alom,Tran Gia Bao Ngo,Murat Kantarcioglu,Cuneyt Gurcan Akcora
GOttack: Universal Adversarial Attacks on Graph Neural Networks via Graph Orbits Learning
Graph Neural Networks (GNNs) have demonstrated superior performance in node classification tasks across diverse applications. However, their vulnerability to adversarial attacks, where minor perturbations can mislead model predictions, poses significant challenges. This study introduces GOttack, a novel adversarial attack framework that exploits the topological structure of graphs to undermine the

integrity of GNN predictions systematically.

By defining a topology-aware method to manipulate graph orbits, our approach gen
erates adversarial modifications that are both subtle and effective, posing a se
vere test to the robustness of GNNs. We evaluate the efficacy of GOttack across
multiple prominent GNN architectures using standard benchmark datasets. Our resu
lts show that GOttack outperforms existing state-of-the-art adversarial techniqu
es and completes training in approximately 55% of the time required by the faste
st competing model, achieving the highest average misclassification rate in 155
tasks.
This work not only sheds light on the susceptibility of GNNs to structured adver
sarial attacks
but also shows that certain topological patterns may play a significant role in
the underlying robustness of the GNNs. Our Python implementation is shared at ht
tps://github.com/cakcora/GOttack.
****
Chen Henry Wu,Rishi Rajesh Shah,Jing Yu Koh,Russ Salakhutdinov,Daniel Fried,Adit
i Raghunathan
Dissecting Adversarial Robustness of Multimodal LM Agents
As language models (LMs) are used to build autonomous agents in real environment
s, ensuring their adversarial robustness becomes a critical challenge. Unlike ch
atbots, agents are compound systems with multiple components taking actions, whi
ch existing LMs safety evaluations do not adequately address. To bridge this gap
, we manually create 200 targeted adversarial tasks and evaluation scripts in a
realistic threat model on top of VisualWebArena, a real environment for web agen
ts. To systematically examine the robustness of agents, we propose the Agent Rob
ustness Evaluation (ARE) framework. ARE views the agent as a graph showing the f
low of intermediate outputs between components and decomposes robustness as the
flow of adversarial information on the graph. We find that we can successfully b
reak latest agents that use black-box frontier LMs, including those that perform
 reflection and tree search. With imperceptible perturbations to a single image
(less than 5% of total web page pixels), an attacker can hijack these agents to
execute targeted adversarial goals with success rates up to 67%. We also use ARE
 to rigorously evaluate how the robustness changes as new components are added.
We find that inference-time compute that typically improves benign performance c
an open up new vulnerabilities and harm robustness. An attacker can compromise t
he evaluator used by the reflexion agent and the value function of the tree sear
ch agent, which increases the attack success relatively by 15% and 20%. Our data
 and code for attacks, defenses, and evaluation are at https://github.com/ChenWu
98/agent-attack
****
Xingyu Zheng,Xianglong Liu,Haotong Qin,Xudong Ma,Mingyuan Zhang,Haojie Hao,Jiaka
i Wang,Zixiang Zhao,Jinyang Guo,Michele Magno
BinaryDM: Accurate Weight Binarization for Efficient Diffusion Models
With the advancement of diffusion models (DMs) and the substantially increased c
omputational requirements, quantization emerges as a practical solution to obtai
n compact and efficient low-bit DMs. However, the highly discrete representation
 leads to severe accuracy degradation, hindering the quantization of diffusion m
odels to ultra-low bit-widths. This paper proposes a novel weight binarization a
pproach for DMs, namely BinaryDM, pushing binarized DMs to be accurate and effic
ient by improving the representation and optimization. From the representation p
erspective, we present an Evolvable-Basis Binarizer (EBB) to enable a smooth evo
lution of DMs from full-precision to accurately binarized. EBB enhances informat
ion representation in the initial stage through the flexible combination of mult
iple binary bases and applies regularization to evolve into efficient single-bas
is binarization. The evolution only occurs in the head and tail of the DM archit
ecture to retain the stability of training. From the optimization perspective, a
 Low-rank Representation Mimicking (LRM) is applied to assist the optimization o
f binarized DMs. The LRM mimics the representations of full-precision DMs in low
-rank space, alleviating the direction ambiguity of the optimization process cau

sed by fine-grained alignment. Comprehensive experiments demonstrate that Binary DM achieves significant accuracy and efficiency gains compared to SOTA quantizat ion methods of DMs under ultra-low bit-widths. With 1-bit weight and 4-bit activ ation (W1A4), BinaryDM achieves as low as 7.74 FID and saves the performance fro m collapse (baseline FID 10.87). As the first binarization method for diffusion models, W1A4 BinaryDM achieves impressive 15.2x OPs and 29.2x model size savings , showcasing its substantial potential for edge deployment.
****
Hui Yuan,Yifan Zeng,Yue Wu,Huazheng Wang,Mengdi Wang,Liu Leqi
A Common Pitfall of Margin-based Language Model Alignment: Gradient Entanglement
Reinforcement Learning from Human Feedback (RLHF) has become the predominant app roach for aligning language models (LMs) to be more helpful and less harmful.
At its core, RLHF uses a margin-based loss for preference optimization, which sp ecifies the ideal LM behavior only in terms of the difference between preferred and dispreferred responses. In this paper, we identify a common pitfall of margi n-based methods---the under-specification of ideal LM behavior on preferred and dispreferred responses individually, which results in two unintended consequence s as the margin increases:
(1) The probability of dispreferred (e.g., unsafe) responses may increase, resul ting in potential safety alignment failures.
(2) The probability of preferred responses may decrease, even when those respons es are ideal.
We demystify the reasons behind these problematic behaviors: margin-based losses couple the change in the preferred probability with the gradient of the dispref erred one, and vice versa, often preventing the preferred probability from incre asing while the dispreferred one decreases, and thus causing a synchronized incr ease or decrease in both probabilities. We term this effect, inherent in margin-based objectives, gradient entanglement.
Formally, we derive conditions for general margin-based alignment objectives und er which gradient entanglement becomes concerning: the inner product between the gradient of preferred log-probability and the gradient of dispreferred log-prob ability is large relative to the individual gradient norms. Furthermore, we theo retically investigate why such inner products can be large when aligning languag e models and empirically validate our findings. Empirical implications of our fr amework further extend to explaining important differences in the training dynam ics of various preference optimization algorithms and suggesting future directio ns for improvement.
****
Hongxin Zhang,Zeyuan Wang,Qiushi Lyu,Zheyuan Zhang,Sunli Chen,Tianmin Shu,Behzad Dariush,Kwonjoon Lee,Yilun Du,Chuang Gan
COMBO: Compositional World Models for Embodied Multi-Agent Cooperation
In this paper, we investigate the problem of embodied multi-agent cooperation, w here decentralized agents must cooperate given only egocentric views of the worl d. To effectively plan in this setting, in contrast to learning world dynamics i n a single-agent scenario, we must simulate world dynamics conditioned on an arb itrary number of agents' actions given only partial egocentric visual observatio ns of the world. To address this issue of partial observability, we first train generative models to estimate the overall world state given partial egocentric o bservations. To enable accurate simulation of multiple sets of actions on this w orld state, we then propose to learn a compositional world model for multi-agent cooperation by factorizing the naturally composable joint actions of multiple a gents and compositionally generating the video conditioned on the world state. B y leveraging this compositional world model, in combination with Vision Language Models to infer the actions of other agents, we can use a tree search procedure to integrate these modules and facilitate online cooperative planning. We evalu ate our methods on three challenging benchmarks with 2-4 agents. The results sho w our compositional world model is effective and the framework enables the embod ied agents to cooperate efficiently with different agents across various tasks a nd an arbitrary number of agents, showing the promising future of our proposed m ethods. More videos can be found at \url{https://embodied-agi.cs.umass.edu/combo

/}.
****
XIANGYU PENG,Congying Xia,Xinyi Yang,Caiming Xiong,Chien-Sheng Wu,Chen Xing
ReGenesis: LLMs can Grow into Reasoning Generalists via Self-Improvement
Post-training Large Language Models (LLMs) with explicit reasoning trajectories can enhance their reasoning abilities. However, acquiring such high-quality trajectory data typically demands meticulous supervision from humans or superior models, which can be either expensive or license-constrained. In this paper, we explore how far an LLM can improve its reasoning by self-synthesizing reasoning paths as training data without any additional supervision. Existing self-synthesizing methods, such as STaR, suffer from poor generalization to out-of-domain (OOD) reasoning tasks. We hypothesize it is due to that their self-synthesized reasoning paths are too task-specific, lacking general task-agnostic reasoning guidance. To address this, we propose **Reasoning Generalist via Self-Improvement (ReGenesis)**, a method to *self-synthesize reasoning paths as post-training data by progressing from abstract to concrete*. More specifically, ReGenesis self-synthesizes reasoning paths by converting general reasoning guidelines into task-specific ones, generating reasoning structures, and subsequently transforming these structures into reasoning paths, without the need for human-designed task-specific examples used in existing methods. We show that ReGenesis achieves superior performance on all in-domain and OOD settings tested compared to existing methods. For six OOD tasks specifically, while previous methods exhibited an average performance decrease of approximately 4.6% after post training, ReGenesis delivers around 6.1% performance improvement. We also conduct an in-depth analysis of our framework and show ReGenesis is effective across various language models and design choices.
****
Bingcong Li,Liang Zhang,Aryan Mokhtari,Niao He
On the Crucial Role of Initialization for Matrix Factorization
This work revisits the classical low-rank matrix factorization problem and unveils the critical role of initialization in shaping convergence rates for such nonconvex and nonsmooth optimization. We introduce Nystrom initialization, which significantly improves the global convergence of Scaled Gradient Descent (ScaledGD) in both symmetric and asymmetric matrix factorization tasks. Specifically, we prove that ScaledGD with Nystrom initialization achieves quadratic convergence in cases where only linear rates were previously known. Furthermore, we extend this initialization to low-rank adapters (LoRA) commonly used for finetuning foundation models. Our approach, NoRA, i.e., LoRA with Nystrom initialization, demonstrates superior performance across various downstream tasks and model scales, from 1B to 7B parameters, in large language and diffusion models.
****
Yichen Wang,Yiyi Zhang,Xinhao Hu,Li Niu,Jianfu Zhang,Yasushi Makihara,Yasushi Yagi,Pai Peng,Wenlong Liao,Tao He,Junchi Yan,Liqing Zhang
Pedestrian Motion Reconstruction: A Large-scale Benchmark via Mixed Reality Rendering with Multiple Perspectives and Modalities
Reconstructing pedestrian motion from dynamic sensors, with a focus on pedestrian intention, is crucial for advancing autonomous driving safety. However, this task is challenging due to data limitations arising from technical complexities, safety, and cost concerns. We introduce the Pedestrian Motion Reconstruction (PMR) dataset, which focuses on pedestrian intention to reconstruct behavior using multiple perspectives and modalities. PMR is developed from a mixed reality platform that combines real-world realism with the extensive, accurate labels of simulations, thereby reducing costs and risks. It captures the intricate dynamics of pedestrian interactions with objects and vehicles, using different modalities for a comprehensive understanding of human-vehicle interaction. Analyses show that PMR can naturally exhibit pedestrian intent and simulate extreme cases. PMR features a vast collection of data from 54 subjects interacting across 12 urban settings with 7 objects, encompassing 12,138 sequences with diverse weather conditions and vehicle speeds. This data provides a rich foundation for modeling pedestrian intent through multi-view and multi-modal insights. We also conduct compr

ehensive benchmark assessments across different modalities to thoroughly evaluat
e pedestrian motion reconstruction methods.
****
Ivan Anokhin,Rishav Rishav,Matthew Riemer,Stephen Chung,Irina Rish,Samira Ebrahi
mi Kahou
Handling Delay in Real-Time Reinforcement Learning
Real-time reinforcement learning (RL) introduces several challenges. First, poli
cies are constrained to a fixed number of actions per second due to hardware lim
itations. Second, the environment may change while the network is still computin
g an action, leading to observational delay. The first issue can partly be addre
ssed with pipelining, leading to higher throughput and potentially better polici
es. However, the second issue remains: if each neuron operates in parallel with
an execution time of τ, an N-layer feed-forward network experiences obser
vation delay of τN.
Reducing the number of layers can decrease this delay, but at the cost of the ne
twork's expressivity. In this work, we explore the trade-off between minimizing
delay and network's expressivity. We present a theoretically motivated solution
that leverages temporal skip connections combined with history-augmented observa
tions. We evaluate several architectures and show that those incorporating temp
oral skip connections achieve strong performance across various neuron execution
 times, reinforcement learning algorithms, and environments, including four Mujo
co tasks and all MinAtar games. Moreover, we demonstrate parallel neuron computa
tion can accelerate inference by 6-350\% on standard hardware. Our investigatio
n into temporal skip connections and parallel computations paves the way for mor
e efficient RL agents in real-time setting.
****
Nan Zhou,Huiqun Wang,Yaoyan Zheng,Di Huang
Progressive Parameter Efficient Transfer Learning for Semantic Segmentation
Parameter Efficient Transfer Learning (PETL) excels in downstream classification
 fine-tuning with minimal computational overhead, demonstrating its potential wi
thin the pre-train and fine-tune paradigm. However, recent PETL methods consiste
ntly struggle when fine-tuning for semantic segmentation tasks, limiting their b
roader applicability. In this paper, we identify that fine-tuning for semantic s
egmentation requires larger parameter adjustments due to shifts in semantic perc
eption granularity. Current PETL approaches are unable to effectively accommodat
e these shifts, leading to significant performance degradation. To address this,
 we introduce ProPETL, a novel approach that incorporates an additional midstrea
m adaptation to progressively align pre-trained models for segmentation tasks. T
hrough this process, ProPETL achieves state-of-the-art performance on most segme
ntation benchmarks and, for the first time, surpasses full fine-tuning on the ch
allenging COCO-Stuff10k dataset. Furthermore, ProPETL demonstrates strong genera
lization across various pre-trained models and scenarios, highlighting its effec
tiveness and versatility for broader adoption in segmentation tasks. Code is ava
ilable at: https://github.com/weeknan/ProPETL.
****
Zijie Geng,Jie Wang,Ziyan Liu,Siyuan Xu,Zhentao Tang,Shixiong Kai,Mingxuan Yuan,
Jianye HAO,Feng Wu
LaMPlace: Learning to Optimize Cross-Stage Metrics in Macro Placement
Machine learning techniques have shown great potential in enhancing macro placem
ent, a critical stage in modern chip design.
However, existing methods primarily focus on *online* optimization of *intermedi
ate surrogate metrics* that are available at the current placement stage, rather
 than directly targeting the *cross-stage metrics*---such as the timing performa
nce---that measure the final chip quality.
This is mainly because of the high computational costs associated with performin
g post-placement stages for evaluating such metrics, making the *online* optimiz
ation impractical.
Consequently, these optimizations struggle to align with actual performance impr
ovements and can even lead to severe manufacturing issues.
To bridge this gap, we propose **LaMPlace**, which **L**earns **a** **M**ask for

optimizing cross-stage metrics in macro placement.
Specifically, LaMPlace trains a predictor on *offline* data to estimate these *cross-stage metrics* and then leverages the predictor to quickly generate a mask, i.e., a pixel-level feature map that quantifies the impact of placing a macro in each chip grid location on the design metrics.
This mask essentially acts as a fast evaluator, enabling placement decisions based on *cross-stage metrics* rather than *intermediate surrogate metrics*.
Experiments on commonly used benchmarks demonstrate that LaMPlace significantly improves the chip quality across several key design metrics, achieving an average improvement of 9.6\%, notably 43.0\% and 30.4\% in terms of WNS and TNS, respectively, which are two crucial cross-stage metrics that reflect the final chip quality in terms of the timing performance.
****
Maxime Burchi,Radu Timofte
Learning Transformer-based World Models with Contrastive Predictive Coding
The DreamerV3 algorithm recently obtained remarkable performance across diverse environment domains by learning an accurate world model based on Recurrent Neural Networks (RNNs). Following the success of model-based reinforcement learning algorithms and the rapid adoption of the Transformer architecture for its superior training efficiency and favorable scaling properties, recent works such as STORM have proposed replacing RNN-based world models with Transformer-based world models using masked self-attention. However, despite the improved training efficiency of these methods, their impact on performance remains limited compared to the Dreamer algorithm, struggling to learn competitive Transformer-based world models. In this work, we show that the next state prediction objective adopted in previous approaches is insufficient to fully exploit the representation capabilities of Transformers. We propose to extend world model predictions to longer time horizons by introducing TWISTER (Transformer-based World model wIth contraSTiv E Representations), a world model using action-conditioned Contrastive Predictive Coding to learn high-level temporal feature representations and improve the agent performance. TWISTER achieves a human-normalized mean score of 162% on the Atari 100k benchmark, setting a new record among state-of-the-art methods that do not employ look-ahead search. We release our code at https://github.com/burchim/TWISTER.
****
Tiankuo Chu,Fudong Lin,Shubo Wang,Jason Jiang,Wiley Jia-Wei Gong,Xu Yuan,Liyun Wang
BoneMet: An Open Large-Scale Multi-Modal Murine Dataset for Breast Cancer Bone Metastasis Diagnosis and Prognosis
Breast cancer bone metastasis (BCBM) affects women’s health globally, calling for the development of effective diagnosis and prognosis solutions. While deep learning has exhibited impressive capacities across various healthcare domains, its applicability in BCBM diseases is consistently hindered by the lack of an open, large-scale, deep learning-ready dataset. As such, we introduce the Bone Metastasis (BoneMet) dataset, the first large-scale, publicly available, high-resolution medical resource, which is derived from a well-accepted murine BCBM model. The unique advantage of BoneMet over existing human datasets is repeated sequential scans per subject over the entire disease development phases. The dataset consists of over 67 terabytes of multi-modal medical data, including 2D X-ray images, 3D CT scans, and detailed biological data (e.g., medical records and bone quantitative analysis), collected from more than five hundreds mice spanning from 2019 to 2024. Our BoneMet dataset is well-organized into six components, i.e., Rotation X-Ray, Recon-CT, Seg-CT, Regist-CT, RoI-CT, and MiceMediRec. We further show that BoneMet can be readily adopted to build versatile, large-scale AI models for managing BCBM diseases in terms of diagnosis using 2D or 3D images, prognos

is of bone deterioration, and sparse-angle 3D reconstruction for safe long-term disease monitoring. Our preliminary results demonstrate that BoneMet has the potentials to jump-start the development and fine-tuning of AI-driven solutions prior to their applications to human patients. To facilitate its easy access and wide dissemination, we have created the BoneMet package, providing three APIs that enable researchers to (i) flexibly process and download the BoneMet data filtered by specific time frames; and (ii) develop and train large-scale AI models for precise BCBM diagnosis and prognosis. The BoneMet dataset is officially available on Hugging Face Datasets at https://huggingface.co/datasets/BoneMet/BoneMet. The BoneMet package is available on the Python Package Index (PyPI) at https://pypi.org/project/BoneMet. Code and tutorials are available at https://github.com/Tiankuo528/BoneMet.

****

Pierre-David Letourneau,Manish Kumar Singh,Hsin-Pai Cheng,Shizhong Han,Yunxiao Shi,Dalton Jones,Matthew Harper Langston,Hong Cai,Fatih Porikli

PADRe: A Unifying Polynomial Attention Drop-in Replacement for Efficient Vision Transformer

We present Polynomial Attention Drop-in Replacement (PADRe), a novel and unifying framework designed to replace the conventional self-attention mechanism in transformer models. Notably, several recent alternative attention mechanisms, including Hyena, Mamba, SimA, Conv2Former, and Castling-ViT, can be viewed as specific instances of our PADRe framework. PADRe leverages polynomial functions and draws upon established results from approximation theory, enhancing computational efficiency without compromising accuracy. PADRe's key components include multiplicative nonlinearities, which we implement using straightforward, hardware-friendly operations such as Hadamard products, incurring only linear computational and memory costs. PADRe further avoids the need for using complex functions such as Softmax, yet it maintains comparable or superior accuracy compared to traditional self-attention. We assess the effectiveness of PADRe as a drop-in replacement for self-attention across diverse computer vision tasks. These tasks include image classification, image-based 2D object detection, and 3D point cloud object detection. Empirical results demonstrate that PADRe runs significantly faster than the conventional self-attention (11x~43x faster on server GPU and mobile NPU) while maintaining similar accuracy when substituting self-attention in the transformer models.

****

Liang Qifan,Yixiang Shan,Haipeng Liu,Zhengbang Zhu,Ting Long,Weinan Zhang,Yuan Tian

Reconstruction-Guided Policy: Enhancing Decision-Making through Agent-Wise State Consistency

An important challenge in multi-agent reinforcement learning is partial observability, where agents cannot access the global state of the environment during execution and can only receive observations within their field of view. To address this issue, previous works typically use the dimensional-wise state, which is obtained by applying MLP or dimensional-based attention on the global state, for decision-making during training and relying on a reconstructed dimensional-wise state during execution. However, dimensional-wise states tend to divert agent attention to specific features, neglecting potential dependencies between agents, making it difficult to make optimal decisions. Moreover, the inconsistency between the states used in training and execution further increases additional errors. To resolve these issues, we propose a method called Reconstruction-Guided Policy (RGP) to reconstruct the agent-wise state, which represents the information of inter-agent relationships, as input for decision-making during both training and execution. This not only preserves the potential dependencies between agents but also ensures consistency between the states used in training and execution. We conducted extensive experiments on both discrete and continuous action environments to evaluate RGP, and the results demonstrates its superior effectiveness. Our code is public in https://anonymous.4open.science/r/RGP-9F79

****
Barbora Barancikova,Zhuoyue Huang,Cristopher Salvi
SigDiffusions: Score-Based Diffusion Models for Time Series via Log-Signature Em
beddings
Score-based diffusion models have recently emerged as state-of-the-art generativ
e
models for a variety of data modalities. Nonetheless, it remains unclear how to
adapt these models to generate long multivariate time series. Viewing a time
series as the discretisation of an underlying continuous process, we introduce
SigDiffusion, a novel diffusion model operating on log-signature embeddings
of the data. The forward and backward processes gradually perturb and denoise
log-signatures while preserving their algebraic structure. To recover a signal f
rom
its log-signature, we provide new closed-form inversion formulae expressing the
coefficients obtained by expanding the signal in a given basis (e.g. Fourier or
orthogonal polynomials) as explicit polynomial functions of the log-signature.
Finally, we show that combining SigDiffusions with these inversion formulae
results in high-quality long time series generation, competitive with the curren
t
state-of-the-art on various datasets of synthetic and real-world examples.
****
Thomas Pethick,Ioannis Mavrothalassitis,Volkan Cevher
Efficient Interpolation between Extragradient and Proximal Methods for Weak MVIs
We study nonmonotone games satisfying the weak Minty variational inequality (MVI
) with parameter $\rho \in (-\tfrac{1}{L}, \infty)$, where L is the Lipschitz
constant of the gradient operator. An error corrected version of the inexact pro
ximal point algorithm is proposed, with which we establish the first $\mathcal O
(1/\epsilon)$ rate for the entire range $\rho \in (-\tfrac{1}{L}, \infty)$, thus
 removing a logarithmic factor compared with the complexity of existing methods.
 The scheme automatically selects the needed accuracy for the proximal computati
on, and can recover the relaxed extragradient method when $\rho > -\tfrac{1}{2L}
$ and the relaxed proximal point algorithm (rPPA) when $\rho > -\tfrac{1}{L}$. D
ue to the error correction, the scheme inherits the strong properties of the _ex
act_ rPPA. Specifically, we show that linear convergence is automatically achiev
ed under appropriate conditions. Tightness for the range of ρ is establishe
d through a lower bound for rPPA. Central to the algorithmic construction is a h
alfspace projection, where the key insight is that the allowed error tolerance c
an both be used to correct for the proximal approximation and to enlarge the pro
blem class.
****
Zhuoxiao Chen,Junjie Meng,Mahsa Baktashmotlagh,Yonggang Zhang,Zi Huang,Yadan Luo
MOS: Model Synergy for Test-Time Adaptation on LiDAR-Based 3D Object Detection
LiDAR-based 3D object detection is crucial for various applications but often ex
periences performance degradation in real-world deployments due to domain shifts
. While most studies focus on cross-dataset shifts, such as changes in environme
nts and object geometries, practical corruptions from sensor variations and weat
her conditions remain underexplored. In this work, we propose a novel online tes
t-time adaptation framework for 3D detectors that effectively tackles these shif
ts, including a challenging $\textit{cross-corruption}$ scenario where cross-dat
aset shifts and corruptions co-occur. By leveraging long-term knowledge from pre
vious test batches, our approach mitigates catastrophic forgetting and adapts ef
fectively to diverse shifts. Specifically, we propose a Model Synergy (MOS) stra
tegy that dynamically selects historical checkpoints with diverse knowledge and
assembles them to best accommodate the current test batch. This assembly is dire
cted by our proposed Synergy Weights (SW), which perform a weighted averaging of
 the selected checkpoints, minimizing redundancy in the composite model. The SWs
 are computed by evaluating the similarity of predicted bounding boxes on the te
st data and the independence of features between checkpoint pairs in the model b
ank. To maintain an efficient and informative model bank, we discard checkpoints
 with the lowest average SW scores, replacing them with newly updated models. Ou

r method was rigorously tested against existing test-time adaptation strategies across three datasets and eight types of corruptions, demonstrating superior adaptability to dynamic scenes and conditions. Notably, it achieved a 67.3% improvement in a challenging cross-corruption scenario, offering a more comprehensive benchmark for adaptation. Source code: https://github.com/zhuoxiao-chen/MOS.
****

Yitong Deng,Winnie Lin,Lingxiao Li,Dmitriy Smirnov,Ryan D Burgert,Ning Yu,Vincent Dedun,Mohammad H. Taghavi

Infinite-Resolution Integral Noise Warping for Diffusion Models

Adapting pretrained image-based diffusion models to generate temporally consistent videos has become an impactful generative modeling research direction. Training-free noise-space manipulation has proven to be an effective technique, where the challenge is to preserve the Gaussian white noise distribution while adding in temporal consistency. Recently, Chang et al. (2024) formulated this problem using an integral noise representation with distribution-preserving guarantees, and proposed an upsampling-based algorithm to compute it. However, while their mathematical formulation is advantageous, the algorithm incurs a high computational cost. Through analyzing the limiting-case behavior of their algorithm as the upsampling resolution goes to infinity, we develop an alternative algorithm that, by gathering increments of multiple Brownian bridges, achieves their infinite-resolution accuracy while simultaneously reducing the computational cost by orders of magnitude. We prove and experimentally validate our theoretical claims, and demonstrate our method's effectiveness in real-world applications. We further show that our method can readily extend to the 3-dimensional space.
****

Abhishek Panigrahi,Nikunj Saunshi,Kaifeng Lyu,Sobhan Miryoosefi,Sashank J. Reddi,Satyen Kale,Sanjiv Kumar

Efficient stagewise pretraining via progressive subnetworks

Recent developments in large language models have sparked interest in efficient pretraining methods. Stagewise training approaches to improve efficiency, like gradual stacking and layer dropping (Reddi et al., 2023; Zhang & He, 2020), have recently garnered attention. The prevailing view suggests that stagewise dropping strategies, such as layer dropping, are ineffective, especially when compared to stacking-based approaches. This paper challenges this notion by demonstrating that, with proper design, dropping strategies can be competitive, if not better, than stacking methods. Specifically, we develop a principled stagewise training framework, progressive subnetwork training, which only trains subnetworks within the model and progressively increases the size of subnetworks during training, until it trains the full network. We propose an instantiation of this framework – Random Part Training (RAPTR) – that selects and trains only a random subnetwork (e.g. depth-wise, width-wise) of the network at each step, progressively increasing the size in stages. We show that this approach not only generalizes prior works like layer dropping but also fixes their key issues. Furthermore, we establish a theoretical basis for such approaches and provide justification for (a) increasing complexity of subnetworks in stages, conceptually diverging from prior works on layer dropping, and (b) stability in loss across stage transitions in presence of key modern architecture components like residual connections and layer norms. Through comprehensive experiments, we demonstrate that RAPTR can significantly speed up training of standard benchmarks like BERT and UL2, up to 33% compared to standard training and, surprisingly, also shows better downstream performance on UL2, improving QA tasks and SuperGLUE by 1.5%; thereby, providing evidence of better inductive bias.
****

Zhenting Qi,Hanlin Zhang,Eric P. Xing,Sham M. Kakade,Himabindu Lakkaraju

Follow My Instruction and Spill the Beans: Scalable Data Extraction from Retrieval-Augmented Generation Systems
Retrieval-Augmented Generation (RAG) improves pre-trained models by incorporating external knowledge at test time to enable customized adaptation.
We study the risk of datastore leakage in Retrieval-In-Context RAG Language Models (LMs). We show that an adversary can exploit LMs' instruction-following capabilities to easily extract text data verbatim from the datastore of RAG systems built with instruction-tuned LMs via prompt injection.
The vulnerability exists for a wide range of modern LMs that span Llama2, Mistral/Mixtral, Vicuna, SOLAR, WizardLM, Qwen1.5, and Platypus2, and the exploitability exacerbates as the model size scales up.
We also study multiple effects of RAG setup on the extractability of data, indicating that following unexpected instructions to regurgitate data can be an outcome of failure in effectively utilizing contexts for modern LMs, and further show that such vulnerability can be greatly mitigated by position bias elimination strategies.
Extending our study to production RAG models, GPTs, we design an attack that can cause datastore leakage with a near-perfect success rate on 25 randomly selected customized GPTs with at most 2 queries, and we extract text data verbatim at a rate of 41\% from a book of 77,000 words and 3\% from a corpus of 1,569,000 words by prompting the GPTs with only 100 queries generated by themselves.
****
Min Shi,Fuxiao Liu,Shihao Wang,Shijia Liao,Subhashree Radhakrishnan,Yilin Zhao,De-An Huang,Hongxu Yin,Karan Sapra,Yaser Yacoob,Humphrey Shi,Bryan Catanzaro,Andrew Tao,Jan Kautz,Zhiding Yu,Guilin Liu
Eagle: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders
The ability to accurately interpret complex visual information is a crucial topic of multimodal large language models (MLLMs). Recent work indicates that enhanced visual perception significantly reduces hallucinations and improves performance on resolution-sensitive tasks, such as optical character recognition and document analysis. A number of recent MLLMs achieve this goal using a mixture of vision encoders. Despite their success, there is a lack of systematic comparisons and detailed ablation studies addressing critical aspects, such as expert selection and the integration of multiple vision experts. This study provides an extensive exploration of the design space for MLLMs using a mixture of vision encoders and resolutions. Our findings reveal several underlying principles common to various existing strategies, leading to a streamlined yet effective design approach. We discover that simply concatenating visual tokens from a set of complementary vision encoders is as effective as more complex mixing architectures or strategies. We additionally introduce Pre-Alignment to bridge the gap between vision-focused encoders and language tokens, enhancing model coherence. The resulting family of MLLMs, Eagle, surpasses other leading open-source models on major MLLM benchmarks.
****
Kaizhi Zheng,Xiaotong Chen,Xuehai He,Jing Gu,Linjie Li,Zhengyuan Yang,Kevin Lin,Jianfeng Wang,Lijuan Wang,Xin Eric Wang
EditRoom: LLM-parameterized Graph Diffusion for Composable 3D Room Layout Editing
Given the steep learning curve of professional 3D software and the time-consuming process of managing large 3D assets, language-guided 3D scene editing has significant potential in fields such as virtual reality, augmented reality, and
gaming. However, recent approaches to language-guided 3D scene editing either require manual interventions or focus only on appearance modifications without supporting comprehensive scene layout changes. In response, we propose EditRoom, a unified framework capable of executing a variety of layout edits through natural language commands, without requiring manual intervention. Specifically, EditRoom leverages Large Language Models (LLMs) for command planning and generates target scenes using a diffusion-based method, enabling six types of edits: rotate, translate, scale, replace, add, and remove. To address

the lack of data for language-guided 3D scene editing, we have developed an auto
matic pipeline to augment existing 3D scene synthesis datasets and introduced
EditRoom-DB, a large-scale dataset with 83k editing pairs, for training and eval
uation. Our experiments demonstrate that our approach consistently outperforms
other baselines across all metrics, indicating higher accuracy and coherence in
language-guided scene layout editing.
****

Jianheng Tang,Qifan Zhang,Yuhan Li,Nuo Chen,Jia Li
GraphArena: Evaluating and Exploring Large Language Models on Graph Computation
The ``arms race'' of Large Language Models (LLMs) demands new benchmarks to exam
ine their progresses. In this paper, we introduce GraphArena, a benchmarking too
l designed to evaluate LLMs on real-world graph computational problems. It offer
s a suite of four polynomial-time tasks (e.g., Shortest Distance) and six NP-com
plete challenges (e.g., Traveling Salesman Problem). GraphArena features a rigor
ous evaluation framework that classifies LLM outputs as correct, suboptimal (fea
sible but not optimal), hallucinatory (properly formatted but infeasible), or mi
ssing. Evaluation of over 10 LLMs reveals that even top-performing LLMs struggle
 with larger, more complex graph problems and exhibit hallucination issues. We f
urther explore four potential solutions to address this issue and improve LLMs o
n graph computation, including chain-of-thought prompting, instruction tuning, c
ode writing, and scaling test-time compute, each demonstrating unique strengths
and limitations. GraphArena complements the existing LLM benchmarks and is open-
sourced at https://github.com/squareRoot3/GraphArena.
****

Maxence Faldor,Jenny Zhang,Antoine Cully,Jeff Clune
OMNI-EPIC: Open-endedness via Models of human Notions of Interestingness with En
vironments Programmed in Code
Open-ended and AI-generating algorithms aim to continuously generate and solve i
ncreasingly complex tasks indefinitely, offering a promising path toward more ge
neral intelligence. To accomplish this grand vision, learning must occur within
a vast array of potential tasks. Existing approaches to automatically generating
 environments are constrained within manually predefined, often narrow distribut
ions of environments, limiting their ability to create any learning environment.
 To address this limitation, we introduce a novel framework, OMNI-EPIC, that aug
ments previous work in Open-endedness via Models of human Notions of Interesting
ness (OMNI) with Environments Programmed in Code (EPIC). OMNI-EPIC leverages fou
ndation models to autonomously generate code specifying the next learnable (i.e.
, not too easy or difficult for the agent’s current skill set) and interesting (
e.g., worthwhile and novel) tasks. OMNI-EPIC generates both environments (e.g.,
an obstacle course) and reward functions (e.g., progress through the obstacle co
urse quickly without touching red objects), enabling it, in principle, to create
 any simulatable learning task. We showcase the explosive creativity of OMNI-EPI
C, which continuously innovates to suggest new, interesting learning challenges.
 We also highlight how OMNI-EPIC can adapt to reinforcement learning agents’ lea
rning progress, generating tasks that are of suitable difficulty. Overall, OMNI-
EPIC has the potential to endlessly create learnable and interesting environment
s, further propelling the development of self-improving AI systems and AI-Genera
ting Algorithms.
****

Hang Chen,Chufeng Tang,Xiao Li,Xiaolin Hu
Efficient Neuron Segmentation in Electron Microscopy by Affinity-Guided Queries
Accurate segmentation of neurons in electron microscopy (EM) images plays a cruc
ial role in understanding the intricate wiring patterns of the brain. Existing a
utomatic neuron segmentation methods rely on traditional clustering algorithms,
where affinities are predicted first, and then watershed and post-processing alg
orithms are applied to yield segmentation results. Due to the nature of watershe
d algorithm, this paradigm has deficiency in both prediction quality and speed.
Inspired by recent advances in natural image segmentation, we propose to use que
ry-based methods to address the problem because they do not necessitate watershe
d algorithms. However, we find that directly applying existing query-based metho

ds faces great challenges due to the large memory requirement of the 3D data and considerably different morphology of neurons. To tackle these challenges, we introduce affinity-guided queries and integrate them into a lightweight query-based framework. Specifically, we first predict affinities with a lightweight branch, which provides coarse neuron structure information. The affinities are then used to construct affinity-guided queries, facilitating segmentation with bottom-up cues. These queries, along with additional learnable queries, interact with the image features to directly predict the final segmentation results. Experiments on benchmark datasets demonstrated that our method achieved better results over state-of-the-art methods with a 2$\sim$3$\times$ speedup in inference. Code is available at https://github.com/chenhang98/AGQ.
****
Vadim Popov,Assel Yermekova,Tasnima Sadekova,Artem Khrapov,Mikhail Sergeevich Kudinov
Improved Sampling Algorithms for Lévy-Itô Diffusion Models
Lévy-Itô denoising diffusion models relying on isotropic α-stable noise instead of Gaussian distribution have recently been shown to improve performance of conventional diffusion models in image generation on imbalanced datasets while performing comparably in the standard settings. However, the stochastic algorithm of sampling from such models consists in solving the stochastic differential equation describing only an approximate inverse of the process of adding α-stable noise to data which may lead to suboptimal performance. In this paper, we derive a parametric family of stochastic differential equations whose solutions have the same marginal densities as those of the forward diffusion and show that the appropriate choice of the parameter values can improve quality of the generated images when the number of reverse diffusion steps is small. Also, we demonstrate that Lévy-Itô diffusion models are applicable to diverse domains and show that a well-trained text-to-speech Lévy-Itô model may have advantages over standard diffusion models on highly imbalanced datasets.
****
Patrick Yin,Tyler Westenbroek,Ching-An Cheng,Andrey Kolobov,Abhishek Gupta
Rapidly Adapting Policies to the Real-World via Simulation-Guided Fine-Tuning
Robot learning requires a considerable amount of high-quality data to realize the promise of generalization. However, large data sets are costly to collect in the real world. Physics simulators can cheaply generate vast data sets with broad coverage over states, actions, and environments. However, physics engines are fundamentally misspecified approximations to reality. This makes direct zero-shot transfer from simulation to reality challenging, especially in tasks where precise and force-sensitive manipulation is necessary. Thus, fine-tuning these policies with small real-world data sets is an appealing pathway for scaling robot learning. However, current reinforcement learning fine-tuning frameworks leverage general, unstructured exploration strategies which are too inefficient to make real-world adaptation practical. This paper introduces the \emph{Simulation-Guided Fine-tuning} (SGFT) framework, which demonstrates how to extract structural priors from physics simulators to substantially accelerate real-world adaptation. Specifically, our approach uses a value function learned in simulation to guide real-world exploration. We demonstrate this approach across five real-world dexterous manipulation tasks where zero-shot sim-to-real transfer fails. We further demonstrate our framework substantially outperforms baseline fine-tuning methods, requiring up to an order of magnitude fewer real-world samples and succeeding at difficult tasks where prior approaches fail entirely. Last but not least, we provide theoretical justification for this new paradigm which underpins how SGFT can rapidly learn high-performance policies in the face of large sim-to-real dynamics gaps.
****
Siyi Jiao,Wenzheng Zeng,Yerong Li,Huayu Zhang,Changxin Gao,Nong Sang,Mike Zheng Shou
MP-Mat: A 3D-and-Instance-Aware Human Matting and Editing Framework with Multiplane Representation
Human instance matting aims to estimate an alpha matte for each human instance i

n an image, which is challenging as it easily fails in complex cases requiring d isentangling mingled pixels belonging to multiple instances along hairy and thin boundary structures. In this work, we address this by introducing MP-Mat, a nov el 3D-and-instance-aware matting framework with multiplane representation, where the multiplane concept is designed from two different perspectives: scene geome try level and instance level. Specifically, we first build feature-level multipl ane representations to split the scene into multiple planes based on depth diffe rences. This approach makes the scene representation 3D-aware, and can serve as an effective clue for splitting instances in different 3D positions, thereby imp roving interpretability and boundary handling ability especially in occlusion ar eas. Then, we introduce another multiplane representation that splits the scene in an instance-level perspective, and represents each instance with both matte a nd color. We also treat background as a special instance, which is often overloo ked by existing methods. Such an instance-level representation facilitates both foreground and background content awareness, and is useful for other down-stream tasks like image editing. Once built, the representation can be reused to reali ze controllable instance-level image editing with high efficiency. Extensive exp eriments validate the clear advantage of MP-Mat in matting task. We also demonst rate its superiority in image editing tasks, an area under-explored by existing matting-focused methods, where our approach under zero-shot inference even outpe rforms trained specialized image editing techniques by large margins. Code is op en-sourced at https://github.com/JiaoSiyi/MPMat.git.
****

Junyu Huang,Qilong Feng,Ziyun Huang,Zhen Zhang,Jinhui Xu,Jianxin Wang
New Algorithms for the Learning-Augmented k-means Problem
In this paper, we study the clustering problems in the learning-augmented settin g, where predicted labels for a d-dimensional dataset with size m are given by a n oracle to serve as auxiliary information to improve the clustering performance . Following the prior work, the given oracle is parameterized by some error rate α, which captures the accuracy of the oracle such that there are at most α frac tion of false positives and false negatives in each predicted cluster. In this s etting, the goal is to design fast and practical algorithms that can break the c omputational barriers of inapproximability. The current state-of-the-art learnin g-augmented k-means algorithm relies on sorting strategies to find good coordina tes approximation, where a $(1+O(\alpha))$-approximation can be achieved with near-line ar running time in the data size. However, the computational demands for sorting may limit the scalability of the algorithm for handling large-scale datasets. T o address this issue, in this paper, we propose new algorithms that can identify good coordinates approximation using sampling-based strategies, where $(1+O(\alpha))$- approximation can be achieved with linear running time in the data size. To obta in a more practical algorithm for the problem with better clustering quality and running time, we propose a sampling-based heuristic which can directly find cen ter approximations using sampling-based strategies. Empirical experiments show t hat our proposed methods are faster than the state-of-the-art learning-augmented k-means algorithms with comparable performances on clustering quality.
****

Nathan W. Henry,Giovanni Luca Marchetti,Kathlén Kohn
Geometry of Lightning Self-Attention: Identifiability and Dimension
We consider function spaces defined by self-attention networks without normaliza tion, and theoretically analyze their geometry. Since these networks are polynom ial, we rely on tools from algebraic geometry. In particular, we study the ident ifiability of deep attention by providing a description of the generic fibers of the parametrization for an arbitrary number of layers and, as a consequence, co mpute the dimension of the function space. Additionally, for a single-layer mode l, we characterize the singular and boundary points. Finally, we formulate a con jectural extension of our results to normalized self-attention networks, prove i t for a single layer, and numerically verify it in the deep case.
****

Hunter Nisonoff,Junhao Xiong,Stephan Allenspach,Jennifer Listgarten
Unlocking Guidance for Discrete State-Space Diffusion and Flow Models

Generative models on discrete state-spaces have a wide range of potential applic
ations, particularly in the domain of natural sciences. In continuous state-spac
es, controllable and flexible generation of samples with desired properties has
been realized using guidance on diffusion and flow models. However, these guidan
ce approaches are not readily amenable to discrete state-space models. Consequen
tly, we introduce a general and principled method for applying guidance on such
models. Our method depends on leveraging continuous-time Markov processes on dis
crete state-spaces, which unlocks computational tractability for sampling from a
 desired guided distribution. We demonstrate the utility of our approach, Discre
te Guidance, on a range of applications including guided generation of small-mol
ecules, DNA sequences and protein sequences.
****
Huihan Li,Arnav Goel,Keyu He,Xiang Ren
Attributing Culture-Conditioned Generations to Pretraining Corpora
In open-ended generative tasks like narrative writing or dialogue, large languag
e models often exhibit cultural biases, showing limited knowledge and generating
 templated outputs for less prevalent cultures. Recent works show that these bia
ses may stem from uneven cultural representation in pretraining corpora. This wo
rk investigates how pretraining leads to biased culture-conditioned generations
by analyzing how models associate entities with cultures based on pretraining da
ta patterns. We propose the MEMOED framework (MEMOrization from prEtraining Docu
ment) to determine whether a generation for a culture arises from memorization.
Using MEMOED on culture-conditioned generations about food and clothing for 110
cultures, we find that high-frequency cultures in pretraining data yield more ge
nerations with memorized symbols, while some low-frequency cultures produce none
. Additionally, the model favors generating entities with extraordinarily high f
requency regardless of the conditioned culture, reflecting biases toward frequen
t pretraining terms irrespective of relevance. We hope that the MEMOED framework
 and our insights will inspire more works on attributing model performance on pr
etraining data.
****
Tao Zhang,Jinyong Wen,Zhen Chen,Kun Ding,Shiming Xiang,Chunhong Pan
UNIP: Rethinking Pre-trained Attention Patterns for Infrared Semantic Segmentati
on
Pre-training techniques significantly enhance the performance of semantic segmen
tation tasks with limited training data. However, the efficacy under a large dom
ain gap between pre-training (e.g. RGB) and fine-tuning (e.g. infrared) remains
underexplored. In this study, we first benchmark the infrared semantic segmentat
ion performance of various pre-training methods and reveal several phenomena dis
tinct from the RGB domain. Next, our layerwise analysis of pre-trained attention
 maps uncovers that: (1) There are three typical attention patterns (local, hybr
id, and global); (2) Pre-training tasks notably influence pattern distribution a
cross layers; (3) The hybrid pattern is crucial for semantic segmentation as it
attends to both nearby and foreground elements; (4) The texture bias impedes mod
el generalization in infrared tasks. Building on these insights, we propose UNIP
, a UNified Infrared Pre-training framework, to enhance the pre-trained model pe
rformance. This framework uses the hybrid-attention distillation NMI-HAD as the
pre-training target, a large-scale mixed dataset InfMix for pre-training, and a
last-layer feature pyramid network LL-FPN for fine-tuning. Experimental results
show that UNIP outperforms various pre-training methods by up to 13.5% in averag
e mIoU on three infrared segmentation tasks, evaluated using fine-tuning and lin
ear probing metrics. UNIP-S achieves performance on par with MAE-L while requiri
ng only 1/10 of the computational cost. Furthermore, with fewer parameters, UNIP
 significantly surpasses state-of-the-art (SOTA) infrared or RGB segmentation me
thods and demonstrates the broad potential for application in other modalities,
such as RGB and depth. Our code is available at https://github.com/casiatao/UNIP
.
****
Vindula Jayawardana,Baptiste Freydt,Ao Qu,Cameron Hickert,Zhongxia Yan,Cathy Wu
IntersectionZoo: Eco-driving for Benchmarking Multi-Agent Contextual Reinforceme

nt Learning
Despite the popularity of multi-agent reinforcement learning (RL) in simulated a nd two-player applications, its success in messy real-world applications has bee n limited. A key challenge lies in its generalizability across problem variation s, a common necessity for many real-world problems. Contextual reinforcement lea rning (CRL) formalizes learning policies that generalize across problem variatio ns. However, the lack of standardized benchmarks for multi-agent CRL has hindere d progress in the field. Such benchmarks are desired to be based on real-world a pplications to naturally capture the many open challenges of real-world problems that affect generalization. To bridge this gap, we propose IntersectionZoo, a c omprehensive benchmark suite for multi-agent CRL through the real-world applicat ion of cooperative eco-driving in urban road networks. The task of cooperative e co-driving is to control a fleet of vehicles to reduce fleet-level vehicular emi ssions. By grounding IntersectionZoo in a real-world application, we naturally c apture real-world problem characteristics, such as partial observability and mul tiple competing objectives. IntersectionZoo is built on data-informed simulation s of 16,334 signalized intersections derived from 10 major US cities, modeled in an open-source industry-grade microscopic traffic simulator. By modeling factor s affecting vehicular exhaust emissions (e.g., temperature, road conditions, tra vel demand), IntersectionZoo provides one million data-driven traffic scenarios. Using these traffic scenarios, we benchmark popular multi-agent RL and human-li ke driving algorithms and demonstrate that the popular multi-agent RL algorithms struggle to generalize in CRL settings.
****
Jonathan Kahana,Eliahu Horwitz,Imri Shuval,Yedid Hoshen
Deep Linear Probe Generators for Weight Space Learning
Weight space learning aims to extract information about a neural network, such a s its training dataset or generalization error. Recent approaches learn directly from model weights, but this presents many challenges as weights are high-dimen sional and include permutation symmetries between neurons. An alternative approa ch, Probing, represents a model by passing a set of learned inputs (probes) thro ugh the model, and training a predictor on top of the corresponding outputs. Alt hough probing is typically not used as a stand alone approach, our preliminary e xperiment found that a vanilla probing baseline worked surprisingly well. Howeve r, we discover that current probe learning strategies are ineffective. We theref ore propose Deep Linear Probe Generators (ProbeGen), a simple and effective modi fication to probing approaches. ProbeGen adds a shared generator module with a d eep linear architecture, providing an inductive bias towards structured probes t hus reducing overfitting. While simple, ProbeGen performs significantly better t han the state-of-the-art and is very efficient, requiring between 30 to 1000 tim es fewer FLOPs than other top approaches.
****
Dehong Xu,Ruiqi Gao,Wenhao Zhang,Xue-Xin Wei,Ying Nian Wu
On Conformal Isometry of Grid Cells: Learning Distance-Preserving Position Embed ding
This paper investigates the conformal isometry hypothesis as a potential explana tion for the hexagonal periodic patterns in grid cell response maps. We posit th at grid cell activities form a high-dimensional vector in neural space, encoding the agent's position in 2D physical space. As the agent moves, this vector rota tes within a 2D manifold in the neural space, driven by a recurrent neural netwo rk. The conformal hypothesis proposes that this neural manifold is a conformal i sometric embedding of 2D physical space, where local physical distance is preser ved by the embedding up to a scaling factor (or unit of metric). Such distance-p reserving position embedding is indispensable for path planning in navigation, e specially planning local straight path segments. We conduct numerical experiment s to show that this hypothesis leads to the hexagonal grid firing patterns by le arning maximally distance-preserving position embedding, agnostic to the choice of the recurrent neural network. Furthermore, we present a theoretical explanati on of why hexagon periodic patterns emerge by minimizing our loss function by sh owing that hexagon flat torus is maximally distance preserving.

****
Tianyi Bao,Xinyu Ye,Hang Ruan,Chang Liu,Wenjie Wu,Junchi Yan
Beyond Circuit Connections: A Non-Message Passing Graph Transformer Approach for Quantum Error Mitigation
Despite the progress in quantum computing, one major bottleneck against the practical utility is its susceptibility to noise, which frequently occurs in current quantum systems. Existing quantum error mitigation (QEM) methods either lack generality to noise and circuit types or fail to capture the global dependencies of entire systems in addition to circuit structure. In this work, we first propose a unique circuit-to-graph encoding scheme with qubit-wise noisy measurement aggregated. Then, we introduce GTranQEM, a non-message passing graph transformer designed to mitigate errors in expected circuit measurement outcomes effectively. GTranQEM is equipped with a quantum-specific positional encoding, a structure matrix as attention bias guiding nonlocal aggregation, and a virtual quantum-representative node to further grasp graph representations, which guarantees to model the long-range entanglement. Experimental evaluations demonstrate that GTranQEM outperforms state-of-the-art QEM methods on both random and structured quantum circuits across noise types and scales among diverse settings.
****
Anh Tong,Thanh Nguyen-Tang,Dongeun Lee,Duc Nguyen,Toan Tran,David Leo Wright Hall,Cheongwoong Kang,Jaesik Choi
Neural ODE Transformers: Analyzing Internal Dynamics and Adaptive Fine-tuning
Recent advancements in large language models (LLMs) based on transformer architectures have sparked significant interest in understanding their inner workings. In this paper, we introduce a novel approach to modeling transformer architectures using highly flexible non-autonomous neural ordinary differential equations (ODEs). Our proposed model parameterizes all weights of attention and feed-forward blocks through neural networks, expressing these weights as functions of a continuous layer index. Through spectral analysis of the model's dynamics, we uncover an increase in eigenvalue magnitude that challenges the weight-sharing assumption prevalent in existing theoretical studies. We also leverage the Lyapunov exponent to examine token-level sensitivity, enhancing model interpretability. Our neural ODE transformer demonstrates performance comparable to or better than vanilla transformers across various configurations and datasets, while offering flexible fine-tuning capabilities that can adapt to different architectural constraints.
****
BO HU,Tat-Jen Cham
TSC-Net: Prediction of Pedestrian Trajectories by Trajectory-Scene-Cell Classification
To predict future trajectories of pedestrians, scene is as important as the history trajectory since i) scene reflects the position of possible goals of the pedestrian ii) trajectories are affected by the semantic information of the scene. It requires the model to capture scene information and learn the relation between scenes and trajectories. However, existing methods either apply Convolutional Neural Networks (CNNs) to summarize the scene to a feature vector, which raises the feature misalignment issue, or convert trajectory to heatmaps to align with the scene map, which ignores the interactions among different pedestrians. In this work, we introduce the trajectory-scene-cell feature to represent both trajectories and scenes in one feature space. By decoupling the trajectory in temporal domain and the scene in spatial domain, trajectory feature and scene feature are re-organized in different types of cell feature, which well aligns trajectory and scene, and allows the framework to model both human-human and human-scene interactions. Moreover, the Trajectory-Scene-Cell Network (TSC-Net) with new trajectory prediction manner is proposed, where both goal and intermediate positions of the trajectory are predict by cell classification and offset regression. Comparative experiments show that TSC-Net achieves the SOTA performance on several datasets with most of the metrics. Especially for the goal estimation, TSC-Net is demonstrated better on predicting goals for trajectories with irregular speed.
****

Fangyu Lei,Jixuan Chen,Yuxiao Ye,Ruisheng Cao,Dongchan Shin,Hongjin SU,ZHAOQING SUO,Hongcheng Gao,Wenjing Hu,Pengcheng Yin,Victor Zhong,Caiming Xiong,Ruoxi Sun, Qian Liu,Sida Wang,Tao Yu
Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Work flows
Real-world enterprise text-to-SQL workflows often involve complex cloud or local data across various database systems, multiple SQL queries in various dialects, and diverse operations from data transformation to analytics.
We introduce Spider 2.0, an evaluation framework comprising 632 real-world tex t-to-SQL workflow problems derived from enterprise-level database use cases.
The databases in Spider 2.0 are sourced from real data applications, often containing over 1,000 columns and stored in local or cloud database systems such as B igQuery and Snowflake.
We show that solving problems in Spider 2.0 frequently requires understanding an d searching through database metadata, dialect documentation, and even project-l evel codebases.
This challenge calls for models to interact with complex SQL workflow environmen ts, process extremely long contexts, perform intricate reasoning, and generate m ultiple SQL queries with diverse operations, often exceeding 100 lines, which goes far beyond traditional text-to-SQL challenges.
Our evaluations indicate that based on o1-preview, our code agent framework succ essfully solves only 21.3\% of the tasks, compared with 91.2\% on Spider 1.0 and 73.0\% on BIRD.
Our results on Spider 2.0 show that while language models have demonstrated rema rkable performance in code generation --- especially in prior text-to-SQL benchm arks --- they require significant improvement in order to achieve adequate perfo rmance for real-world enterprise usage.
Progress on Spider 2.0 represents crucial steps towards developing intelligent, autonomous, code agents for real-world enterprise settings.
Our code, baseline models, and data are available at [spider2-sql.github.io](spi der2-sql.github.io) .
****
Timofei Gritsaev,Nikita Morozov,Sergey Samsonov,Daniil Tiapkin
Optimizing Backward Policies in GFlowNets via Trajectory Likelihood Maximization
Generative Flow Networks (GFlowNets) are a family of generative models that lear n to sample objects with probabilities proportional to a given reward function.
The key concept behind GFlowNets is the use of two stochastic policies: a forwar d policy, which incrementally constructs compositional objects, and a backward p olicy, which sequentially deconstructs them. Recent results show a close relatio nship between GFlowNet training and entropy-regularized reinforcement learning (RL) problems with a particular reward design. However, this connection applies o nly in the setting of a fixed backward policy, which might be a significant limi tation. As a remedy to this problem, we introduce a simple backward policy optim ization algorithm that involves direct maximization of the value function in an entropy-regularized Markov Decision Process (MDP) over intermediate rewards. We provide an extensive experimental evaluation of the proposed approach across var ious benchmarks in combination with both RL and GFlowNet algorithms and demonstr ate its faster convergence and mode discovery in complex environments.
****
Core Francisco Park,Ekdeep Singh Lubana,Hidenori Tanaka
Competition Dynamics Shape Algorithmic Phases of In-Context Learning
In-Context Learning (ICL) has significantly expanded the general-purpose nature of large language models, allowing them to adapt to novel tasks using merely the inputted context. This has motivated a series of papers that analyze tractable synthetic domains and postulate precise mechanisms that may underlie ICL. Howeve r, the use of relatively distinct setups that often lack a sequence modeling nat ure to them makes it unclear how general the reported insights from such studies are. Motivated by this, we propose a synthetic sequence modeling task that invo lves learning to simulate a finite mixture of Markov chains. As we show, models trained on this task reproduce most well-known results on ICL, hence offering a

unified setting for studying the concept. Building on this setup, we demonstrate we can explain a model’s behavior by decomposing it into four broad algorithms that combine a fuzzy retrieval vs. inference approach with either unigram or bigram statistics of the context. These algorithms engage in a competitive dynamics to dominate model behavior, with the precise experimental conditions dictating which algorithm ends up superseding others: e.g., we find merely varying context size or amount of training yields (at times sharp) transitions between which algorithm dictates the model behavior, revealing a mechanism that explains the transient nature of ICL. In this sense, we argue ICL is best thought of as a mixture of different algorithms, each with its own peculiarities, instead of a monolithic capability. This also implies that making general claims about ICL that hold universally across all settings may be infeasible.
****
Haocheng Xi,Han Cai,Ligeng Zhu,Yao Lu,Kurt Keutzer,Jianfei Chen,Song Han
COAT: Compressing Optimizer states and Activations for Memory-Efficient FP8 Training
FP8 training has emerged as a promising method for improving training efficiency. Existing frameworks accelerate training by applying FP8 computation to linear layers while leaving optimizer states and activations in higher precision, which fails to fully optimize memory usage. This paper introduces COAT (**C**ompressing **O**ptimizer States and **A**ctivations for FP8 **T**raining), a novel FP8 training framework designed to significantly reduce memory footprint when training large models. COAT addresses current limitations through two key innovations: (1) **Dynamic Range Expansion**, which aligns optimizer state distributions more closely with the FP8 representation range, thereby reducing quantization error, and (2) **Mixed-Granularity Activation Quantization**, which optimizes activation memory using a combination of per-tensor and per-group quantization strategies. Experiments demonstrate that COAT effectively reduces end-to-end training memory footprint by **1.54×** compared to BF16 while achieving nearly lossless performance across various tasks, such as Large Language Model pretraining and fine-tuning and Vision Language Model training. COAT also achieves a **1.43×** end-to-end training speedup compared to BF16, performing on par with or surpassing TransformerEngine's speedup. COAT enables efficient full-parameter training of large models on fewer GPUs, and facilitates doubling the batch size in distributed training settings, providing a practical solution for scaling large-scale model training. Code will be released upon publication.
****
Wonhyeok Choi,Kyumin Hwang,Wei Peng,Minwoo Choi,Sunghoon Im
Self-supervised Monocular Depth Estimation Robust to Reflective Surface Leveraged by Triplet Mining
Self-supervised monocular depth estimation (SSMDE) aims to predict the dense depth map of a monocular image, by learning depth from RGB image sequences, eliminating the need for ground-truth depth labels.
Although this approach simplifies data acquisition compared to supervised methods, it struggles with reflective surfaces, as they violate the assumptions of Lambertian reflectance, leading to inaccurate training on such surfaces.
To tackle this problem, we propose a novel training strategy for an SSMDE by leveraging triplet mining to pinpoint reflective regions at the pixel level, guided by the camera geometry between different viewpoints.
The proposed reflection-aware triplet mining loss specifically penalizes the inappropriate photometric error minimization on the localized reflective regions while preserving depth accuracy on non-reflective areas.
We also incorporate a reflection-aware knowledge distillation method that enables a student model to selectively learn the pixel-level knowledge from reflective and non-reflective regions. This results in robust depth estimation across areas.
Evaluation results on multiple datasets demonstrate that our method effectively enhances depth quality on reflective surfaces and outperforms state-of-the-art SSMDE baselines.
****

Amith Ananthram,Elias Stengel-Eskin,Mohit Bansal,Kathleen McKeown
See It from My Perspective: How Language Affects Cultural Bias in Image Understanding
Vision-language models (VLMs) can respond to queries about images in many languages. However, beyond language, culture affects how we see things. For example, individuals from Western cultures focus more on the central figure in an image while individuals from East Asian cultures attend more to scene context (Nisbett 2001). In this work, we characterize the Western bias of VLMs in image understanding and investigate the role that language plays in this disparity. We evaluate VLMs across subjective and objective visual tasks with culturally diverse images and annotations. We find that VLMs perform better on the Western split than on the East Asian split of each task. Through controlled experimentation, we trace one source of this bias in image understanding to the lack of diversity in language model construction. While inference in a language nearer to a culture can lead to reductions in bias, we show it is much more effective when that language was well-represented during text-only pre-training. Interestingly, this yields bias reductions even when prompting in English. Our work highlights the importance of richer representation of all languages in building equitable VLMs.
****
Xueyan Zou,Yuchen Song,Ri-Zhao Qiu,Xuanbin Peng,Jianglong Ye,Sifei Liu,Xiaolong Wang
3D-SPATIAL MULTIMODAL MEMORY
We present 3D Spatial MultiModal Memory (M3), a multimodal memory system designed to retain information about medium-sized static scenes through video sources for visual perception. By integrating 3D Gaussian Splatting techniques with foundation models, M3 builds a multimodal memory capable of rendering feature representations across granularities, encompassing a wide range of knowledge. In our exploration, we identify two key challenges in previous works on feature splatting: (1) computational constraints in storing high-dimensional features for each Gaussian primitive, and (2) misalignment or information loss between distilled features and foundation model features. To address these challenges, we propose M3 with key components of principal scene components and Gaussian memory attention, enabling efficient training and inference. To validate M3, we conduct comprehensive quantitative evaluations of feature similarity and downstream tasks, as well as qualitative visualizations to highlight the pixel trace of Gaussian memory attention. Our approach encompasses a diverse range of foundation models, including vision-language models (VLMs), perception models, and large multimodal and language models (LMMs/LLMs). Furthermore, to demonstrate real-world applicability, we deploy M3’s feature field in indoor scenes on a quadruped robot. Notably, we claim that M3 is the first work to address the core compression challenges in 3D feature distillation.
****
Yipei Wang,Xiaoqian Wang
Agree to Disagree: Demystifying Homogeneous Deep Ensembles through Distributional Equivalence
Deep ensembles improve the performance of the models by taking the average predictions of a group of ensemble members. However, the origin of these capabilities remains a mystery and deep ensembles are used as a reliable “black box” to improve the performance. Existing studies typically attribute such improvement to Jensen gaps of the deep ensemble method, where the loss of the mean does not exceed the mean of the loss for any convex loss metric. In this work, we demonstrate that Jensen’s inequality is not responsible for the effectiveness of deep ensembles, and convexity is not a necessary condition. Instead, Jensen Gap focuses on the “average loss” of individual models, which provides no practical meaning. Thus it fails to explain the core phenomena of deep ensembles such as the superiority to any single ensemble member, the decreasing loss with the number of ensemble members, etc. Regarding this mystery, we provide theoretical analysis and comprehensive empirical results from a statistical perspective that reveal the true mechanism of deep ensembles. Our results highlight that deep ensembles originate from the homogeneous output distribution across all ensemble members. Specific

ally, the predictions of homogeneous models (Abe et al., 2022b) have the distrib
utional equivalence property – Although the predictions of independent ensemble
members are point-wise different, they form an identical distribution. Such agre
ement and disagreement contribute to deep ensembles' "magical power". Based on t
his discovery, we provide rigorous proof of the effectiveness of deep ensembles
and analytically quantify the extent to which ensembles improve performance. The
 derivations not only theoretically quantify the effectiveness of deep ensembles
 for the first time, but also enable estimation schemes that foresee the perform
ance of ensembles with different capacities. Furthermore, different from existin
g studies, our results also point out that deep ensembles work in a different me
chanism from model scaling a single model, even though significant correlations
between them have been observed.
****
Siqi Wan,Jingwen Chen,Yingwei Pan,Ting Yao,Tao Mei
Incorporating Visual Correspondence into Diffusion Model for Virtual Try-On
Diffusion models have shown preliminary success in virtual try-on (VTON) task. T
he typical dual-branch architecture comprises two UNets for implicit garment def
ormation and synthesized image generation respectively, and has emerged as the r
ecipe for VTON task. Nevertheless, the problem remains challenging to preserve t
he shape and every detail of the given garment due to the intrinsic stochasticit
y of diffusion model. To alleviate this issue, we novelly propose to explicitly
capitalize on visual correspondence as the prior to tame diffusion process inste
ad of simply feeding the whole garment into UNet as the appearance reference. Sp
ecifically, we interpret the fine-grained appearance and texture details as a se
t of structured semantic points, and match the semantic points rooted in garment
 to the ones over target person through local flow warping. Such 2D points are t
hen augmented into 3D-aware cues with depth/normal map of target person. The cor
respondence mimics the way of putting clothing on human body and the 3D-aware cu
es act as semantic point matching to supervise diffusion model training. A point
-focused diffusion loss is further devised to fully take the advantage of semant
ic point matching. Extensive experiments demonstrate strong garment detail prese
rvation of our approach, evidenced by state-of-the-art VTON performances on both
 VITON-HD and DressCode datasets. Code is publicly available at: https://github.
com/HiDream-ai/SPM-Diff.
****
Dominik Fuchsgruber,Tim Postuvan,Stephan Günnemann,Simon Geisler
Graph Neural Networks for Edge Signals: Orientation Equivariance and Invariance
Many applications in traffic, civil engineering, or electrical engineering revol
ve around edge-level signals. Such signals can be categorized as inherently dire
cted, for example, the water flow in a pipe network, and undirected, like the di
ameter of a pipe. Topological methods model edge signals with inherent direction
 by representing them relative to a so-called *orientation* assigned to each edg
e.
They can neither model undirected edge signals nor distinguish if an edge itself
 is directed or undirected. We address these shortcomings by (i) revising the no
tion of *orientation equivariance* to enable edge direction-aware topological mo
dels, (ii) proposing *orientation invariance* as an additional requirement to de
scribe signals without inherent direction, and (iii) developing EIGN, an archite
cture composed of novel direction-aware edge-level graph shift operators, that p
rovably fulfils the aforementioned desiderata. It is the first work that discuss
es modeling directed and undirected signals while distinguishing between directe
d and undirected edges. A comprehensive evaluation shows that EIGN outperforms p
rior work in edge-level tasks, improving in RMSE on flow simulation tasks by up
to 23.5%.
****
Ruoxuan Feng,Jiangyu Hu,Wenke Xia,TianciGao,Ao Shen,Yuhao Sun,Bin Fang,Di Hu
AnyTouch: Learning Unified Static-Dynamic Representation across Multiple Visuo-t
actile Sensors
Visuo-tactile sensors aim to emulate human tactile perception, enabling robots t
o precisely understand and manipulate objects. Over time, numerous meticulously

designed visuo-tactile sensors have been integrated into robotic systems, aiding in completing various tasks. However, the distinct data characteristics of these low-standardized visuo-tactile sensors hinder the establishment of a powerful tactile perception system. We consider that the key to addressing this issue lies in learning unified multi-sensor representations, thereby integrating the sensors and promoting tactile knowledge transfer between them. To achieve unified representation of this nature, we introduce TacQuad, an aligned multi-modal multi-sensor tactile dataset from four different visuo-tactile sensors, which enables the explicit integration of various sensors. Recognizing that humans perceive the physical environment by acquiring diverse tactile information such as texture and pressure changes, we further propose to learn unified multi-sensor representations from both static and dynamic perspectives. By integrating tactile images and videos, we present AnyTouch, a unified static-dynamic multi-sensor representation learning framework with a multi-level structure, aimed at both enhancing comprehensive perceptual abilities and enabling effective cross-sensor transfer. This multi-level architecture captures pixel-level details from tactile data via masked modeling and enhances perception and transferability by learning semantic-level sensor-agnostic features through multi-modal alignment and cross-sensor matching. We provide a comprehensive analysis of multi-sensor transferability, and validate our method on various offline datasets and in the real-world pouring task. Experimental results show that our method outperforms existing methods, exhibits outstanding static and dynamic perception capabilities across various sensors. The code, TacQuad dataset and AnyTouch model are fully available at gewu-lab.github.io/AnyTouch/.
****

Manuel Cherep,Nikhil Singh
Contrastive Learning from Synthetic Audio Doppelgängers
Learning robust audio representations currently demands extensive datasets of real-world sound recordings. By applying artificial transformations to these recordings, models can learn to recognize similarities despite subtle variations through techniques like contrastive learning. However, these transformations are only approximations of the true diversity found in real-world sounds, which are generated by complex interactions of physical processes, from vocal cord vibrations to the resonance of musical instruments. We propose a solution to both the data scale and transformation limitations, leveraging synthetic audio. By randomly perturbing the parameters of a sound synthesizer, we generate audio doppelgängers—synthetic positive pairs with causally manipulated variations in timbre, pitch, and temporal envelopes. These variations, difficult to achieve through augmentations of existing audio, provide a rich source of contrastive information. Despite the shift to randomly generated synthetic data, our method produces strong representations, outperforming real data on several standard audio classification tasks. Notably, our approach is lightweight, requires no data storage, and has only a single hyperparameter, which we extensively analyze. We offer this method as a complement to existing strategies for contrastive learning in audio, using synthesized sounds to reduce the data burden on practitioners.
****

Jiashuo Liu,Nabeel Seedat,Peng Cui,Mihaela van der Schaar
Going Beyond Static: Understanding Shifts with Time-Series Attribution
Distribution shifts in time-series data are complex due to temporal dependencies, multivariable interactions, and trend changes.
However, robust methods often rely on structural assumptions that lack thorough empirical validation, limiting their practical applicability.
In order to support an empirically grounded inductive approach to research, we introduce our **T**ime-**S**eries **S**hift **A**ttribution (TSSA) framework, which analyzes *problem-specific* patterns of distribution shifts. Our framework attributes performance degradation from various types of shifts to each *temporal data property* in a detailed manner, supported by theoretical analysis of unbiasedness and asymptotic properties. Empirical studies in real-world healthcare applications highlight how the TSSA framework enhances the understanding of time-series shifts, facilitating reliable model deployment and driving targeted improv

ements from both algorithmic and data-centric perspectives.
****
Kyeongmin Yeo,Jaihoon Kim,Minhyuk Sung
StochSync: Stochastic Diffusion Synchronization for Image Generation in Arbitrary Spaces
We propose a zero-shot method for generating images in arbitrary spaces (e.g., a sphere for 360■ panoramas and a mesh surface for texture) using a pretrained image diffusion model. The zero-shot generation of various visual content using a pretrained image diffusion model has been explored mainly in two directions. First, Diffusion Synchronization–performing reverse diffusion processes jointly across different projected spaces while synchronizing them in the target space–generates high-quality outputs when enough conditioning is provided, but it struggles in its absence. Second, Score Distillation Sampling–gradually updating the target space data through gradient descent–results in better coherence but often lacks detail. In this paper, we reveal for the first time the interconnection between these two methods while highlighting their differences. To this end, we propose StochSync, a novel approach that combines the strengths of both, enabling effective performance with weak conditioning. Our experiments demonstrate that StochSync provides the best performance in 360■ panorama generation (where image conditioning is not given), outperforming previous finetuning-based methods, and also delivers comparable results in 3D mesh texturing (where depth conditioning is provided) with previous methods.
****
Kaifeng Zhao,Gen Li,Siyu Tang
DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control
Text-conditioned human motion generation, which allows for user interaction through natural language, has become increasingly popular. Existing methods typically generate short, isolated motions based on a single input sentence. However, human motions are continuous and can extend over long periods, carrying rich semantics. Creating long, complex motions that precisely respond to streams of text descriptions, particularly in an online and real-time setting, remains a significant challenge. Furthermore, incorporating spatial constraints into text-conditioned motion generation presents additional challenges, as it requires aligning the motion semantics specified by text descriptions with geometric information, such as goal locations and 3D scene geometry. To address these limitations, we propose **DartC**ontrol, in short **DART**, a **D**iffusion-based **A**utoregressive motion primitive model for **R**eal-time **T**ext-driven motion **C**ontrol. Our model, DART, effectively learns a compact motion primitive space jointly conditioned on motion history and text inputs using latent diffusion models. By autoregressively generating motion primitives based on the preceding history and current text input, DART enables real-time, sequential motion generation driven by natural language descriptions. Additionally, the learned motion primitive space allows for precise spatial motion control, which we formulate either as a latent noise optimization problem or as a Markov decision process addressed through reinforcement learning. We present effective algorithms for both approaches, demonstrating our model’s versatility and superior performance in various motion synthesis tasks. Experiments show our method outperforms existing baselines in motion realism, efficiency, and controllability. Video results and code are available at https://zkf1997.github.io/DART/.
****
Xinyue Wang,Biwei Huang
Modeling Unseen Environments with Language-guided Composable Causal Components in Reinforcement Learning
Generalization in reinforcement learning (RL) remains a significant challenge, especially when agents encounter novel environments with unseen dynamics. Drawing inspiration from human compositional reasoning—where known components are reconfigured to handle new situations—we introduce World Modeling with Compositional Causal Components (WM3C). This novel framework enhances RL generalization by learning and leveraging compositional causal components. Unlike previous approaches

focusing on invariant representation learning or meta-learning, WM3C identifies and utilizes causal dynamics among composable elements, facilitating robust adaptation to new tasks. Our approach integrates language as a compositional modality to decompose the latent space into meaningful components and provides theoretical guarantees for their unique identification under mild assumptions. Our practical implementation uses a masked autoencoder with mutual information constraints and adaptive sparsity regularization to capture high-level semantic information and effectively disentangle transition dynamics. Experiments on numerical simulations and real-world robotic manipulation tasks demonstrate that WM3C significantly outperforms existing methods in identifying latent processes, improving policy learning, and generalizing to unseen tasks.
****
Yixiang Shan,Zhengbang Zhu,Ting Long,Liang Qifan,Yi Chang,Weinan Zhang,Liang Yin
ContraDiff: Planning Towards High Return States via Contrastive Learning
The performance of offline reinforcement learning (RL) is sensitive to the proportion of high-return trajectories in the offline dataset. However, in many simulation environments and real-world scenarios, there are large ratios of low-return trajectories rather than high-return trajectories, which makes learning an efficient policy challenging. In this paper, we propose a method called Contrastive Diffuser (ContraDiff) to make full use of low-return trajectories and improve the performance of offline RL algorithms. Specifically, ContraDiff groups the states of trajectories in the offline dataset into high-return states and low-return states and treats them as positive and negative samples correspondingly. Then, it designs a contrastive mechanism to pull the planned trajectory of an agent toward high-return states and push them away from low-return states. Through the contrast mechanism, trajectories with low returns can serve as negative examples for policy learning, guiding the agent to avoid areas associated with low returns and achieve better performance. Through the contrast mechanism, trajectories with low returns provide a ``counteracting force'' guides the agent to avoid areas associated with low returns and achieve better performance.
Experiments on 27 sub-optimal datasets demonstrate the effectiveness of our proposed method. Our code is publicly available at https://github.com/Looomo/contradiff.
****
Renxi Wang,Xudong Han,Lei Ji,Shu Wang,Timothy Baldwin,Haonan Li
ToolGen: Unified Tool Retrieval and Calling via Generation
As large language models (LLMs) advance, their inability to autonomously execute tasks by directly interacting with external tools remains a critical limitation. Traditional methods rely on inputting tool descriptions as context, which is constrained by context length and requires separate, often inefficient, retrieval mechanisms. We introduce ToolGen, a paradigm shift that integrates tool knowledge directly into the LLM’s parameters by representing each tool as a unique token. This enables the LLM to generate tool calls and arguments as part of its next token prediction capabilities, seamlessly blending tool invocation with language generation. Our framework allows the LLM to access and utilize a vast amount of tools with no additional retrieval step, significantly enhancing both performance and scalability. Experimental results with over 47,000 tools show that ToolGen not only achieves superior results in both tool retrieval and autonomous task completion but also sets the stage for a new era of AI agents that can adapt to tools across diverse domains. By fundamentally transforming tool retrieval into a generative process, ToolGen paves the way for more versatile, efficient, and autonomous AI systems. ToolGen enables end-to-end tool learning and opens opportunities for integration with other advanced techniques such as chain-of-thought and reinforcement learning, thereby expanding the practical capabilities of LLMs
****
Norah Alballa,Wenxuan Zhang,Ziquan Liu,Ahmed M. Abdelmoniem,Mohamed Elhoseiny,Marco Canini
Query-based Knowledge Transfer for Heterogeneous Learning Environments
Decentralized collaborative learning under data heterogeneity and privacy constr

aints has rapidly advanced. However, existing solutions like federated learning, ensembles, and transfer learning, often fail to adequately serve the unique n eeds of clients, especially when local data representation is limited.
To address this issue, we propose a novel framework called Query-based Knowledge Transfer (QKT) that enables tailored knowledge acquisition to fulfill specific client needs without direct data exchange.
It employs a data-free masking strategy to facilitate the communication-efficien t query-focused knowledge transformation while refining task-specific parameters to mitigate knowledge interference and forgetting. Our experiments, conducted o n both standard and clinical benchmarks, show that QKT significantly outperforms existing collaborative learning methods by an average of 20.91% points in singl e-class query settings and an average of 14.32% points in multi-class query scen arios.
Further analysis and ablation studies reveal that QKT effectively balances the l earning of new and existing knowledge, showing strong potential for its applicat ion in decentralized learning.
****
Rachel Mikulinsky,Morris Alper,Shai Gordin,Enrique Jiménez,Yoram Cohen,Hadar Ave rbuch-Elor
ProtoSnap: Prototype Alignment For Cuneiform Signs
The cuneiform writing system served as the medium for transmitting knowledge in the ancient Near East for a period of over three thousand years. Cuneiform signs have a complex internal structure which is the subject of expert paleograp hic analysis, as variations in sign shapes bear witness to historical developments a nd transmission of writing and culture over time. However, prior automated techniqu es mostly treat sign types as categorical and do not explicitly model their highly varied internal configurations. In this work, we present an unsupervised approach for recovering the fine-grained internal configuration of cuneiform signs by leverag ing powerful generative models and the appearance and structure of prototype font images as priors. Our approach, ProtoSnap, enforces structural consistency on matches found with deep image features to estimate the diverse configurations of cuneiform characters, snapping a skeleton-based template to photographed cuneiform signs. We provide a new benchmark of expert annotations and evaluate our method on this task. Our evaluation shows that our approach succeeds in aligning prototype skeletons to a wide variety of cuneiform signs. Moreover, we show that conditioning on structures produced by our method allows for generatin g synthetic data with correct structural configurations, significantly boosting th e performance of cuneiform sign recognition beyond existing techniques, in particu lar over rare signs. Our code, data, and trained models are available at the project page:
https://tau-vailab.github.io/ProtoSnap/
****
Xiuwei Xu,Huangxing Chen,Linqing Zhao,Ziwei Wang,Jie Zhou,Jiwen Lu
EmbodiedSAM: Online Segment Any 3D Thing in Real Time
Embodied tasks require the agent to fully understand 3D scenes simultaneously wi th its exploration, so an online, real-time, fine-grained and highly-generalized 3D perception model is desperately needed. Since high-quality 3D data is limite d, directly training such a model in 3D is infeasible. Meanwhile, vision foundat ion models (VFM) has revolutionized the field of 2D computer vision with superio r performance, which makes the use of VFM to assist embodied 3D perception a pro mising direction. However, most existing VFM-assisted 3D perception methods are either offline or too slow that cannot be applied in practical embodied tasks. I

n this paper, we aim to leverage Segment Anything Model (SAM) for real-time 3D i
nstance segmentation in an online setting. This is a challenging problem since f
uture frames are not available in the input streaming RGB-D video, and an instan
ce may be observed in several frames so efficient object matching between frames
 is required. To address these challenges, we first propose a geometric-aware qu
ery lifting module to represent the 2D masks generated by SAM by 3D-aware querie
s, which is then iteratively refined by a dual-level query decoder. In this way,
 the 2D masks are transferred to fine-grained shapes on 3D point clouds. Benefit
 from the query representation for 3D masks, we can compute the similarity matri
x between the 3D masks from different views by efficient matrix operation, which
 enables real-time inference. Experiments on ScanNet, ScanNet200, SceneNN and 3R
Scan show our method achieves state-of-the-art performance among online 3D perce
ption models, even outperforming offline VFM-assisted 3D instance segmentation m
ethods by a large margin. Our method also demonstrates great generalization abil
ity in several zero-shot dataset transferring experiments and show great potenti
al in data-efficient setting.
****

Zeping Min,Xinshang Wang
DOCS: Quantifying Weight Similarity for Deeper Insights into Large Language Mode
ls
We introduce a novel index, the Distribution of Cosine Similarity (DOCS), for qu
antitatively assessing the similarity between weight matrices in Large Language
Models (LLMs), aiming to facilitate the analysis of their complex architectures
. Leveraging DOCS, our analysis uncovers intriguing patterns in the latest open-
source LLMs: adjacent layers frequently exhibit high weight similarity and tend
to form clusters, suggesting depth-wise functional specialization. Additionally,
 we prove that DOCS is theoretically effective in quantifying similarity for ort
hogonal matrices, a crucial aspect given the prevalence of orthogonal initializa
tions in LLMs. This research contributes to a deeper understanding of LLM archit
ecture and behavior, offering tools with potential implications for developing m
ore efficient and interpretable models.
****

Donglei Yu,Yang Zhao,Jie Zhu,Yangyifan Xu,Yu Zhou,Chengqing Zong
SimulPL: Aligning Human Preferences in Simultaneous Machine Translation
Simultaneous Machine Translation (SiMT) generates translations while receiving s
treaming source inputs. This requires the SiMT model to learn a read/write polic
y, deciding when to translate and when to wait for more source input. Numerous l
inguistic studies indicate that audiences in SiMT scenarios have distinct prefer
ences, such as accurate translations, simpler syntax, and no unnecessary latency
. Aligning SiMT models with these human preferences is crucial to improve their
performances. However, this issue still remains unexplored. Additionally, prefer
ence optimization for SiMT task is also challenging. Existing methods focus sole
ly on optimizing the generated responses, ignoring human preferences related to
latency and the optimization of read/write policy during the preference optimiza
tion phase. To address these challenges, we propose Simultaneous Preference Lear
ning (SimulPL), a preference learning framework tailored for the SiMT task. In t
he SimulPL framework, we categorize SiMT human preferences into five aspects: **
translation quality preference**, **monotonicity preference**, **key point prefe
rence**, **simplicity preference**, and **latency preference**. By leveraging th
e first four preferences, we construct human preference prompts to efficiently g
uide GPT-4/4o in generating preference data for the SiMT task. In the preference
 optimization phase, SimulPL integrates **latency preference** into the optimiza
tion objective and enables SiMT models to improve the read/write policy, thereby
 aligning with human preferences more effectively. Experimental results indicate
 that SimulPL exhibits better alignment with human preferences across all latenc
y levels in Zh$\rightarrow$En, De$\rightarrow$En and En$\rightarrow$Zh SiMT task
s. Our data and code will be available at https://github.com/EurekaForNLP/Simul
PL.
****

Tim Lawson,Lucy Farnik,Conor Houghton,Laurence Aitchison

Residual Stream Analysis with Multi-Layer SAEs
Sparse autoencoders (SAEs) are a promising approach to interpreting the internal representations of transformer language models. However, SAEs are usually trained separately on each transformer layer, making it difficult to use them to study how information flows across layers. To solve this problem, we introduce the multi-layer SAE (MLSAE): a single SAE trained on the residual stream activation vectors from every transformer layer. Given that the residual stream is understood to preserve information across layers, we expected MLSAE latents to 'switch on' at a token position and remain active at later layers. Interestingly, we find that individual latents are often active at a single layer for a given token or prompt, but the layer at which an individual latent is active may differ for different tokens or prompts. We quantify these phenomena by defining a distribution over layers and considering its variance. We find that the variance of the distributions of latent activations over layers is about two orders of magnitude greater when aggregating over tokens compared with a single token. For larger underlying models, the degree to which latents are active at multiple layers increases, which is consistent with the fact that the residual stream activation vectors at adjacent layers become more similar. Finally, we relax the assumption that the residual stream basis is the same at every layer by applying pre-trained tuned-lens transformations, but our findings remain qualitatively similar. Our results represent a new approach to understanding how representations change as they flow through transformers. We release our code to train and analyze MLSAEs at https://github.com/tim-lawson/mlsae.
****
Zirui Hu,Yingjie Wang,Zheng Zhang,Hong Chen,Dacheng Tao
A Statistical Approach for Controlled Training Data Detection
Detecting training data for large language models (LLMs) is receiving growing attention, especially in applications requiring high reliability. While numerous efforts have been made to address this issue, they typically focus on accuracy without ensuring controllable results.
To fill this gap, we propose **K**nockoff Inference-based **T**raining data **D**etector (KTD), a novel method that achieves rigorous false discovery rate (FDR) control in training data detection. Specifically, KTD generates synthetic knockoff samples that seamlessly replace original data points without compromising contextual integrity. A novel knockoff statistic, which incorporates multiple knockoff draws, is then calculated to ensure FDR control while maintaining high power.
Our theoretical analysis demonstrates KTD's asymptotic optimality in terms of FDR control and power. Empirical experiments on real-world datasets such as WikiMIA, XSum and Real Time BBC News further validate KTD's superior performance compared to existing methods.
****
Zonglin Yang,Wanhao Liu,Ben Gao,Tong Xie,Yuqiang Li,Wanli Ouyang,Soujanya Poria,Erik Cambria,Dongzhan Zhou
MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses
Scientific discovery contributes largely to the prosperity of human society, and recent progress shows that LLMs could potentially catalyst the process. However, it is still unclear whether LLMs can discover novel and valid hypotheses in chemistry. In this work, we investigate this main research question: whether LLMs can automatically discover novel and valid chemistry research hypotheses, given only a research question? With extensive discussions with chemistry experts, we adopt the assumption that a majority of chemistry hypotheses can be resulted from a research background question and several inspirations. With this key insight, we break the main question into three smaller fundamental questions. In brief, they are: (1) given a background question, whether LLMs can retrieve good inspirations; (2) with background and inspirations, whether LLMs can lead to hypothesis; and (3) whether LLMs can identify good hypotheses to rank them higher. To investigate these questions, we construct a benchmark consisting of 51 chemistry papers published in Nature or a similar level in 2024 (all papers are only availa

ble online since 2024). Every paper is divided by chemistry PhD students into th
ree components: background, inspirations, and hypothesis. The goal is to redisco
ver the hypothesis given only the background and a large chemistry literature co
rpus consisting the ground truth inspiration papers, with LLMs trained with data
 up to 2023. We also develop an LLM-based multi-agent framework that leverages t
he assumption, consisting of three stages reflecting the more smaller questions.
 The proposed method can rediscover many hypotheses with very high similarity wi
th the ground truth ones, covering the main innovations.
****

Hengxiang Zhang,Songxin Zhang,Bingyi Jing,Hongxin Wei
Fine-tuning can Help Detect Pretraining Data from Large Language Models
In the era of large language models (LLMs), detecting pretraining data has been
increasingly important due to concerns about fair evaluation and ethical risks.
Current methods differentiate members and non-members by designing scoring funct
ions, like Perplexity and Min-k%. However, the diversity and complexity of train
ing data magnifies the difficulty of distinguishing, leading to suboptimal perfo
rmance in detecting pretraining data. In this paper, we first explore the benefi
ts of unseen data, which can be easily collected after the release of the LLM. W
e find that the perplexities of LLMs shift differently for members and non-membe
rs, after fine-tuning with a small amount of previously unseen data. In light of
 this, we introduce a novel and effective method termed Fine-tuned Score Deviati
on (FSD), which improves the performance of current scoring functions for pretra
ining data detection. In particular, we propose to measure the deviation distanc
e of current scores after fine-tuning on a small amount of unseen data within th
e same domain. In effect, using a few unseen data can largely decrease the score
s of all non-members, leading to a larger deviation distance than members. Exten
sive experiments demonstrate the effectiveness of our method, significantly impr
oving the AUC score on common benchmark datasets across various models.
****

Bin Liang,Shiwei Chen,Lin Gui,Hui Wang,Yue Yu,Ruifeng Xu,Kam-Fai Wong
Centrality-guided Pre-training for Graph
Self-supervised learning (SSL) has shown great potential in learning generalizab
le representations for graph-structured data. However, existing SSL-based graph
pre-training methods largely focus on improving graph representations by learnin
g the structure information based on disturbing or reconstructing graphs, which
ignores an important issue: the importance of different nodes in the graph struc
ture may vary. To fill this gap, we propose a Centrality-guided Graph Pre-traini
ng (CenPre) framework to integrate the distinct importance of nodes in graph str
ucture into the corresponding representations of nodes based on the centrality i
n graph theory. In this way, the different roles played by different nodes can b
e effectively leveraged when learning graph structure. The proposed CenPre conta
ins three modules for node representation pre-training and alignment. The node-l
evel importance learning module fuses the fine-grained node importance into node
 representation based on degree centrality, allowing the aggregation of node rep
resentations with equal/similar importance. The graph-level importance learning
module characterizes the importance between all nodes in the graph based on eige
nvector centrality, enabling the exploitation of graph-level structure similarit
ies/differences when learning node representation. Finally, a representation ali
gnment module aligns the pre-trained node representation using the original one,
 essentially allowing graph representations to learn structural information with
out losing their original semantic information, thereby leading to better graph
representations. Extensive experiments on a series of real-world datasets demons
trate that the proposed CenPre outperforms the state-of-the-art baselines in the
 tasks of node classification, link prediction, and graph classification.
****

Victor Priser,PASCAL BIANCHI,Adil Salim
Long-time asymptotics of noisy SVGD outside the population limit
Stein Variational Gradient Descent (SVGD) is a widely used sampling algorithm th
at has been successfully applied in several areas of Machine Learning. SVGD oper
ates by iteratively moving a set of n interacting particles (which represent t

he samples) to approximate the target distribution. Despite recent studies on th
e complexity of SVGD and its variants, their long-time asymptotic behavior (i.e.
, after numerous iterations k) is still not understood in the finite number of
 particles regime. We study the long-time asymptotic behavior of a noisy variant
 of SVGD. First, we establish that the limit set of noisy SVGD for large k is
well-defined. We then characterize this limit set, showing that it approaches th
e target distribution as n increases. In particular, noisy SVGD avoids the var
iance collapse observed for SVGD. Our approach involves demonstrating that the t
rajectories of noisy SVGD closely resemble those described by a McKean-Vlasov pr
ocess.
****
Rubén Ballester,Ernst Röell,Daniel Bin Schmid,Mathieu Alain,Sergio Escalera,Carl
es Casacuberta,Bastian Rieck
MANTRA: The Manifold Triangulations Assemblage
The rising interest in leveraging higher-order interactions present in complex s
ystems has led to a surge in more expressive models exploiting high-order struct
ures in the data, especially in topological deep learning (TDL), which designs n
eural networks on high-order domains such as simplicial complexes. However, prog
ress in this field is hindered by the scarcity of datasets for benchmarking thes
e architectures. To address this gap, we introduce MANTRA, the first large-scale
, diverse, and intrinsically high-order dataset for benchmarking high-order mode
ls, comprising over 43,000 and 250,000 triangulations of surfaces and three-dime
nsional manifolds, respectively. With MANTRA, we assess several graph- and simpl
icial complex-based models on three topological classification tasks. We demonst
rate that while simplicial complex-based neural networks generally outperform th
eir graph-based counterparts in capturing simple topological invariants, they al
so struggle, suggesting a rethink of TDL. Thus, MANTRA serves as a benchmark fo
r assessing and advancing topological
methods, leading the way for more effective high-order models.
****
Fawaz Sammani,Nikos Deligiannis
Zero-Shot Natural Language Explanations
Natural Language Explanations (NLEs) interpret the decision-making process of a
given model through textual sentences. Current NLEs suffer from a severe limitat
ion; they are unfaithful to the model’s actual reasoning process, as a separate
textual decoder is explicitly trained to generate those explanations using annot
ated datasets for a specific task, leading them to reflect what annotators desir
e. In this work, we take the first step towards generating faithful NLEs for any
 visual classification model without any training data. Our approach models the
relationship between class embeddings from the classifier of the vision model an
d their corresponding class names via a simple MLP which trains in seconds. Afte
r training, we can map any new text to the classifier space and measure its asso
ciation with the visual features. We conduct experiments on 38 vision models, in
cluding both CNNs and Transformers. In addition to NLEs, our method offers other
 advantages such as zero-shot image classification and fine-grained concept disc
overy.
****
Jingcheng Deng,Zihao Wei,Liang Pang,Hanxing Ding,Huawei Shen,Xueqi Cheng
Everything is Editable: Extend Knowledge Editing to Unstructured Data in Large L
anguage Models
Recent knowledge editing methods have primarily focused on modifying structured
knowledge in large language models. However, this task setting overlooks the fac
t that a significant portion of real-world knowledge is stored in an unstructure
d format, characterized by long-form content, noise, and a complex yet comprehen
sive nature.
Techniques like "local layer key-value storage" and "term-driven optimization",
as used in previous methods like MEMIT, are not effective for handling unstructu
red knowledge.
To address these challenges, we propose a novel Unstructured Knowledge Editing m
ethod, namely UnKE, which extends previous assumptions in the layer dimension an

d token dimension.
Firstly, in the layer dimension, we propose non-local block key-value storage to replace local layer key-value storage, increasing the representation ability of key-value pairs and incorporating attention layer knowledge.
Secondly, in the token dimension, we replace "term-driven optimization" with "cause-driven optimization", which edits the last token directly while preserving context, avoiding the need to locate terms and preventing the loss of context information.
Results on newly proposed unstructured knowledge editing dataset (UnKEBench) and traditional structured datasets demonstrate that UnKE achieves remarkable performance, surpassing strong baselines. In addition, UnKE has robust batch editing and sequential editing capabilities.
****
Louis Bradshaw,Simon Colton
Aria-MIDI: A Dataset of Piano MIDI Files for Symbolic Music Modeling
We introduce an extensive new dataset of MIDI files, created by transcribing audio recordings of piano performances into their constituent notes. The data pipeline we use is multi-stage, employing a language model to autonomously crawl and score audio recordings from the internet based on their metadata, followed by a stage of pruning and segmentation using an audio classifier. The resulting dataset contains over one million distinct MIDI files, comprising roughly 100,000 hours of transcribed audio. We provide an in-depth analysis of our techniques, offering statistical insights, and investigate the content by extracting metadata tags, which we also provide. Dataset available at https://github.com/loubbrad/aria-midi.
****
Shijie Liu,Andrew Craig Cullen,Paul Montague,Sarah Monazam Erfani,Benjamin I. P. Rubinstein
Multi-level Certified Defense Against Poisoning Attacks in Offline Reinforcement Learning
Similar to other machine learning frameworks, Offline Reinforcement Learning (RL) is shown to be vulnerable to poisoning attacks, due to its reliance on externally sourced datasets, a vulnerability that is exacerbated by its sequential nature. To mitigate the risks posed by RL poisoning, we extend certified defenses to provide larger guarantees against adversarial manipulation, ensuring robustness for both per-state actions, and the overall expected cumulative reward. Our approach leverages properties of Differential Privacy, in a manner that allows this work to span both continuous and discrete spaces, as well as stochastic and deterministic environments---significantly expanding the scope and applicability of achievable guarantees. Empirical evaluations demonstrate that our approach ensures the performance drops to no more than 50% with up to 7% of the training data poisoned, significantly improving over the 0.008% in prior work (Wu et al., 2022), while producing certified radii that is 5 times larger as well. This highlights the potential of our framework to enhance safety and reliability in offline RL.
****
Massimo Bini,Leander Girrbach,Zeynep Akata
Decoupling Angles and Strength in Low-rank Adaptation
Parameter-Efficient FineTuning (PEFT) methods have recently gained significant popularity thanks to the widespread availability of large-scale pretrained models. These methods allow for quick adaptation to downstream tasks with minimal computational cost. However, popular finetuning methods such as LoRA exhibit limited robustness when it comes to hyperparameter choices or extended training regimes, preventing optimal out-of-the-box performance. In contrast, bounded approaches, such as ETHER, provide greater robustness but are limited to extremely low-rank adaptations and fixed-strength transformations, reducing their adaptation expressive power. In this work, we propose Decoupled Low-rank Adaptation (DeLoRA), a novel finetuning method that normalizes and scales learnable low-rank matrices. By bounding the distance of the transformation, DeLoRA effectively decouples the angular learning from the adaptation strength, enhancing robustness without co

mpromising performance. Through evaluations on subject-driven image generation, natural language understanding, and instruction tuning, we show that DeLoRA matches or surpasses performance of competing PEFT methods, while exhibiting stronger robustness. Code is available at https://github.com/ExplainableML/DeLoRA.
****

Yue Yang,Shuibo Zhang,Kaipeng Zhang,Yi Bin,Yu Wang,Ping Luo,Wenqi Shao
Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping
Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities across multimodal tasks such as visual perception and reasoning, leading to good performance on various multimodal evaluation benchmarks. However, these benchmarks keep a static nature and overlap with the pre-training data, resulting in fixed complexity constraints and data contamination issues. This raises the concern regarding the validity of the evaluation. To address these two challenges, we introduce a dynamic multimodal evaluation protocol called Vision-Language Bootstrapping (VLB). VLB provides a robust and comprehensive assessment for LVLMs with reduced data contamination and flexible complexity. To this end, VLB dynamically generates new visual question-answering samples through a multimodal bootstrapping module that modifies both images and language, while ensuring that newly generated samples remain consistent with the original ones by a judge module. By composing various bootstrapping strategies, VLB offers dynamic variants of existing benchmarks with diverse complexities, enabling the evaluation to co-evolve with the ever-evolving capabilities of LVLMs. Extensive experimental results across multiple benchmarks, including SEEDBench, MMBench, and MME, show that VLB significantly reduces data contamination and exposes performance limitations of LVLMs.
****

Xu Zheng,Farhad Shirani,Zhuomin Chen,Chaohao Lin,Wei Cheng,Wenbo Guo,Dongsheng Luo
F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI
Recent research has developed a number of eXplainable AI (XAI) techniques, such as gradient-based approaches, input perturbation-base methods, and black-box explanation methods. While these XAI techniques can extract meaningful insights from deep learning models, how to properly evaluate them remains an open problem. The most widely used approach is to perturb or even remove what the XAI method considers to be the most important features in an input and observe the changes in the output prediction. This approach, although straightforward, suffers the Out-of-Distribution (OOD) problem as the perturbed samples may no longer follow the original data distribution. A recent method RemOve And Retrain (ROAR) solves the OOD issue by retraining the model with perturbed samples guided by explanations. However, using the model retrained based on XAI methods to evaluate these explainers may cause information leakage and thus lead to unfair comparisons. We propose Fine-tuned Fidelity (F-Fidelity), a robust evaluation framework for XAI, which utilizes i) an explanation-agnostic fine-tuning strategy, thus mitigating the information leakage issue, and ii) a random masking operation that ensures that the removal step does not generate an OOD input. We also design controlled experiments with state-of-the-art (SOTA) explainers and their degraded version to verify the correctness of our framework. We conduct experiments on multiple data modalities, such as images, time series, and natural language. The results demonstrate that F-Fidelity significantly improves upon prior evaluation metrics in recovering the ground-truth ranking of the explainers. Furthermore, we show both theoretically and empirically that, given a faithful explainer, F-Fidelity metric can be used to compute the sparsity of influential input components, i.e., to extract the true explanation size.
****

Muthu Chidambaram,Rong Ge
Reassessing How to Compare and Improve the Calibration of Machine Learning Models
A machine learning model is calibrated if its predicted probability for an outcome matches the observed frequency for that outcome conditional on the model pred

iction. This property has become increasingly important as the impact of machine learning models has continued to spread to various domains. As a result, there are now a dizzying number of recent papers on measuring and improving the calibration of (specifically deep learning) models. In this work, we reassess the reporting of calibration metrics in the recent literature. We show that there exist trivial recalibration approaches that can appear seemingly state-of-the-art unless calibration and prediction metrics (i.e. test accuracy) are accompanied by additional generalization metrics such as negative log-likelihood. We then use a calibration-based decomposition of Bregman divergences to develop a new extension to reliability diagrams that jointly visualizes calibration and generalization error, and show how our visualization can be used to detect trade-offs between calibration and generalization. Along the way, we prove novel results regarding the relationship between full calibration error and confidence calibration error for Bregman divergences. We also establish the consistency of the kernel regression estimator for calibration error used in our visualization approach, which generalizes existing consistency results in the literature.
****

Xin Zhang,Jiawei Du,Ping Liu,Joey Tianyi Zhou
Breaking Class Barriers: Efficient Dataset Distillation via Inter-Class Feature Compensator
Dataset distillation has emerged as a technique aiming to condense informative features from large, natural datasets into a compact and synthetic form. While recent advancements have refined this technique, its performance is bottlenecked by the prevailing class-specific synthesis paradigm. Under this paradigm, synthetic data is optimized exclusively for a pre-assigned one-hot label, creating an implicit class barrier in feature condensation. This leads to inefficient utilization of the distillation budget and oversight of inter-class feature distributions, which ultimately limits the effectiveness and efficiency, as demonstrated in our analysis.
To overcome these constraints, this paper presents the Inter-class Feature Compensator (INFER), an innovative distillation approach that transcends the class-specific data-label framework widely utilized in current dataset distillation methods. Specifically, INFER leverages a Universal Feature Compensator (UFC) to enhance feature integration across classes, enabling the generation of multiple additional synthetic instances from a single UFC input. This significantly improves the efficiency of the distillation budget.
Moreover, INFER enriches inter-class interactions during the distillation, thereby enhancing the effectiveness and generalizability of the distilled data. By allowing for the linear interpolation of labels similar to those in the original dataset, INFER meticulously optimizes the synthetic data and dramatically reduces the size of soft labels in the synthetic dataset to almost zero, establishing a new benchmark for efficiency and effectiveness in dataset distillation. In practice, INFER demonstrates state-of-the-art performance across benchmark datasets. For instance, in the $\texttt{ipc} = 50$ setting on ImageNet-1k with the same compression level, it outperforms SRe2L by 34.5\% using ResNet18. Codes are available at https://github.com/zhangxin-xd/UFC.
****

Maosheng Yang
Topological Schrödinger Bridge Matching
Given two boundary distributions, the \emph{Schrödinger Bridge} (SB) problem seeks the “most likely” random evolution between them with respect to a reference process. It has revealed rich connections to recent machine learning methods for generative modeling and distribution matching. While these methods perform well in Euclidean domains, they are not directly applicable to topological domains such as graphs and simplicial complexes, which are crucial for data defined over network entities, such as node signals and edge flows. In this work, we propose the \emph{Topological Schrödinger Bridge problem} ($\mathcal{T}$SBP) for matching signal distributions on a topological domain. We set the reference process to follow some linear tractable \emph{topology-aware} stochastic dynamics such as topological heat diffusion. For the case of Gaussian boundary distributions, we de

rive a \emph{closed-form} topological SB ($\mathcal{T}$SB) in terms of its time-marginal and stochastic differential. In the general case, leveraging the well-known result, we show that the optimal process follows the forward-backward topological dynamics governed by some unknowns. Building on these results, we develop $\mathcal{T}$SB-based models for matching topological signals by parameterizing the unknowns in the optimal process as \emph{(topological) neural networks} and learning them through \emph{likelihood training}. We validate the theoretical results and demonstrate the practical applications of $\mathcal{T}$SB-based models on both synthetic and real-world networks, emphasizing the role of topology. Additionally, we discuss the connections of $\mathcal{T}$SB-based models to other emerging models, and outline future directions for topological signal matching.
****

Fahad Sarfraz,Elahe Arani,Bahram Zonooz

Semantic Aware Representation Learning for Lifelong Learning

The human brain excels at lifelong learning by not only encoding information in sparse activation codes but also leveraging rich semantic structures and relationships between newly encountered and previously learned objects. This ability to utilize semantic similarities is crucial for efficient learning and knowledge consolidation, yet is often underutilized in current continual learning approaches. To bridge this gap, we propose Semantic-Aware Representation Learning (SARL) which employs sparse activations and a principled approach to evaluate similarities between objects encountered across different tasks and subsequently uses them to guide representation learning. Using these relationships, SARL enhances the reusability of features and reduces interference between tasks. This approach empowers the model to adapt to new information while maintaining stability, significantly improving performance in complex incremental learning scenarios. Our analysis demonstrates that SARL achieves a superior balance between plasticity and stability by harnessing the underlying semantic structure.
****

Sangmin Bae,Adam Fisch,Hrayr Harutyunyan,Ziwei Ji,Seungyeon Kim,Tal Schuster

Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA

Large language models (LLMs) are expensive to deploy. Parameter sharing offers a possible path towards reducing their size and cost, but its effectiveness in modern LLMs remains fairly limited. In this work, we revisit "layer tying" as form of parameter sharing in Transformers, and introduce novel methods for converting existing LLMs into smaller "Recursive Transformers" that share parameters across layers, with minimal loss of performance. Here, our Recursive Transformers are efficiently initialized from standard pretrained Transformers, but only use a single block of unique layers that is then repeated multiple times in a loop. We further improve performance by introducing Relaxed Recursive Transformers that add flexibility to the layer tying constraint via depth-wise low-rank adaptation (LoRA) modules, yet still preserve the compactness of the overall model. We show that our recursive models (e.g., recursive Gemma 1B) outperform both similar-sized vanilla pretrained models (such as TinyLlama 1.1B and Pythia 1B) and knowledge distillation baselines---and can even recover most of the performance of the original "full-size" model (e.g., Gemma 2B with no shared parameters). Finally, we propose Continuous Depth-wise Batching, a promising new inference paradigm enabled by the Recursive Transformer when paired with early exiting. In a theoretical analysis, we show that this has the potential to lead to significant (2-3 $\times$) gains in inference throughput.
****

Aishwarya Jayagopal,Yanrong Zhang,Robert John Walsh,Tuan Zea Tan,Anand D Jeyasekharan,Vaibhav Rajan

GANDALF: Generative AttentioN based Data Augmentation and predictive modeLing Framework for personalized cancer treatment

Effective treatment of cancer is a major challenge faced by healthcare providers, due to the highly individualized nature of patient responses to treatment. This is caused by the heterogeneity seen in cancer-causing alterations (mutations) across patient genomes. Limited availability of response data in patients makes it difficult to train personalized treatment recommendation models on mutations

from clinical genomic sequencing reports. Prior methods tackle this by utilising larger, labelled pre-clinical laboratory datasets (‘cell lines’), via transfer learning. These methods augment patient data by learning a shared, domain-invariant representation, between the cell line and patient domains, which is then used to train a downstream drug response prediction (DRP) model. This approach augments data in the shared space but fails to model patient-specific characteristics, which have a strong influence on their drug response. We propose a novel generative attention-based data augmentation and predictive modeling framework, GANDALF, to tackle this crucial shortcoming of prior methods. GANDALF not only augments patient genomic data directly, but also accounts for its domain-specific characteristics. GANDALF outperforms state-of-the-art DRP models on publicly available patient datasets and emerges as the front-runner amongst SOTA cancer DRP models.
****
Insoo Kim,Hana Lee,Hyong-Euk Lee,Jinwoo Shin
Controllable Blur Data Augmentation Using 3D-Aware Motion Estimation
Existing realistic blur datasets provide insufficient variety in scenes and blur patterns to be trained, while expanding data diversity demands considerable time and effort due to complex dual-camera systems. To address the challenge, data augmentation can be an effective way to artificially increase data diversity. However, existing methods on this line are typically designed to estimate motions from a 2D perspective, e.g., estimating 2D non-uniform kernels disregarding 3D aspects of blur modeling, which leads to unrealistic motion patterns due to the fact that camera and object motions inherently arise in 3D space. In this paper, we propose a 3D-aware blur synthesizer capable of generating diverse and realistic blur images for blur data augmentation. Specifically, we estimate 3D camera positions within the motion blur interval, generate the corresponding scene images, and aggregate them to synthesize a realistic blur image. Since the 3D camera positions projected onto the 2D image plane inherently lie in 2D space, we can represent the 3D transformation as a combination of 2D transformation and projected 3D residual component. This allows for 3D transformation without requiring explicit depth measurements, as the 3D residual component is directly estimated via a neural network. Furthermore, our blur synthesizer allows for controllable blur data augmentation by modifying blur magnitude, direction, and scenes, resulting in diverse blur images. As a result, our method significantly improves deblurring performance, making it more practical for real-world scenarios.
****
Jongmin Lee,Ernest K. Ryu
Optimal Non-Asymptotic Rates of Value Iteration for Average-Reward Markov Decision Processes
While there is an extensive body of research on the analysis of Value Iteration (VI) for discounted cumulative-reward MDPs, prior work on analyzing VI for (undiscounted) average-reward MDPs has been limited, and most prior results focus on asymptotic rates in terms of Bellman error. In this work, we conduct refined non-asymptotic analyses of average-reward MDPs, obtaining a collection of convergence results advancing our understanding of the setup. Among our new results, most notable are the $\mathcal{O}(1/k)$-rates of Anchored Value Iteration on the Bellman error under the multichain setup and the span-based complexity lower bound that matches the $\mathcal{O}(1/k)$ upper bound up to a constant factor of 8 in the weakly communicating and unichain setups.
****
Shi Fu,Yingjie Wang,Yuzhu Chen,Xinmei Tian,Dacheng Tao
A Theoretical Perspective: How to Prevent Model Collapse in Self-consuming Training Loops
High-quality data is essential for training large generative models, yet the vast reservoir of real data available online has become nearly depleted. Consequently, models increasingly generate their own data for further training, forming Self-consuming Training Loops (STLs). However, the empirical results have been strikingly inconsistent: some models degrade or even collapse, while others successfully avoid these failures, leaving a significant gap in theoretical understandi

ng to explain this discrepancy. This paper introduces the intriguing notion of * recursive stability* and presents the first theoretical generalization analysis, revealing how both model architecture and the proportion between real and synthetic data influence the success of STLs. We further extend this analysis to transformers in in-context learning, showing that even a constant-sized proportion of real data ensures convergence, while also providing insights into optimal synthetic data sizing.
****
Fanqing Meng,Jin Wang,Chuanhao Li,Quanfeng Lu,Hao Tian,Tianshuo Yang,Jiaqi Liao, Xizhou Zhu,Jifeng Dai,Yu Qiao,Ping Luo,Kaipeng Zhang,Wenqi Shao
MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models
The capability to process multiple images is crucial for Large Vision-Language Models (LVLMs) to develop a more thorough and nuanced understanding of a scene. Recent multi-image LVLMs have begun to address this need. However, their evaluation has not kept pace with their development. To fill this gap, we introduce the Multimodal Multi-image Understanding (MMIU) benchmark, a comprehensive evaluation suite designed to assess LVLMs across a wide range of multi-image tasks. MMIU encompasses 7 types of multi-image relationships, 52 tasks, 77K images, and 11K meticulously curated multiple-choice questions, making it the most extensive benchmark of its kind. Our evaluation of nearly 30 popular LVLMs, including both open-source and proprietary models, reveals significant challenges in multi-image comprehension, particularly in tasks involving spatial understanding. Even the most advanced models, such as GPT-4o, achieve only 55.7\% accuracy on MMIU. Through multi-faceted analytical experiments, we identify key performance gaps and limitations, providing valuable insights for future model and data improvements. We aim for MMIU to advance the frontier of LVLM research and development. We release the data and code at https://github.com/MMIUBenchmark/MMIU.
****
Hanyu Wang,Saksham Suri,Yixuan Ren,Hao Chen,Abhinav Shrivastava
LARP: Tokenizing Videos with a Learned Autoregressive Generative Prior
We present LARP, a novel video tokenizer designed to overcome limitations in current video tokenization methods for autoregressive (AR) generative models. Unlike traditional patchwise tokenizers that directly encode local visual patches into discrete tokens, LARP introduces a holistic tokenization scheme that gathers information from the visual content using a set of learned holistic queries. This design allows LARP to capture more global and semantic representations, rather than being limited to local patch-level information. Furthermore, it offers flexibility by supporting an arbitrary number of discrete tokens, enabling adaptive and efficient tokenization based on the specific requirements of the task. To align the discrete token space with downstream AR generation tasks, LARP integrates a lightweight AR transformer as a training-time prior model that predicts the next token on its discrete latent space. By incorporating the prior model during training, LARP learns a latent space that is not only optimized for video reconstruction but is also structured in a way that is more conducive to autoregressive generation. Moreover, this process defines a sequential order for the discrete tokens, progressively pushing them toward an optimal configuration during training, ensuring smoother and more accurate AR generation at inference time. Comprehensive experiments demonstrate LARPs strong performance, achieving state-of-the-art FVD on the UCF101 class-conditional video generation benchmark. LARP enhances the compatibility of AR models with videos and opens up the potential to build unified high-fidelity multimodal large language models (MLLMs). Project page: https://hywang66.github.io/larp/
****
John Xavier Morris,Alexander M Rush
Contextual Document Embeddings
Dense document embeddings are central to neural retrieval. The dominant paradigm is to train and construct embeddings by running encoders directly on individual documents. In this work, we argue that these embeddings, while effective, are implicitly out-of-context for targeted use cases of retrieval, and that a context

ualized document embedding should take into account both the document and neighb
oring documents in context - analogous to contextualized word embeddings. We pro
pose two complementary methods for contextualized document embeddings: first, an
 alternative contrastive learning objective that explicitly incorporates the doc
ument neighbors into the intra-batch contextual loss; second, a new contextual a
rchitecture that explicitly encodes neighbor document information into the encod
ed representation. Results show that both methods achieve better performance tha
n biencoders in several settings, with differences especially pronounced out-of-
domain. We achieve state-of-the-art results on the MTEB benchmark with no hard n
egative mining, score distillation, dataset-specific instructions, intra-GPU exa
mple-sharing, or extremely large batch sizes. Our method can be applied to impr
ove performance on any contrastive learning dataset and any biencoder.
****

Yiyang Zhou,Zhaoyang Wang,Tianle Wang,Shangyu Xing,Peng Xia,Bo Li,Kaiyuan Zheng,
Zijian Zhang,Zhaorun Chen,Wenhao Zheng,Xuchao Zhang,Chetan Bansal,Weitong Zhang,
Ying Wei,Mohit Bansal,Huaxiu Yao
Anyprefer: An Agentic Framework for Preference Data Synthesis
High-quality preference data is essential for aligning foundation models with hu
man values through preference learning. However, manual annotation of such data
is often time-consuming and costly. Recent methods often adopt a self-rewarding
approach, where the target model generates and annotates its own preference data
, but this can lead to inaccuracies since the reward model shares weights with t
he target model, thereby amplifying inherent biases. To address these issues, we
 propose Anyprefer, a framework designed to synthesize high-quality preference d
ata for aligning the target model. Anyprefer frames the data synthesis process a
s a cooperative two-player Markov Game, where the target model and the judge mod
el collaborate together. Here, a series of external tools are introduced to assi
st the judge model in accurately rewarding the target model’s responses, mitigat
ing biases in the rewarding process. In addition, a feedback mechanism is introd
uced to optimize prompts for both models, enhancing collaboration and improving
data quality.
The synthesized data is compiled into a new preference dataset, Anyprefer-V1, co
nsisting of 58K high-quality preference pairs.
Extensive experiments show that Anyprefer significantly improves model alignment
 performance across four main applications, covering 21 datasets, achieving aver
age improvements of 18.55% in five natural language generation datasets, 3.66% i
n nine vision-language understanding datasets, 30.05% in three medical image ana
lysis datasets, and 16.00% in four visuo-motor control tasks.
****

Jiahuan Yan,Jintai Chen,Chaowen Hu,Bo Zheng,Yaojun Hu,Jimeng Sun,Jian Wu
Small Models are LLM Knowledge Triggers for Medical Tabular Prediction
Recent development in large language models (LLMs) has demonstrated impressive d
omain proficiency on unstructured textual or multi-modal tasks. However, despite
 with intrinsic world knowledge, their application on structured tabular data pr
ediction still lags behind, primarily due to the numerical insensitivity and mod
ality discrepancy that brings a gap between LLM reasoning and statistical tabula
r learning. Unlike textual or vision data (e.g., electronic clinical notes or me
dical imaging data), tabular data is often presented in heterogeneous numerical
values (e.g., CBC reports). This ubiquitous data format requires intensive exper
t annotation, and its numerical nature limits LLMs' capability to effectively tr
ansfer untapped domain expertise. In this paper, we propose SERSAL, a general se
lf-prompting method by synergy learning with small models to enhance LLM tabular
 prediction in an unsupervised manner. Specifically, SERSAL utilizes the LLM's p
rior outcomes as original soft noisy annotations, which are dynamically leverage
d to teach a better small student model. Reversely, the outcomes from the traine
d small model are used to teach the LLM to further refine its real capability. T
his process can be repeatedly applied to gradually distill refined knowledge for
 continuous progress. Comprehensive experiments on widely used medical domain ta
bular datasets show that, without access to gold labels, applying SERSAL to Open
AI GPT reasoning process attains substantial improvement compared to linguistic

prompting methods, which serves as an orthogonal direction for tabular LLM, and increasing prompting bonus is observed as more powerful LLMs appear. Codes are available at https://github.com/jyansir/sersal.
****

Zhiteng Li,Lele Chen,Jerone Andrews,Yunhao Ba,Yulun Zhang,Alice Xiang
GenDataAgent: On-the-fly Dataset Augmentation with Synthetic Data
We propose a generative agent that augments training datasets with synthetic data for model fine-tuning. Unlike prior work, which uniformly samples synthetic data, our agent iteratively generates relevant samples on-the-fly, aligning with the target distribution. It prioritizes synthetic data that complements difficult training samples, focusing on those with high variance in gradient updates. Experiments across several image classification tasks demonstrate the effectiveness of our approach.
****

Jiaxi Yang,Wenlong Deng,Benlin Liu,Yangsibo Huang,James Zou,Xiaoxiao Li
GMValuator: Similarity-based Data Valuation for Generative Models
Data valuation plays a crucial role in machine learning. Existing data valuation methods, mainly focused on discriminative models, overlook generative models that have gained attention recently. In generative models, data valuation measures the impact of training data on generated datasets. Very few existing attempts at data valuation methods designed for deep generative models either concentrate on specific models or lack robustness in their outcomes. Moreover, efficiency still reveals vulnerable shortcomings. We formulate the data valuation problem in generative models from a similarity matching perspective to bridge the gaps. Specifically, we introduce Generative Model Valuator (GMValuator), the first training-free and model-agnostic approach to providing data valuation for image generation tasks. It empowers efficient data valuation through our innovative similarity matching module, calibrates biased contributions by incorporating image quality assessment, and attributes credits to all training samples based on their contributions to the generated samples. Additionally, we introduce four evaluation criteria for assessing data valuation methods in generative models. GMValuator is extensively evaluated on benchmark and high-resolution datasets and various mainstream generative architectures to demonstrate its effectiveness. Our code is available at: https://github.com/ubc-tea/GMValuator.
****

Hongwei Wen,Annika Betken,Hanyuan Hang
Optimal Learning of Kernel Logistic Regression for Complex Classification Scenarios
Complex classification scenarios, including long-tailed learning, domain adaptation, and transfer learning, present substantial challenges for traditional algorithms. Conditional class probability (CCP) predictions have recently become critical components of many state-of-the-art algorithms designed to address these challenging scenarios. Among kernel methods, kernel logistic regression (KLR) is distinguished by its effectiveness in predicting CCPs through the minimization of the cross-entropy (CE) loss. Despite the empirical success of CCP-based approaches, the theoretical understanding of their performance, particularly regarding the CE loss, remains limited. In this paper, we bridge this gap by demonstrating that KLR-based algorithms achieve minimax optimal convergence rates for the CE loss under mild assumptions in these complex tasks, thereby establishing their theoretical efficiency in such demanding contexts.
****

Wei Xiong,Chengshuai Shi,Jiaming Shen,Aviv Rosenberg,Zhen Qin,Daniele Calandriello,Misha Khalman,Rishabh Joshi,Bilal Piot,Mohammad Saleh,Chi Jin,Tong Zhang,Tianqi Liu
Building Math Agents with Multi-Turn Iterative Preference Learning
Recent studies have shown that large language models' (LLMs) mathematical problem-solving capabilities can be enhanced by integrating external tools, such as code interpreters, and employing multi-turn Chain-of-Thought (CoT) reasoning. While current methods focus on synthetic data generation and Supervised Fine-Tuning (SFT), this paper studies the complementary direct preference learning approach

to further improve model performance. However, existing direct preference learni
ng algorithms are originally designed for the single-turn chat task, and do not
fully address the complexities of multi-turn reasoning and external tool integra
tion required for tool-integrated mathematical reasoning tasks. To fill in this
gap, we introduce a multi-turn direct preference learning framework, tailored fo
r this context, that leverages feedback from code interpreters and optimizes tra
jectory-level preferences. This framework includes multi-turn DPO and multi-turn
 KTO as specific implementations. The effectiveness of our framework is validate
d through training of various language models using an augmented prompt set from
 the GSM8K and MATH datasets. Our results demonstrate substantial improvements:
a supervised fine-tuned Gemma-1.1-it-7B model's performance increased from 77.5%
 to 83.9% on GSM8K and from 46.1% to 51.2% on MATH. Similarly, a Gemma-2-it-9B m
odel improved from 84.1% to 86.3% on GSM8K and from 51.0% to 54.5% on MATH.
****

Yue Cheng,Bochen Lyu,Weiwei Xing,Zhanxing Zhu
DyCAST: Learning Dynamic Causal Structure from Time Series
Understanding the dynamics of causal structures is crucial for uncovering the un
derlying processes in time series data. Previous approaches rely on static assum
ptions, where contemporaneous and time-lagged dependencies are assumed to have i
nvariant topological structures. However, these models fail to capture the evolv
ing causal relationship between variables when the underlying process exhibits s
uch dynamics. To address this limitation, we propose DyCAST, a novel framework d
esigned to learn dynamic causal structures in time series using Neural Ordinary
Differential Equations (Neural ODEs). The key innovation lies in modeling the te
mporal dynamics of the contemporaneous structure, drawing inspiration from recen
t advances in Neural ODEs on constrained manifolds. We reformulate the task of l
earning causal structures at each time step as solving the solution trajectory o
f a Neural ODE on the directed acyclic graph (DAG) manifold. To accommodate high
-dimensional causal structures, we extend DyCAST by learning the temporal dynami
cs of the hidden state for contemporaneous causal structure. Experiments on both
 synthetic and real-world datasets demonstrate that DyCAST achieves superior or
comparable performance compared to existing causal discovery models.
****

Yizhou Xu,Liu Ziyin
Three Mechanisms of Feature Learning in a Linear Network
Understanding the dynamics of neural networks in different width regimes is cruc
ial for improving their training and performance. We present an exact solution f
or the learning dynamics of a one-hidden-layer linear network, with one-dimensio
nal data, across any finite width, uniquely exhibiting both kernel and feature l
earning phases. This study marks a technical advancement by enabling the analysi
s of the training trajectory from any initialization and a detailed phase diagra
m under varying common hyperparameters such as width, layer-wise learning rates,
 and scales of output and initialization. We identify three novel prototype mech
anisms specific to the feature learning regime: (1) learning by alignment, (2) l
earning by disalignment, and (3) learning by rescaling, which contrast starkly w
ith the dynamics observed in the kernel regime. Our theoretical findings are sub
stantiated with empirical evidence showing that these mechanisms also manifest i
n deep nonlinear networks handling real-world tasks, enhancing our understanding
 of neural network training dynamics and guiding the design of more effective le
arning strategies.
****

Olga Solodova,Nick Richardson,Deniz Oktay,Ryan P Adams
Graph Neural Networks Gone Hogwild
Graph neural networks (GNNs) appear to be powerful tools to learn state represen
tations for agents in distributed, decentralized multi-agent systems, but genera
te catastrophically incorrect predictions when nodes update asynchronously durin
g inference.
  This failure under asynchrony effectively excludes these architectures from ma
ny potential applications where synchrony is difficult or impossible to enforce,
 e.g., robotic swarms or sensor networks.

In this work we identify ''implicitly-defined'' GNNs as a class of architectur
es which is provably robust to asynchronous ''hogwild'' inference, adapting conv
ergence guarantees from work in asynchronous and distributed optimization.
We then propose a novel implicitly-defined GNN architecture, which we call an
energy GNN.
We show that this architecture outperforms other GNNs from this class on a var
iety of synthetic tasks inspired by multi-agent systems.
****
Sijin Chen,Omar Hagrass,Jason Matthew Klusowski
Decoding Game: On Minimax Optimality of Heuristic Text Generation Strategies
Decoding strategies play a pivotal role in text generation for modern language m
odels, yet a puzzling gap divides theory and practice. Surprisingly, strategies
that should intuitively be optimal, such as Maximum a Posteriori (MAP), often pe
rform poorly in practice. Meanwhile, popular heuristic approaches like Top-k a
nd Nucleus sampling, which employ truncation and normalization of the conditiona
l next-token probabilities, have achieved great empirical success but lack theor
etical justifications. In this paper, we propose Decoding Game, a comprehensive
theoretical framework which reimagines text generation as a two-player zero-sum
game between Strategist, who seeks to produce text credible in the true distribu
tion, and Nature, who distorts the true distribution adversarially. After discus
sing the decomposibility of multi-step generation, we derive the optimal strateg
y in closed form for one-step Decoding Game. It is shown that the adversarial Na
ture imposes an implicit regularization on likelihood maximization, and truncati
on-normalization methods are first-order approximations to the optimal strategy
under this regularization. Additionally, by generalizing the objective and param
eters of Decoding Game, near-optimal strategies encompass diverse methods such a
s greedy search, temperature scaling, and hybrids thereof. Numerical experiments
are conducted to complement our theoretical analysis.
****
Shaden Naif Alshammari,John R. Hershey,Axel Feldmann,William T. Freeman,Mark Ham
ilton
A Unifying Framework for Representation Learning
As the field of representation learning grows, there has been a proliferation of
different loss functions to solve different classes of problems. We introduce a
single information-theoretic equation that generalizes a large collection of mo
d- ern loss functions in machine learning. In particular, we introduce a framewo
rk that shows that several broad classes of machine learning methods are precise
ly minimizing an integrated KL divergence between two conditional distributions:
the supervisory and learned representations. This viewpoint exposes a hidden in
formation geometry underlying clustering, spectral methods, dimensionality re- d
uction, contrastive learning, and supervised learning. This framework enables th
e development of new loss functions by combining successful techniques from acro
ss the literature. We not only present a wide array of proofs, connecting over 2
3 different approaches, but we also leverage these theoretical results to create
state-of-the-art unsupervised image classifiers that achieve a +8% improvement
over the prior state-of-the-art on unsupervised classification on ImageNet-1K. W
e also demonstrate that I-Con can be used to derive principled debiasing methods
which improve contrastive representation learners.
****
Marcus Williams,Micah Carroll,Adhyyan Narang,Constantin Weisser,Brendan Murphy,A
nca Dragan
On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback
As LLMs become more widely deployed, there is increasing interest in directly op
timizing for feedback from end users (e.g. thumbs up) in addition to feedback fr
om paid annotators. However, training to maximize human feedback creates a perve
rse incentive structure for the AI to resort to manipulative or deceptive tactic
s to obtain positive feedback from users who are vulnerable to such strategies.
We study this phenomenon by training LLMs with Reinforcement Learning with simul
ated user feedback in environments of practical LLM usage. In our settings, we f
ind that: 1) Extreme forms of "feedback gaming" such as manipulation and decepti

on are learned reliably; 2) Even if only 2% of users are vulnerable to manipulat
ive strategies, LLMs learn to identify and target them while behaving appropriat
ely with other users, making such behaviors harder to detect; 3) To mitigate thi
s issue, it may seem promising to leverage continued safety training or LLM-as-j
udges during training to filter problematic outputs. Instead, we found that whil
e such approaches help in some of our settings, they backfire in others, sometim
es even leading to subtler manipulative behaviors. We hope our results can serve
 as a case study which highlights the risks of using gameable feedback sources -
- such as user feedback -- as a target for RL. Our code is publicly available. W
arning: some of our examples may be upsetting.
****

Can Rong,Jingtao Ding,Yan Liu,Yong Li
A Large-scale Dataset and Benchmark for Commuting Origin-Destination Flow Genera
tion
Commuting Origin-Destination~(OD) flows are critical inputs for urban planning a
nd transportation, providing crucial information about the population residing i
n one region and working in another within an interested area. Due to the high c
ost of data collection, researchers have developed physical and computational mo
dels to generate commuting OD flows using readily available urban attributes, su
ch as sociodemographics and points of interest, for cities lacking historical OD
 flows \textemdash commuting OD flow generation. Existing works developed models
 based on different techniques and achieved improvement on different datasets wi
th different evaluation metrics, which hinderes establishing a unified standard
for comparing model performance. To bridge this gap, we introduce a large-scale
dataset containing commuting OD flows for 3,333 areas including a wide range of
urban environments around the United States. Based on that, we benchmark widely
used models for commuting OD flow generation. We surprisingly find that the netw
ork-based generative models achieve the optimal performance in terms of both pre
cision and generalization ability, which may inspire new research directions of
graph generative modeling in this field. The dataset and benchmark are available
 at https://anonymous.4open.science/r/CommutingODGen-Dataset-0D4C/.
****

Hung Le,Dung Nguyen,Kien Do,Sunil Gupta,Svetha Venkatesh
Stable Hadamard Memory: Revitalizing Memory-Augmented Agents for Reinforcement L
earning
Effective decision-making in partially observable environments demands robust me
mory management. Despite their success in supervised learning, current deep-lear
ning memory models struggle in reinforcement learning environments that are part
ially observable and long-term. They fail to efficiently capture relevant past i
nformation, adapt flexibly to changing observations, and maintain stable updates
 over long episodes. We theoretically analyze the limitations of existing memory
 models within a unified framework and introduce the Stable Hadamard Memory, a n
ovel memory model for reinforcement learning agents. Our model dynamically adjus
ts memory by erasing no longer needed experiences and reinforcing crucial ones c
omputationally efficiently. To this end, we leverage the Hadamard product for ca
librating and updating memory, specifically designed to enhance memory capacity
while mitigating numerical and learning challenges. Our approach significantly o
utperforms state-of-the-art memory-based methods on challenging partially observ
able benchmarks, such as meta-reinforcement learning, long-horizon credit assign
ment, and POPGym, demonstrating superior performance in handling long-term and e
volving contexts.
****

Anders Aamand,Justin Y. Chen,Siddharth Gollapudi,Sandeep Silwal,Hao WU
Learning-Augmented Frequent Directions
An influential paper of Hsu et al. (ICLR'19) introduced the study of learning-au
gmented streaming algorithms in the context of frequency estimation. A fundament
al problem in the streaming literature, the goal of frequency estimation is to a
pproximate the number of occurrences of items appearing in a long stream of data
 using only a small amount of memory. Hsu et al. develop a natural framework to
combine the worst-case guarantees of popular solutions such as CountMin and Coun

tSketch with learned predictions of high frequency elements. They demonstrate th
at learning the underlying structure of data can be used to yield better streami
ng algorithms, both in theory and practice.

We simplify and generalize past work on learning-augmented frequency estimation.
 Our first contribution is a learning-augmented variant of the Misra-Gries algor
ithm which improves upon the error of learned CountMin and learned CountSketch a
nd achieves the state-of-the-art performance of randomized algorithms (Aamand et
 al., NeurIPS'23) with a simpler, deterministic algorithm. Our second contributi
on is to adapt learning-augmentation to a high-dimensional generalization of fre
quency estimation corresponding to finding important directions (top singular ve
ctors) of a matrix given its rows one-by-one in a stream. We analyze a learning-
augmented variant of the Frequent Directions algorithm, extending the theoretica
l and empirical understanding of learned predictions to matrix streaming.
****
Wei Liu,Zhiying Deng,Zhongyu Niu,Jun Wang,Haozhao Wang,Zhigang Zeng,Ruixuan Li
Breaking Free from MMI: A New Frontier in Rationalization by Probing Input Utili
zation
Extracting a small subset of crucial rationales from the full input is a key pro
blem in explainability research. The most widely used fundamental criterion for
rationale extraction is the maximum mutual information (MMI) criterion. In this
paper, we first demonstrate that MMI suffers from diminishing marginal returns.
Once part of the rationale has been identified, finding the remaining portions c
ontributes only marginally to increasing the mutual information, making it diffi
cult to use MMI to locate the rest. In contrast to MMI that aims to reproduce th
e prediction, we seek to identify the parts of the input that the network can ac
tually utilize. This is achieved by comparing how different rationale candidates
 match the capability space of the weight matrix. The weight matrix of a neural
network is typically low-rank, meaning that the linear combinations of its colum
n vectors can only cover part of the directions in a high-dimensional space (hig
h-dimension: the dimensions of an input vector). If an input is fully utilized b
y the network, it generally matches these directions (e.g., a portion of a hyper
sphere), resulting in a representation with a high norm. Conversely, if an input
 primarily falls outside (orthogonal to) these directions, its representation no
rm will approach zero, behaving like noise that the network cannot effectively u
tilize.
Building on this, we propose using the norms of rationale candidates as an alter
native objective to MMI.
Through experiments on four text classification datasets and one graph classific
ation dataset using three network architectures (GRUs, BERT, and GCN), we show t
hat our method outperforms MMI and its improved variants in identifying better r
ationales. We also compare our method with a representative LLM (llama-3.1-8b-in
struct) and find that our simple method gets comparable results to it and can so
metimes even outperform it.
****
Johan Bjorck,Alon Benhaim,Vishrav Chaudhary,Furu Wei,Xia Song
Scaling Optimal LR Across Token Horizons
State-of-the-art LLMs are powered by scaling -- scaling model size, training tok
ens, and cluster size. It is economically infeasible to extensively tune hyperpa
rameters for the largest runs. Instead, approximately optimal hyperparameters mu
st be inferred or transferred from smaller experiments. Hyperparameter transfer
across model sizes has been studied in muP. However, hyperparameter transfer acr
oss training tokens -- or token horizon -- has not been studied yet. To remedy t
his we conduct a large-scale empirical study on how optimal learning rate (LR) d
epends on the token horizon in LLM training. We first demonstrate that the optim
al LR changes significantly with token horizon -- longer training necessitates s
maller LR. Secondly, we demonstrate that the optimal LR follows a scaling law an
d that the optimal LR for longer horizons can be accurately estimated from short
er horizons via such scaling laws. We also provide a rule-of-thumb for transferr
ing LR across token horizons with zero overhead over current practices. Lastly,

we provide evidence that LLama-1 used too high LR, and thus argue that hyperpara
meter transfer across data size is an overlooked component of LLM training.
****
Yekun Chai,Haoran Sun,Huang Fang,Shuohuan Wang,Yu Sun,Hua Wu
MA-RLHF: Reinforcement Learning from Human Feedback with Macro Actions
Reinforcement learning from human feedback (RLHF) has demonstrated effectiveness
 in aligning large language models (LLMs) with human preferences. However, token
-level RLHF suffers from the credit assignment problem over long sequences, wher
e delayed rewards make it challenging for the model to discern which actions con
tributed to preferred outcomes. This hinders learning efficiency and slows conve
rgence.In this paper, we propose MA-RLHF, a simple yet effective RLHF framework
that incorporates macro actions --- sequences of tokens or higher-level language
 constructs --- into the learning process. By operating at higher level of abstr
action, our approach reduces the temporal distance between actions and rewards,
facilitating faster and more accurate credit assignment. This results in more st
able policy gradient estimates and enhances learning efficiency within each epis
ode, all without increasing computational complexity during training or inferenc
e. We validate our approach through extensive experiments across various model s
izes and tasks, including text summarization, dialogue generation, question answ
ering, and program synthesis. Our method achieves substantial performance improv
ements over standard RLHF, with performance gains of up to 30\% in text summariz
ation and code generation, 18\% in dialogue, and 8\% in question answering tasks
. Notably, our approach reaches parity with vanilla RLHF $1.7 \sim 2$ times fast
er in terms of training time and continues to outperform it with further trainin
g. We make our code and data publicly available at \url{https://github.com/ernie
-research/MA-RLHF}.
****
Tengfei Ma,Xiang song,Wen Tao,Mufei Li,Jiani Zhang,Xiaoqin Pan,Yijun Wang,Boshen
g Song,xiangxiang Zeng
Towards Synergistic Path-based Explanations for Knowledge Graph Completion: Expl
oration and Evaluation
Knowledge graph completion (KGC) aims to alleviate the inherent incompleteness o
f knowledge graphs (KGs), a crucial task for numerous applications such as recom
mendation systems and drug repurposing. The success of knowledge graph embedding
 (KGE) models provokes the question about the explainability: ``\textit{Which th
e patterns of the input KG are most determinant to the prediction}?'' Particular
ly, path-based explainers prevail in existing methods because of their strong ca
pability for human understanding. In this paper, based on the observation that a
 fact is usually determined by the synergy of multiple reasoning chains, we prop
ose a novel explainable framework, dubbed KGExplainer, to explore synergistic pa
thways. KGExplainer is a model-agnostic approach that employs a perturbation-bas
ed greedy search algorithm to identify the most crucial synergistic paths as exp
lanations within the local structure of target predictions. To evaluate the qual
ity of these explanations, KGExplainer distills an evaluator from the target KGE
 model, allowing for the examination of their fidelity. We experimentally demons
trate that the distilled evaluator has comparable predictive performance to the
target KGE. Experimental results on benchmark datasets demonstrate the effective
ness of KGExplainer, achieving a human evaluation accuracy of 83.3\% and showing
 promising improvements in explainability. Code is available at \url{https://git
hub.com/xiaomingaaa/KGExplainer}
****
Kiyoung Seong,Seonghyun Park,Seonghwan Kim,Woo Youn Kim,Sungsoo Ahn
Transition Path Sampling with Improved Off-Policy Training of Diffusion Path Sam
plers
Understanding transition pathways between two meta-stable states of a molecular
system is crucial to advance drug discovery and material design. However, unbias
ed molecular dynamics (MD) simulations are computationally infeasible because of
 the high energy barriers that separate these states. Although recent machine le
arning techniques are proposed to sample rare events, they are often limited to
simple systems and rely on collective variables (CVs) derived from costly domain

expertise. In this paper, we introduce a novel approach that trains diffusion p
ath samplers (DPS) to address the transition path sampling (TPS) problem without
requiring CVs. We reformulate the problem as an amortized sampling from the tra
nsition path distribution by minimizing the log-variance divergence between the
path distribution induced by DPS and the transition path distribution. Based on
the log-variance divergence, we propose learnable control variates to reduce the
variance of gradient estimators and the off-policy training objective with repl
ay buffers and simulated annealing techniques to improve sample efficiency and d
iversity. We also propose a scale-based equivariant parameterization of the bias
forces to ensure scalability for large systems. We extensively evaluate our app
roach, termed TPS-DPS, on a synthetic system, small peptide, and challenging fas
t-folding proteins, demonstrating that it produces more realistic and diverse tr
ansition pathways than existing baselines. We also provide links to [project pag
e](https://kiyoung98.github.io/tps-dps/) and [code](https://github.com/kiyoung98
/tps-dps).
****
Yifan Wang,Yifei Liu,Yingdong Shi,Changming Li,Anqi Pang,Sibei Yang,Jingyi Yu,Ka
n Ren
Discovering Influential Neuron Path in Vision Transformers
Vision Transformer models exhibit immense power yet remain opaque to human under
standing, posing challenges and risks for practical applications.
While prior research has attempted to demystify these models through input attri
bution and neuron role analysis,
there's been a notable gap in considering layer-level information and the holist
ic path of information flow across layers.
In this paper, we investigate the significance of influential neuron paths withi
n vision Transformers, which is a path of neurons from the model input to output
that impacts the model inference most significantly.
We first propose a joint influence measure to assess the contribution of a set o
f neurons to the model outcome.
And we further provide a
layer-progressive neuron locating
approach that efficiently selects the most influential neuron at each layer tryi
ng to discover the crucial neuron path from input to output within the target mo
del.
Our experiments demonstrate the superiority of our method finding the most influ
ential neuron path along which the information flows, over the existing baseline
solutions.
Additionally, the neuron paths have illustrated that vision Transformers exhibit
some specific inner working mechanism for processing the visual information wit
hin the same image category.
We further analyze the key effects of these neurons on the image classification
task, showcasing that the found neuron paths have already preserved the model ca
pability on downstream tasks, which may also shed some lights on real-world appl
ications like model pruning.
The project website including implementation code is available at https://founda
tion-model-research.github.io/NeuronPath/.
****
Rundong He,Yicong Dong,Lan-Zhe Guo,Yilong Yin,Tailin Wu
Re-Evaluating the Impact of Unseen-Class Unlabeled Data on Semi-Supervised Learn
ing Model
Semi-supervised learning (SSL) effectively leverages unlabeled data and has been
proven successful across various fields. Current safe SSL methods believe that
unseen classes in unlabeled data harm the performance of SSL models. However, pr
evious methods for assessing the impact of unseen classes on SSL model performan
ce are flawed. They fix the size of the unlabeled dataset and adjust the proport
ion of unseen classes within the unlabeled data to assess the impact. This proce
ss contravenes the principle of controlling variables. Adjusting the proportion
of unseen classes in unlabeled data alters the proportion of seen classes, meani
ng the decreased classification performance of seen classes may not be due to an

increase in unseen class samples in the unlabeled data, but rather a decrease in seen class samples. Thus, the prior flawed assessment standard that "unseen classes in unlabeled data can damage SSL model performance" may not always hold true. This paper strictly adheres to the principle of controlling variables, maintaining the proportion of seen classes in unlabeled data while only changing the unseen classes across five critical dimensions, to investigate their impact on SSL models from global robustness and local robustness. Experiments demonstrate that unseen classes in unlabeled data do not necessarily impair the performance of SSL models; in fact, under certain conditions, unseen classes may even enhance them.

****

Xin Gu,Yaojie Shen,Chenxi Luo,Tiejian Luo,Yan Huang,Yuewei Lin,Heng Fan,Libo Zhang

Knowing Your Target: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding

Transformer has attracted increasing interest in spatio-temporal video grounding, or STVG, owing to its end-to-end pipeline and promising result. Existing Transformer-based STVG approaches often leverage a set of object queries, which are initialized simply using zeros and then gradually learn target position information via iterative interactions with multimodal features, for spatial and temporal localization. Despite simplicity, these zero object queries, due to lacking target-specific cues, are hard to learn discriminative target information from interactions with multimodal features in complicated scenarios (e.g., with distractors or occlusion), resulting in degradation. Addressing this, we introduce a novel $\textbf{T}$arget-$\textbf{A}$ware Transformer for $\textbf{STVG}$ ($\textbf{TA-STVG}$), which seeks to adaptively generate object queries via exploring target-specific cues from the given video-text pair, for improving STVG. The key lies in two simple yet effective modules, comprising text-guided temporal sampling (TTS) and attribute-aware spatial activation (ASA), working in a cascade. The former focuses on selecting target-relevant temporal cues from a video utilizing holistic text information, while the latter aims at further exploiting the fine-grained visual attribute information of the object from previous target-aware temporal cues, which is applied for object query initialization. Compared to existing methods leveraging zero-initialized queries, object queries in our TA-STVG, directly generated from a given video-text pair, naturally carry target-specific cues, making them adaptive and better interact with multimodal features for learning more discriminative information to improve STVG. In our experiments on three benchmarks, including HCSTVG-v1/-v2 and VidSTG, TA-STVG achieves state-of-the-art performance and significantly outperforms the baseline, validating its efficacy. Moreover, TTS and ASA are designed for general purpose. When applied to existing methods such as TubeDETR and STCAT, we show substantial performance gains, verifying its generality. Code is released at https://github.com/HengLan/TA-STVG.

****

Qi Le,Enmao Diao,Ziyan Wang,Xinran Wang,Jie Ding,Li Yang,Ali Anwar

Probe Pruning: Accelerating LLMs through Dynamic Pruning via Model-Probing

We introduce Probe Pruning (PP), a novel framework for online, dynamic, structured pruning of Large Language Models (LLMs) applied in a batch-wise manner. PP leverages the insight that not all samples and tokens contribute equally to the model's output, and probing a small portion of each batch effectively identifies crucial weights, enabling tailored dynamic pruning for different batches. It comprises three main stages: probing, history-informed pruning, and full inference. In the probing stage, PP selects a small yet crucial set of hidden states, based on residual importance, to run a few model layers ahead. During the history-informed pruning stage, PP strategically integrates the probing states with historical states. Subsequently, it structurally prunes weights based on the integrated states and the PP importance score, a metric developed specifically to assess the importance of each weight channel in maintaining performance. In the final stage, full inference is conducted on the remaining weights. A major advantage of PP is its compatibility with existing models, as it operates without requiring a

dditional neural network modules or fine-tuning. Comprehensive evaluations of PP on LLaMA-2/3 and OPT models reveal that even minimal probing—using just 1.5% of FLOPs—can substantially enhance the efficiency of structured pruning of LLMs. For instance, when evaluated on LLaMA-2-7B with WikiText2, PP achieves a 2.56 times lower ratio of performance degradation per unit of latency reduction compared to the state-of-the-art method at a 40\% pruning ratio.
****
Shen Nie,Fengqi Zhu,Chao Du,Tianyu Pang,Qian Liu,Guangtao Zeng,Min Lin,Chongxuan Li
Scaling up Masked Diffusion Models on Text
Masked diffusion models (MDMs) have shown promise in language modeling, yet their scalability and effectiveness in core language tasks, such as text generation and language understanding, remain underexplored. This paper establishes the first scaling law for MDMs, demonstrating a scaling rate comparable to autoregressive models (ARMs) and a relatively small compute gap. Motivated by their scalability, we train a family of MDMs with up to 1.1 billion (B) parameters to systematically evaluate their performance against ARMs of comparable or larger sizes. Fully leveraging the probabilistic formulation of MDMs, we propose a simple yet effective *unsupervised classifier-free guidance* that effectively exploits large-scale unpaired data, boosting performance for conditional inference. In language understanding, the 1.1B MDM outperforms the 1.1B TinyLlama model trained on the same data across four of eight zero-shot benchmarks. Notably, it achieves competitive math reasoning ability with the 7B Llama-2 model on the GSM8K dataset.
In text generation, MDMs with 16 times more pre-training time offer a flexible trade-off against ARMs with the accelerated sampling technique KV-Cache: MDMs match ARMs in performance while being 1.4 times faster during sampling.
Moreover, MDMs address challenging tasks for ARMs by effectively handling bidirectional reasoning and adapting to temporal shifts in data. Notably, a 1.1B MDM breaks the *reverse curse* encountered by much larger ARMs with significantly more data and computation, such as 13B Llama-2 and 175B GPT-3. Our code is available at https://github.com/ML-GSAI/SMDM.
****
Taewon Kim,Hyunjin Seo,Sungsoo Ahn,Eunho Yang
REBIND: Enhancing Ground-state Molecular Conformation Prediction via Force-Based Graph Rewiring
Predicting the ground-state 3D molecular conformations from 2D molecular graphs is critical in computational chemistry due to its profound impact on molecular properties. Deep learning (DL) approaches have recently emerged as promising alternatives to computationally-heavy classical methods such as density functional theory (DFT). However, we discover that existing DL methods inadequately model inter-atomic forces, particularly for non-bonded atomic pairs, due to their naive usage of bonds and pairwise distances. Consequently, significant prediction errors occur for atoms with low degree (ie., low coordination numbers) whose conformations are primarily influenced by non-bonded interactions. To address this, we propose ReBIND, a novel framework that rewires molecular graphs by adding edges based on the Lennard-Jones potential to capture non-bonded interactions for low-degree atoms. Experimental results demonstrate that ReBIND significantly outperforms state-of-the-art methods across various molecular sizes, achieving up to a 20% reduction in prediction error. The code is available in: https://github.com/holymollyhao/ReBIND
****
Yi-Chen Li,Fuxiang Zhang,Wenjie Qiu,Lei Yuan,Chengxing Jia,Zongzhang Zhang,Yang Yu,Bo An
Q-Adapter: Customizing Pre-trained LLMs to New Preferences with Forgetting Mitigation
Large Language Models (LLMs), trained on a large amount of corpus, have demonstrated remarkable abilities. However, it may not be sufficient to directly apply open-source LLMs like Llama to certain real-world scenarios, since most of them are trained for \emph{general} purposes. Thus, the demands for customizing publicly available LLMs emerge, but are currently under-studied. In this work, we cons

ider customizing pre-trained LLMs with new human preferences. Specifically, the LLM should not only meet the new preference but also preserve its original capabilities after customization. Drawing inspiration from the observation that human preference can be expressed as a reward model, we propose to cast LLM customization as optimizing the sum of two reward functions, one of which (denoted as r_1) was used to pre-train the LLM while the other (denoted as r_2) characterizes the new human preference. The obstacle here is that both reward functions are unknown, making the application of modern reinforcement learning methods infeasible. Thanks to the residual Q-learning framework, we can restore the customized LLM with the pre-trained LLM and the \emph{residual Q-function} without the reward function r_1. Moreover, we find that for a fixed pre-trained LLM, the reward function r_2 can be derived from the residual Q-function, enabling us to directly learn the residual Q-function from the new human preference data upon the Bradley-Terry model. We name our method Q-Adapter as it introduces an adapter module to approximate the residual Q-function for customizing the pre-trained LLM towards the new preference. Experiments based on the Llama-3.1 model on the DSP dataset and HH-RLHF dataset illustrate the superior effectiveness of Q-Adapter on both retaining existing knowledge and learning new preferences. Our code is available at \url{https://github.com/LAMDA-RL/Q-Adapter}.
****
Chengan He,Xin Sun,Zhixin Shu,Fujun Luan,Soren Pirk,Jorge Alejandro Amador Herrera,Dominik Michels,Tuanfeng Yang Wang,Meng Zhang,Holly Rushmeier,Yi Zhou
Perm: A Parametric Representation for Multi-Style 3D Hair Modeling
We present Perm, a learned parametric representation of human 3D hair designed to facilitate various hair-related applications. Unlike previous work that jointly models the global hair structure and local curl patterns, we propose to disentangle them using a PCA-based strand representation in the frequency domain, thereby allowing more precise editing and output control. Specifically, we leverage our strand representation to fit and decompose hair geometry textures into low- to high-frequency hair structures, termed guide textures and residual textures, respectively. These decomposed textures are later parameterized with different generative models, emulating common stages in the hair grooming process. We conduct extensive experiments to validate the architecture design of Perm, and finally deploy the trained model as a generic prior to solve task-agnostic problems, further showcasing its flexibility and superiority in tasks such as single-view hair reconstruction, hairstyle editing, and hair-conditioned image generation. More details can be found on our project page: https://cs.yale.edu/homes/che/projects/perm/.
****
Ali Ebrahimpour-Boroojeny,Hari Sundaram,Varun Chandrasekaran
Training Robust Ensembles Requires Rethinking Lipschitz Continuity
Transferability of adversarial examples is a well-known property that endangers all classification models, even those that are only accessible through black-box queries. Prior work has shown that an ensemble of models is more resilient to transferability: the probability that an adversarial example is effective against most models of the ensemble is low. Thus, most ongoing research focuses on improving ensemble diversity. Another line of prior work has shown that Lipschitz continuity of the models can make models more robust since it limits how a model's output changes with small input perturbations. {\em In this paper, we study the effect of Lipschitz continuity on transferability rates.} We show that although a lower Lipschitz constant increases the robustness of a single model, it is not as beneficial in training robust ensembles as it increases the transferability rate of adversarial examples across models in the ensemble. Therefore, we introduce LOTOS, a new training paradigm for ensembles, which counteracts this adverse effect. It does so by promoting orthogonality among the top-k sub-spaces of the transformations of the corresponding affine layers of any pair of models in the ensemble. We theoretically show that k does not need to be large for convolutional layers, which makes the computational overhead negligible. Through various experiments, we show LOTOS increases the robust accuracy of ensembles of ResNet-18 models by 6 percentage points (p.p) against black-box attacks on CIFAR-

10. It is also capable of combining with the robustness of prior state-of-the-art methods for training robust ensembles to enhance their robust accuracy by 10.7 p.p.
****
Santhosh Kumar Ramakrishnan,Erik Wijmans,Philipp Kraehenbuehl,Vladlen Koltun
Does Spatial Cognition Emerge in Frontier Models?
Not yet. We present SPACE, a benchmark that systematically evaluates spatial cognition in frontier models. Our benchmark builds on decades of research in cognitive science. It evaluates large-scale mapping abilities that are brought to bear when an organism traverses physical environments, smaller-scale reasoning about object shapes and layouts, and cognitive infrastructure such as spatial attention and memory. For many tasks, we instantiate parallel presentations via text and images, allowing us to benchmark both large language models and large multimodal models. Results suggest that contemporary frontier models fall short of the spatial intelligence of animals, performing near chance level on a number of classic tests of animal cognition.
****
Audrey Huang,Adam Block,Dylan J Foster,Dhruv Rohatgi,Cyril Zhang,Max Simchowitz,Jordan T. Ash,Akshay Krishnamurthy
Self-Improvement in Language Models: The Sharpening Mechanism
Recent work in language modeling has raised the possibility of “self-improvement,” where an LLM evaluates and refines its own generations to achieve higher performance without external feedback. It is impossible for this self-improvement to create information that is not already in the model, so why should we expect that this will lead to improved capabilities? We offer a new theoretical perspective on the capabilities of self-improvement through a lens we refer to as “sharpening.” Motivated by the observation that language models are often better at verifying response quality than they are at generating correct responses, we formalize self-improvement as using the model itself as a verifier during post-training in order to ‘sharpen’ the model to one placing large mass on high-quality sequences, thereby amortizing the expensive inference-time computation of generating good sequences. We begin by introducing a new statistical framework for sharpening in which the learner has sample access to a pre-trained base policy. Then, we analyze two natural families of self improvement algorithms based on SFT and RLHF. We find that (i) the SFT-based approach is minimax optimal whenever the initial model has sufficient coverage, but (ii) the RLHF-based approach can improve over SFT-based self- improvement by leveraging online exploration, bypassing the need for coverage. We view these findings as a starting point toward a foundational understanding that can guide the design and evaluation of self-improvement algorithms.
****
Alireza Mousavi-Hosseini,Denny Wu,Murat A Erdogdu
Learning Multi-Index Models with Neural Networks via Mean-Field Langevin Dynamics
We study the problem of learning multi-index models in high-dimensions using a two-layer neural network trained with the mean-field Langevin algorithm. Under mild distributional assumptions on the data, we characterize the effective dimension d_{eff} that controls both sample and computational complexity by utilizing the adaptivity of neural networks to latent low-dimensional structures. When the data exhibit such a structure, d_{eff} can be significantly smaller than the ambient dimension. We prove that the sample complexity grows almost linearly with d_{eff}, bypassing the limitations of the information and generative exponents that appeared in recent analyses of gradient-based feature learning. On the other hand, the computational complexity may inevitably grow exponentially with d_{eff} in the worst-case scenario. Motivated by improving computational complexity, we take the first steps towards polynomial time convergence of the mean-field Langevin algorithm by investigating a setting where the weights are constrained to be on a compact manifold with positive Ricci curvature, such as the hypersphere. There, we study assumptions under which polynomial time convergence is achievable, whereas similar assumptions in t

he Euclidean setting lead to exponential time complexity.
****
Ke Yi,Zengke Liu,jianwei zhang,Chengyuan Li,Tong Zhang,Junyang Lin,Jingren Zhou
Rotated Runtime Smooth: Training-Free Activation Smoother for accurate INT4 infe
rence
Large language models have demonstrated promising capabilities upon scaling up p
arameters. However, serving large language models incurs substantial computation
 and memory movement costs due to their large scale. Quantization methods have b
een employed to reduce service costs and latency. Nevertheless, outliers in acti
vations hinder the development of INT4 weight-activation quantization. Existing
approaches separate outliers and normal values into two matrices or migrate outl
iers from activations to weights, suffering from high latency or accuracy degrad
ation. Based on observing activations from large language models, outliers can b
e classified into channel-wise and spike outliers.
In this work, we propose Rotated Runtime Smooth (**RRS**), a plug-and-play activ
ation smoother for quantization, consisting of Runtime Smooth and the Rotation o
peration. Runtime Smooth (**RS**) is introduced to eliminate **channel-wise outl
iers** by smoothing activations with channel-wise maximums during runtime. The R
otation operation can narrow the gap between **spike outliers** and normal value
s, alleviating the effect of victims caused by channel-wise smoothing.
The proposed method outperforms the state-of-the-art method in the LLaMA and Qwe
n families and improves WikiText-2 perplexity from 57.33 to 6.66 for INT4 infere
nce.
****
Sungnyun Kim,Sungwoo Cho,Sangmin Bae,Kangwook Jang,Se-Young Yun
Multi-Task Corrupted Prediction for Learning Robust Audio-Visual Speech Represen
tation
Audio-visual speech recognition (AVSR) incorporates auditory and visual modaliti
es to improve recognition accuracy, particularly in noisy environments where aud
io-only speech systems are insufficient. While previous research has largely add
ressed audio disruptions, few studies have dealt with visual corruptions, e.g.,
lip occlusions or blurred videos, which are also detrimental. To address this re
al-world challenge, we propose CAV2vec, a novel self-supervised speech represent
ation learning framework particularly designed to handle audio-visual joint corr
uption. CAV2vec employs a self-distillation approach with a corrupted prediction
 task, where the student model learns to predict clean targets, generated by the
 teacher model, with corrupted input frames. Specifically, we suggest a unimodal
 multi-task learning, which distills cross-modal knowledge and aligns the corrup
ted modalities, by predicting clean audio targets with corrupted videos, and cle
an video targets with corrupted audios. This strategy mitigates the dispersion i
n the representation space caused by corrupted modalities, leading to more relia
ble and robust audio-visual fusion. Our experiments on robust AVSR benchmarks de
monstrate that the corrupted representation learning method significantly enhanc
es recognition accuracy across generalized environments involving various types
of corruption. Our code is available at https://github.com/sungnyun/cav2vec.
****
Lei Chen,Joan Bruna,Alberto Bietti
Distributional Associations vs In-Context Reasoning: A Study of Feed-forward and
 Attention Layers
Large language models have been successful at tasks involving basic forms of in-
context reasoning, such as generating coherent language, as well as storing vast
 amounts of knowledge. At the core of the Transformer architecture behind such m
odels are feed-forward and attention layers, which are often associated to knowl
edge and reasoning, respectively. In this paper, we study this distinction empir
ically and theoretically in a controlled synthetic setting where certain next-to
ken predictions involve both distributional and in-context information. We find
that feed-forward layers tend to learn simple distributional associations such a
s bigrams, while attention layers focus on in-context reasoning. Our theoretical
 analysis identifies the noise in the gradients as a key factor behind this disc
repancy. Finally, we illustrate how similar disparities emerge in pre-trained mo

dels through ablations on the Pythia model family on simple reasoning tasks.
****
Javier Ferrando,Oscar Balcells Obeso,Senthooran Rajamanoharan,Neel Nanda
Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models
Hallucinations in large language models are a widespread problem, yet the mechan isms behind whether models will hallucinate are poorly understood, limiting our ability to solve this problem. Using sparse autoencoders as an interpretability tool, we discover that a key part of these mechanisms is entity recognition, whe re the model detects if an entity is one it can recall facts about. Sparse autoe ncoders uncover meaningful directions in the representation space, these detect whether the model recognizes an entity, e.g. detecting it doesn't know about an athlete or a movie. This shows that models can have self-knowledge: internal rep resentations about their own capabilities. These directions are causally relevan t: capable of steering the model to refuse to answer questions about known entit ies, or to hallucinate attributes of unknown entities when it would otherwise re fuse. We demonstrate that despite the sparse autoencoders being trained on the b ase model, these directions have a causal effect on the chat model's refusal beh avior, suggesting that chat finetuning has repurposed this existing mechanism. F urthermore, we provide an initial exploration into the mechanistic role of these directions in the model, finding that they disrupt the attention of downstream heads that typically move entity attributes to the final token.
****
Mehdi Naouar,Hanne Raum,Jens Rahnfeld,Yannick Vogt,Joschka Boedecker,Gabriel Kal weit,Maria Kalweit
Salvage: Shapley-distribution Approximation Learning Via Attribution Guided Expl oration for Explainable Image Classification
The integration of deep learning into critical vision application areas has give n rise to a necessity for techniques that can explain the rationale behind predi ctions. In this paper, we address this need by introducing Salvage, a novel remo val-based explainability method for image classification. Our approach involves training an explainer model that learns the prediction distribution of the class ifier on masked images. We first introduce the concept of Shapley-distributions, which offers a more accurate approximation of classification probability distri butions than existing methods. Furthermore, we address the issue of unbalanced i mportant and unimportant features. In such settings, naive uniform sampling of f eature subsets often results in a highly unbalanced ratio of samples with high a nd low prediction likelihoods, which can hinder effective learning. To mitigate this, we propose an informed sampling strategy that leverages approximated featu re importance scores, thereby reducing imbalance and facilitating the estimation of underrepresented features. After incorporating these two principles into our method, we conducted an extensive analysis on the ImageNette, MURA, WBC, and Pe t datasets. The results show that Salvage outperforms various baseline explainab ility methods, including attention-, gradient-, and removal-based approaches, bo th qualitatively and quantitatively. Furthermore, we demonstrate that our explai ner model can serve as a fully explainable classifier without a major decrease i n classification performance, paving the way for fully explainable image classif ication.
****
Haohan Weng,Yikai Wang,Tong Zhang,C. L. Philip Chen,Jun Zhu
PivotMesh: Generic 3D Mesh Generation via Pivot Vertices Guidance
Generating compact and sharply detailed 3D meshes poses a significant challenge for current 3D generative models. Different from extracting dense meshes from ne ural representation, some recent works try to model the native mesh distribution (i.e., a set of triangles), which generates more compact results as humans craf ted. However, due to the complexity and variety of mesh topology, most of these methods are typically limited to generating meshes with simple geometry. In this paper, we introduce a generic and scalable mesh generation framework PivotMesh, which makes an initial attempt to extend the native mesh generation to large-sc ale datasets. We employ a transformer-based autoencoder to encode meshes into di screte tokens and decode them from face level to vertex level hierarchically. Su

bsequently, to model the complex typology, our model first learns to generate pi
vot vertices as coarse mesh representation and then generate the complete mesh t
okens with the same auto-regressive Transformer. This reduces the difficulty com
pared with directly modeling the mesh distribution and further improves the mode
l controllability. PivotMesh demonstrates its versatility by effectively learnin
g from both small datasets like Shapenet, and large-scale datasets like Objavers
e and Objaverse-xl. Extensive experiments indicate that PivotMesh can generate c
ompact and sharp 3D meshes across various categories, highlighting its great pot
ential for native mesh modeling.
****

Yizhuo Ding,Ke Fan,Yikai Wang,Xinwei Sun,Yanwei Fu
Adaptive Pruning of Pretrained Transformer via Differential Inclusions
Large transformers have demonstrated remarkable success, making it necessary to
compress these models to reduce inference costs while preserving their performan
ce. Current compression algorithms prune transformers at fixed compression ratio
s, requiring a unique pruning process for each ratio, which results in high comp
utational costs. In contrast, we propose pruning of pretrained transformers at a
ny desired ratio within a single pruning stage, based on a differential inclusio
n for a mask parameter. This dynamic can generate the whole regularization solut
ion path of the mask parameter, whose support set identifies the network structu
re. Therefore, the solution path identifies a Transformer weight family with var
ious sparsity levels, offering greater flexibility and customization.In this pap
er, weintroduce such an effective pruning method, termed SPP (Solution Path Prun
ing). To achieve effective pruning, we segment the transformers into paired modu
les, including query-key pairs, value-projection pairs, and sequential linear la
yers, and apply low-rank compression to these pairs, maintaining the output stru
cture while enabling structural compression within the inner states. Extensive e
xperiments conducted on various well-known transformer backbones have demonstrat
ed the efficacy of SPP.
****

Afra Amini,Tim Vieira,Elliott Ash,Ryan Cotterell
Variational Best-of-N Alignment
Best-of-N (BoN) is a popular and effective algorithm for aligning language model
s to human preferences. The algorithm works as follows: at inference time, N sam
ples are drawn from the language model, and the sample with the highest reward,
as judged by a reward model, is returned as the output. Despite its effectivenes
s, BoN is computationally expensive; it reduces sampling throughput by a factor
of N.
To make BoN more efficient at inference time, one strategy is to fine-tune the l
anguage model to mimic what BoN does during inference.
To achieve this, we derive the distribution induced by the BoN algorithm. We the
n propose to fine-tune the language model to minimize backward KL divergence to
the BoN distribution. Our approach is analogous to mean-field variational infere
nce and, thus, we term it variational BoN (vBoN). To the extent this fine-tuning
 is successful and we end up with a good approximation, we have reduced the infe
rence cost by a factor of N. Our experiments on controlled generation and summar
ization tasks show that BoN is the most effective alignment method, and our vari
ational approximation to BoN achieves the closest performance to BoN and surpass
es models fine-tuned using the standard KL-constrained RL objective. In the cont
rolled generation task, vBoN appears more frequently on the Pareto frontier of r
eward and KL divergence compared to other alignment methods. In the summarizatio
n task, vBoN achieves high reward values across various sampling temperatures.
****

Xianyuan Zhan,Xiangyu Zhu,Peng Cheng,Xiao Hu,Ziteng He,Hanfei Geng,Jichao Leng,H
uiwen Zheng,Chenhui Liu,Tianshun Hong,Yan Liang,Yunxin Liu,Feng Zhao
Data Center Cooling System Optimization Using Offline Reinforcement Learning
The recent advances in information technology and artificial intelligence have f
ueled a rapid expansion of the data center (DC) industry worldwide, accompanied
by an immense appetite for electricity to power the DCs. In a typical DC, around
 30-40% of the energy is spent on the cooling system rather than on computer ser

vers, posing a pressing need for developing new energy-saving optimization techn
ologies for DC cooling systems. However, optimizing such real-world industrial s
ystems faces numerous challenges, including but not limited to a lack of reliabl
e simulation environments, limited historical data, and stringent safety and con
trol robustness requirements. In this work, we present a novel physics-informed
offline reinforcement learning (RL) framework for energy efficiency optimization
 of DC cooling systems. The proposed framework models the complex dynamical patt
erns and physical dependencies inside a server room using a purposely designed g
raph neural network architecture that is compliant with the fundamental time-rev
ersal symmetry. Because of its well-behaved and generalizable state-action repre
sentations, the model enables sample-efficient and robust latent space offline p
olicy learning using limited real-world operational data. Our framework has been
 successfully deployed and verified in a large-scale production DC for closed-lo
op control of its air-cooling units (ACUs). We conducted a total of 2000 hours o
f short and long-term experiments in the production DC environment. The results
show that our method achieves 14-21% energy savings in the DC cooling system, wi
thout any violation of the safety or operational constraints. We have also condu
cted a comprehensive evaluation of our approach in a real-world DC testbed envir
onment. Our results have demonstrated the significant potential of offline RL in
 solving a broad range of data-limited, safety-critical real-world industrial co
ntrol problems.
****

Quoc-Vinh Lai-Dang,Taemin Kang,Seungah Son

Adaptive Transformer Programs: Bridging the Gap Between Performance and Interpre
tability in Transformers

Balancing high performance with interpretability in increasingly powerful Transf
ormer-based models remains a challenge. While mechanistic interpretability aims
to specify neural network computations in explicit, pseudocode-like formats, exi
sting methods often involve laborious manual analysis or struggle to fully eluci
date learned internal algorithms. Recent efforts to build intrinsically interpre
table models have introduced considerable expressivity and optimization challeng
es. This work introduces Adaptive Transformer Programs, an enhanced framework bu
ilding upon RASP language and Transformer Programs to create more robust and int
erpretable models. The proposed method increases expressivity by redesigning two
 primary attention modules to improve categorical and numerical reasoning capabi
lities. To overcome optimization hurdles, we introduce a novel reparameterizatio
n scheme that enhances the exploration-exploitation trade-off during training. W
e validate our approach through extensive experiments on diverse tasks, includin
g in-context learning, algorithmic problems (e.g., sorting and Dyck languages),
and NLP benchmarks such as named entity recognition and text classification. Res
ults demonstrate that Adaptive Transformer Programs substantially narrow the per
formance gap between black-box Transformers and interpretable models, enhancing
transparency. This work advances the development of high-performing, transparent
 AI systems for critical applications, addressing crucial ethical concerns in AI
 development.
****

Shaotian Yan,Chen Shen,Wenxiao Wang,Liang Xie,Junjie Liu,Jieping Ye

Don't Take Things Out of Context: Attention Intervention for Enhancing Chain-of-
Thought Reasoning in Large Language Models

Few-shot Chain-of-Thought (CoT) significantly enhances the reasoning capabilitie
s of large language models (LLMs), functioning as a whole to guide these models
in generating reasoning steps toward final answers. However, we observe that iso
lated segments, words, or tokens within CoT demonstrations can unexpectedly disr
upt the generation process of LLMs. The model may overly concentrate on certain
local information present in the demonstration, introducing irrelevant noise int
o the reasoning process and potentially leading to incorrect answers. In this pa
per, we investigate the underlying mechanism of CoT through dynamically tracing
and manipulating the inner workings of LLMs at each output step, which demonstra
tes that tokens exhibiting specific attention characteristics are more likely to
 induce the model to take things out of context; these tokens directly attend to

the hidden states tied with prediction, without substantial integration of non-local information. Building upon these insights, we propose a Few-shot Attention Intervention method (FAI) that dynamically analyzes the attention patterns of demonstrations to accurately identify these tokens and subsequently make targeted adjustments to the attention weights to effectively suppress their distracting effect on LLMs. Comprehensive experiments across multiple benchmarks demonstrate consistent improvements over baseline methods, with a remarkable 5.91\% improvement on the AQuA dataset, further highlighting the effectiveness of FAI.
****
Zhengyao Lv,Chenyang Si,Junhao Song,Zhenyu Yang,Yu Qiao,Ziwei Liu,Kwan-Yee K. Wong
FasterCache: Training-Free Video Diffusion Model Acceleration with High Quality
In this paper, we present \textbf{\textit{FasterCache}}, a novel training-free strategy designed to accelerate the inference of video diffusion models with high-quality generation. By analyzing existing cache-based methods, we observe that \textit{directly reusing adjacent-step features degrades video quality due to the loss of subtle variations}. We further perform a pioneering investigation of the acceleration potential of classifier-free guidance (CFG) and reveal significant redundancy between conditional and unconditional features within the same timestep. Capitalizing on these observations, we introduce FasterCache to substantially accelerate diffusion-based video generation. Our key contributions include a dynamic feature reuse strategy that preserves both feature distinction and temporal continuity, and CFG-Cache which optimizes the reuse of conditional and unconditional outputs to further enhance inference speed without compromising video quality. We empirically evaluate FasterCache on recent video diffusion models. Experimental results show that FasterCache can significantly accelerate video generation (\eg 1.67$\times$ speedup on Vchitect-2.0) while keeping video quality comparable to the baseline, and consistently outperform existing methods in both inference speed and video quality. \textit{Our code will be made public upon publication.}
****
Yansen Zhang,Qingcan Kang,Wing Yin YU,HaileiGong,Xiaojin Fu,Xiongwei Han,Tao Zhong,Chen Ma
Decision Information Meets Large Language Models: The Future of Explainable Operations Research
Operations Research (OR) is vital for decision-making in many industries. While recent OR methods have seen significant improvements in automation and efficiency through integrating Large Language Models (LLMs), they still struggle to produce meaningful explanations. This lack of clarity raises concerns about transparency and trustworthiness in OR applications. To address these challenges, we propose a comprehensive framework, Explainable Operations Research (EOR), emphasizing actionable and understandable explanations accompanying optimization. The core of EOR is the concept of Decision Information, which emerges from what-if analysis and focuses on evaluating the impact of complex constraints (or parameters) changes on decision-making. Specifically, we utilize bipartite graphs to quantify the changes in the OR model and adopt LLMs to improve the explanation capabilities. Additionally, we introduce the first industrial benchmark to rigorously evaluate the effectiveness of explanations and analyses in OR, establishing a new standard for transparency and clarity in the field.
****
Dong Wang,Haris Šiki■,Lothar Thiele,Olga Saukh
Forget the Data and Fine-Tuning! Just Fold the Network to Compress
We introduce model folding, a novel data-free model compression technique that merges structurally similar neurons across layers, significantly reducing the model size without the need for fine-tuning or access to training data. Unlike existing methods, model folding preserves data statistics during compression by leveraging k-means clustering, and using novel data-free techniques to prevent variance collapse or explosion. Our theoretical framework and experiments across standard benchmarks, including ResNet18 and LLaMA-7B, demonstrate that model folding achieves comparable performance to data-driven compression techniques and outpe

rforms recently proposed data-free methods, especially at high sparsity levels. This approach is particularly effective for compressing large-scale models, maki ng it suitable for deployment in resource-constrained environments.
****

Shahrzad Kiani,Nupur Kulkarni,Adam Dziedzic,Stark Draper,Franziska Boenisch
Differentially Private Federated Learning with Time-Adaptive Privacy Spending
Federated learning (FL) with differential privacy (DP) provides a framework for collaborative machine learning, enabling clients to train a shared model while a dhering to strict privacy constraints. The framework allows each client to have an individual privacy guarantee, e.g., by adding different amounts of noise to e ach client's model updates. One underlying assumption is that all clients spend their privacy budgets uniformly over time (learning rounds). However, it has bee n shown in the literature that learning in early rounds typically focuses on mor e coarse-grained features that can be learned at lower signal-to-noise ratios wh ile later rounds learn fine-grained features that benefit from higher signal-to-noise ratios. Building on this intuition, we propose a time-adaptive DP-FL frame work that expends the privacy budget non-uniformly across both time and clients. Our framework enables each client to save privacy budget in early rounds so as to be able to spend more in later rounds when additional accuracy is beneficial in learning more fine-grained features. We theoretically prove utility improveme nts in the case that clients with stricter privacy budgets spend budgets unevenl y across rounds, compared to clients with more relaxed budgets, who have suffici ent budgets to distribute their spend more evenly. Our practical experiments on standard benchmark datasets support our theoretical results and show that, in pr actice, our algorithms improve the privacy-utility trade-offs compared to baseli ne schemes.
****

Yizi Zhang,Jingyan Shen,Xiaoxue Xiong,Yongchan Kwon
TimeInf: Time Series Data Contribution via Influence Functions
Evaluating the contribution of individual data points to a model's prediction is critical for interpreting model predictions and improving model performance. Ex isting data contribution methods have been applied to various data types, includ ing tabular data, images, and text; however, their primary focus has been on i.i .d. settings. Despite the pressing need for principled approaches tailored to ti me series datasets, the problem of estimating data contribution in such settings remains under-explored, possibly due to challenges associated with handling inh erent temporal dependencies. This paper introduces TimeInf, a model-agnostic dat a contribution estimation method for time-series datasets. By leveraging influen ce scores, TimeInf attributes model predictions to individual time points while preserving temporal structures between the time points. Our empirical results sh ow that TimeInf effectively detects time series anomalies and outperforms existi ng data attribution techniques as well as state-of-the-art anomaly detection met hods. Moreover, TimeInf offers interpretable attributions of data values, allowi ng us to distinguish diverse anomalous patterns through visualizations. We also showcase a potential application of TimeInf in identifying mislabeled anomalies in the ground truth annotations.
****

Longrong Yang,Dong Shen,Chaoxiang Cai,Fan Yang,Tingting Gao,Di ZHANG,Xi Li
Solving Token Gradient Conflict in Mixture-of-Experts for Large Vision-Language Model
The Mixture-of-Experts (MoE) has gained increasing attention in studying Large V ision-Language Models (LVLMs). It uses a sparse model to replace the dense model , achieving comparable performance while activating fewer parameters during infe rence, thus significantly reducing the inference cost. Existing MoE methods in L VLM encourage different experts to specialize in different tokens, and they usua lly employ a router to predict the routing of each token. However, the router is not optimized concerning distinct parameter optimization directions generated f rom tokens within an expert. This may lead to severe interference between tokens within an expert. To address this problem, we propose to use the token-level gr adient analysis to Solving Token Gradient Conflict (STGC) in this paper. Specifi

cally, we first use token-level gradients to identify conflicting tokens in expe
rts. After that, we add a regularization loss tailored to encourage conflicting
tokens routing from their current experts to other experts, for reducing interfe
rence between tokens within an expert. Our method can serve as a plug-in for div
erse LVLM methods, and extensive experimental results demonstrate its effectiven
ess. demonstrate its effectiveness.
The code will be publicly available at https://github.com/longrongyang/STGC.
****
Xiaoling Zhou,Mingjie Zhang,Zhemg Lee,Wei Ye,Shikun Zhang
HaDeMiF: Hallucination Detection and Mitigation in Large Language Models
The phenomenon of knowledge hallucinations has raised substantial concerns about
 the security and reliability of deployed large language models (LLMs). Current
methods for detecting hallucinations primarily depend on manually designed indiv
idual metrics, such as prediction uncertainty and consistency, and fall short in
 effectively calibrating model predictions, thus constraining their detection ac
curacy and applicability in practical applications. In response, we propose an a
dvanced framework, termed HaDeMiF, for detecting and mitigating hallucinations i
n LLMs. Specifically, hallucinations within the output and semantic spaces of LL
Ms are comprehensively captured through two compact networks—a novel, interpreta
ble tree model known as the Deep Dynamic Decision Tree (D3T) and a Multilayer Pe
rceptron (MLP)—which take as input a set of prediction characteristics and the h
idden states of tokens, respectively. The predictions of LLMs are subsequently c
alibrated using the outputs from the D3T and MLP networks, aiming to mitigate ha
llucinations and enhance model calibration. HaDeMiF can be applied during both t
he inference and fine-tuning phases of LLMs, introducing less than 2% of the par
ameters relative to the LLMs through the training of two small-scale networks. E
xtensive experiments conclusively demonstrate the effectiveness of our framework
 in hallucination detection and model calibration across text generation tasks w
ith responses of varying lengths.
****
Yangning Li,Yinghui Li,Xinyu Wang,Yong Jiang,Zhen Zhang,Xinran Zheng,Hui Wang,Ha
i-Tao Zheng,Fei Huang,Jingren Zhou,Philip S. Yu
Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset
and Self-adaptive Planning Agent
Multimodal Retrieval Augmented Generation (mRAG) plays an important role in miti
gating the “hallucination” issue inherent in multimodal large language models (M
LLMs). Although promising, existing heuristic mRAGs typically predefined fixed r
etrieval processes, which causes two issues: (1) Non-adaptive Retrieval Queries.
 (2) Overloaded Retrieval Queries. However, these flaws cannot be adequately ref
lected by current knowledge-seeking visual question answering (VQA) datasets, si
nce the most required knowledge can be readily obtained with a standard two-step
 retrieval. To bridge the dataset gap, we first construct Dyn-VQA dataset, consi
sting of three types of ``dynamic'' questions, which require complex knowledge r
etrieval strategies variable in query, tool, and time: (1) Questions with rapidl
y changing answers. (2) Questions requiring multi-modal knowledge. (3) Multi-hop
 questions. Experiments on Dyn-VQA reveal that existing heuristic mRAGs struggle
 to provide sufficient and precisely relevant knowledge for dynamic questions du
e to their rigid retrieval processes. Hence, we further propose the first self-a
daptive planning agent for multimodal retrieval, **OmniSearch**. The underlying
idea is to emulate the human behavior in question solution which dynamically dec
omposes complex multimodal questions into sub-question chains with retrieval act
ion. Extensive experiments prove the effectiveness of our OmniSearch, also provi
de direction for advancing mRAG. Code and dataset will be open-sourced.
****
Peidong Li,Dixiao Cui
Navigation-Guided Sparse Scene Representation for End-to-End Autonomous Driving
End-to-End Autonomous Driving (E2EAD) methods typically rely on supervised perce
ption tasks to extract explicit scene information (e.g., objects, maps). This re
liance necessitates expensive annotations and constrains deployment and data sca
lability in real-time applications. In this paper, we introduce SSR, a novel fra

mework that utilizes only 16 navigation-guided tokens as Sparse Scene Representa
tion, efficiently extracting crucial scene information for E2EAD. Our method eli
minates the need for human-designed supervised sub-tasks, allowing computational
 resources to concentrate on essential elements directly related to navigation i
ntent. We further introduce a temporal enhancement module, aligning predicted fu
ture scenes with actual future scenes through self-supervision. SSR achieves a 2
7.2\% relative reduction in L2 error and a 51.6\% decrease in collision rate to
UniAD in nuScenes, with a 10.9× faster inference speed and 13× faster training t
ime. Moreover, SSR outperforms VAD-Base with a 48.6-point improvement on driving
 score in CARLA's Town05 Long benchmark. This framework represents a significant
 leap in real-time autonomous driving systems and paves the way for future scala
ble deployment. Code is available at https://github.com/PeidongLi/SSR.
****

Xiaoqiang Wang,Bang Liu
OSCAR: Operating System Control via State-Aware Reasoning and Re-Planning
Large language models (LLMs) and large multimodal models (LMMs) have shown great
 potential in automating complex tasks like web browsing and gaming. However, th
eir ability to generalize across diverse applications remains limited, hindering
 broader utility. To address this challenge, we present OSCAR: Operating System
Control via state-Aware reasoning and Re-planning. OSCAR is a generalist agent d
esigned to autonomously navigate and interact with various desktop and mobile ap
plications through standardized controls, such as mouse and keyboard inputs, whi
le processing screen images to fulfill user commands. OSCAR translates human ins
tructions into executable Python code, enabling precise control over graphical u
ser interfaces (GUIs). To enhance stability and adaptability, OSCAR operates as
a state machine, equipped with error-handling mechanisms and dynamic task re-pla
nning, allowing it to efficiently adjust to real-time feedback and exceptions. W
e demonstrate OSCAR’s effectiveness through extensive experiments on diverse ben
chmarks across desktop and mobile platforms, where it transforms complex workflo
ws into simple natural language commands, significantly boosting user productivi
ty. Our code will be open-source upon publication.
****

Zewei Zhang,Huan Liu,Jun Chen,Xiangyu Xu
GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models
In this paper, we introduce GoodDrag, a novel approach to improve the stability
and image quality of drag editing. Unlike existing methods that struggle with ac
cumulated perturbations and often result in distortions, GoodDrag introduces an
AlDD framework that alternates between drag and denoising operations within the
diffusion process, effectively improving the fidelity of the result. We also pro
pose an information-preserving motion supervision operation that maintains the o
riginal features of the starting point for precise manipulation and artifact red
uction. In addition, we contribute to the benchmarking of drag editing by introd
ucing a new dataset, Drag100, and developing dedicated quality assessment metric
s, Dragging Accuracy Index and Gemini Score, utilizing Large Multimodal Models.
Extensive experiments demonstrate that the proposed GoodDrag compares favorably
against the state-of-the-art approaches both qualitatively and quantitatively. T
he source code and data are available at https://gooddrag.github.io.
****

Alberto Pepe,Mattia Montanari,Joan Lasenby
Fengbo: a Clifford Neural Operator pipeline for 3D PDEs in Computational Fluid D
ynamics
We introduce Fengbo, a pipeline entirely in Clifford Algebra to solve 3D partial
 differential equations (PDEs) specifically for computational fluid dynamics (CF
D). Fengbo is an architecture composed of only 3D convolutional and Fourier Neur
al Operator (FNO) layers, all working in 3D Clifford Algebra. It models the PDE
solution problem as an interpretable mapping from the geometry to the physics of
 the problem. Despite having just few layers, Fengbo achieves competitive accura
cy, superior to 5 out of 6 proposed models reported in \cite{li2024geometry} for
 the $\emph{ShapeNet Car}$ dataset, and it does so with only 42 million trainabl
e parameters, at a reduced computational complexity compared to graph-based meth

ods, and estimating jointly pressure \emph{and} velocity fields. In addition, th
e output of each layer in Fengbo can be clearly visualised as objects and physic
al quantities in 3D space, making it a whitebox model.
By leveraging Clifford Algebra and establishing a direct mapping from the geomet
ry to the physics of the PDEs, Fengbo provides an efficient, geometry- and physi
cs-aware approach to solving complex PDEs.
****
Jui-Nan Yen,Si Si,Zhao Meng,Felix Yu,Sai Surya Duvvuri,Inderjit S Dhillon,Cho-Ju
i Hsieh,Sanjiv Kumar
LoRA Done RITE: Robust Invariant Transformation Equilibration for LoRA Optimizat
ion
Low-rank adaption (LoRA) is a widely used parameter-efficient finetuning method
for LLM that reduces memory requirements. However, current LoRA optimizers lack
transformation invariance, meaning the updates depending on how the two LoRA fac
tors are scaled or rotated. This deficiency leads to inefficient learning and su
b-optimal solutions in practice. This paper introduces LoRA-RITE, a novel adapti
ve matrix preconditioning method for LoRA optimization, which can achieve transf
ormation invariance and remain computationally efficient. We provide theoretical
 analysis to demonstrate the benefit of our method and conduct experiments on va
rious LLM tasks with different models including Gemma 2B, 7B, and mT5-XXL. The r
esults demonstrate consistent improvements against existing optimizers. For exam
ple, replacing Adam with LoRA-RITE during LoRA fine-tuning of Gemma-2B yielded 4
.6% accuracy gain on Super-Natural Instructions and 3.5% accuracy gain across ot
her four LLM benchmarks (HellaSwag, ArcChallenge, GSM8K, OpenBookQA).
****
Manuel Brenner,Elias Weber,Georgia Koppe,Daniel Durstewitz
Learning Interpretable Hierarchical Dynamical Systems Models from Time Series Da
ta
In science, we are often interested in obtaining a generative model of the under
lying system dynamics from observed time series. While powerful methods for dyna
mical systems reconstruction (DSR) exist when data come from a single domain, ho
w to best integrate data from multiple dynamical regimes and leverage it for gen
eralization is still an open question. This becomes particularly important when
individual time series are short, and group-level information may help to fill i
n for gaps in single-domain data. Here we introduce a hierarchical framework tha
t enables to harvest group-level (multi-domain) information while retaining all
single-domain characteristics, and showcase it on popular DSR benchmarks, as wel
l as on neuroscience and medical data. In addition to faithful reconstruction of
 all individual dynamical regimes, our unsupervised methodology discovers common
 low-dimensional feature spaces in which datasets with similar dynamics cluster.
 The features spanning these spaces were further dynamically highly interpretabl
e, surprisingly in often linear relation to control parameters that govern the d
ynamics of the underlying system. Finally, we illustrate transfer learning and g
eneralization to new parameter regimes, paving the way toward DSR foundation mod
els.
****
Wonje Choi,Jinwoo Park,Sanghyun Ahn,Daehee Lee,Honguk Woo
NeSyC: A Neuro-symbolic Continual Learner For Complex Embodied Tasks in Open Dom
ains
We explore neuro-symbolic approaches to generalize actionable knowledge, enablin
g embodied agents to tackle complex tasks more effectively in open-domain enviro
nments. A key challenge for embodied agents is the generalization of knowledge a
cross diverse environments and situations, as limited experiences often confine
them to their prior knowledge. To address this issue, we introduce a novel frame
work, NeSyC, a neuro-symbolic continual learner that emulates the hypothetico-de
ductive model by continually formulating and validating knowledge from limited e
xperiences through the combined use of Large Language Models (LLMs) and symbolic
 tools. Specifically, we devise a contrastive generality improvement scheme with
in NeSyC, which iteratively generates hypotheses using LLMs and conducts contras
tive validation via symbolic tools. This scheme reinforces the justification for

admissible actions while minimizing the inference of inadmissible ones. Additio
nally, we incorporate a memory-based monitoring scheme that efficiently detects
action errors and triggers the knowledge refinement process across domains. Expe
riments conducted on diverse embodied task benchmarks—including ALFWorld, Virtua
lHome, Minecraft, RLBench, and a real-world robotic scenario—demonstrate that Ne
SyC is highly effective in solving complex embodied tasks across a range of open
-domain environments.
****

Arjun Subramonian,Samuel Bell,Levent Sagun,Elvis Dohmatob

An Effective Theory of Bias Amplification

Machine learning models can capture and amplify biases present in data, leading
to disparate test performance across social groups. To better understand, evalua
te, and mitigate these biases, a deeper theoretical understanding of how model d
esign choices and data distribution properties contribute to bias is needed. In
this work, we contribute a precise analytical theory in the context of ridge reg
ression, both with and without random projections, where the former models feedf
orward neural networks in a simplified regime. Our theory offers a unified and r
igorous explanation of machine learning bias, providing insights into phenomena
such as bias amplification and minority-group bias in various feature and parame
ter regimes. For example, we observe that there may be an optimal regularization
 penalty or training time to avoid bias amplification, and there can be differen
ces in test error between groups that are not alleviated with increased paramete
rization. Importantly, our theoretical predictions align with empirical observa
tions reported in the literature on machine learning bias. We extensively empiri
cally validate our theory on synthetic and semi-synthetic datasets.
****

Jaehyung Seo,Heuiseok Lim

K-HALU: Multiple Answer Korean Hallucination Benchmark for Large Language Models

Recent researchers and companies have been developing large language models (LLM
s) specifically designed for particular purposes and have achieved significant a
dvancements in various natural language processing tasks. However, LLMs are stil
l prone to generating hallucinations—results that are unfaithful or inconsistent
 with the given input. As a result, the need for datasets to evaluate and demons
trate the hallucination detection capabilities of LLMs is increasingly recognize
d. Nonetheless, the Korean NLP community lacks publicly available benchmark data
sets demonstrating the faithfulness of knowledge-based information. Furthermore,
 the few existing datasets that evaluate hallucination are limited in their acce
ss to the entire dataset, restricting detailed analysis beyond simple scoring, a
nd are based on translated English knowledge. To address these challenges, we in
troduce K-HALU, a Korean benchmark designed to evaluate LLMs' hallucination dete
ction in Korean. This benchmark contains seven domains, considering the faithful
ness of statements based on knowledge documents compiled from Korean news, magaz
ines, and books. For more strict evaluation, 40% of the dataset is structured as
 multiple-answer questions, requiring models to select all possible correct answ
ers from the given options. Our empirical results show that open-source LLMs sti
ll struggle with hallucination detection in Korean knowledge, emphasizing the ne
ed for a more detailed analysis of their limitations.
****

Amin Karimi Monsefi,Mengxi Zhou,Nastaran Karimi Monsefi,Ser-Nam Lim,Wei-Lun Chao
,Rajiv Ramnath

Frequency-Guided Masking for Enhanced Vision Self-Supervised Learning

We present a novel frequency-based Self-Supervised Learning (SSL) approach that
significantly enhances its efficacy for pre-training. Prior work in this directi
on masks out pre-defined frequencies in the input image and employs a reconstruc
tion loss to pre-train the model. While achieving promising results, such an imp
lementation has two fundamental limitations as identified in our paper. First, u
sing pre-defined frequencies overlooks the variability of image frequency respon
ses. Second, pre-trained with frequency-filtered images, the resulting model nee
ds relatively more data to adapt to naturally looking images during fine-tuning.
 To address these drawbacks, we propose FOurier transform compression with seLf-

Knowledge distillation (FOLK), integrating two dedicated ideas. First, inspired by image compression, we adaptively select the masked-out frequencies based on image frequency responses, creating more suitable SSL tasks for pre-training. Second, we employ a two-branch framework empowered by knowledge distillation, enabling the model to take both the filtered and original images as input, largely reducing the burden of downstream tasks. Our experimental results demonstrate the effectiveness of FOLK in achieving competitive performance to many state-of-the-art SSL methods across various downstream tasks, including image classification, few-shot learning, and semantic segmentation.
****
Letian Peng,Chenyang An,Jingbo Shang
Correlation and Navigation in the Vocabulary Key Representation Space of Language Models
Language model (LM) decoding is based on the next-token prediction (NTP) probability distribution. For neural LMs (e.g., Transformer-based), NTP distribution is essentially a softmax-regularized dot product between an encoded input context (query) and fixed vocabulary representations (keys). In this paper, we study the effect of the key distribution on the NTP distribution, with a focus on whether the similarity between keys will trigger spurious correlations in NTP. Through knowledge-probing tasks, we show that in the NTP distribution, the few top-ranked tokens are typically accurate. However, the middle-ranked prediction is highly biased towards the tokens that are distributionally (not necessarily semantically) similar to these top ones. For instance, if “P” is predicted as the top-1 token, “A”-“Z” will all be ranked high in NTP, no matter whether they can lead to correct decoding results. This hurts the sampling diversity and makes the sampling of correct, long-tail results hopeless and noisy. We attempt to alleviate this issue via a novel in-context method that iteratively pushes the query representation away from explored regions. Specifically, we include the explored decoding results in the context and prompt the LM to generate something else, which encourages the LM to produce a query representation that has small dot products with explored keys. Experiments on knowledge-probing tasks show that our method leads to efficient navigation away from explored keys to correct new keys. We further extend our method to open-ended and chain-of-thought (for reasoning) generation. Experiment results show that ICN contributes to better generation diversity and improved self-consistency voting performance. Finally, we discuss potential training issues caused by the fixed key space together with the challenges and possible ways to address them in future research.
****
Zhaoyang Wang,Weilei He,Zhiyuan Liang,Xuchao Zhang,Chetan Bansal,Ying Wei,Weitong Zhang,Huaxiu Yao
CREAM: Consistency Regularized Self-Rewarding Language Models
Recent self-rewarding large language models (LLM) have successfully applied LLM-as-a-Judge to iteratively improve the alignment performance without the need of human annotations for preference data. These methods commonly utilize the same LLM to act as both the policy model (which generates responses) and the reward model (which scores and ranks those responses). The ranked responses are then used as preference pairs to train the LLM via direct alignment technologies (e.g. DPO). However, it is noteworthy that throughout this process, there is no guarantee of accuracy in the rewarding and ranking, which is critical for ensuring accurate rewards and high-quality preference data. Empirical results from relatively small LLMs (e.g., 7B parameters) also indicate that improvements from self-rewar

ding may diminish after several iterations in certain situations, which we hypot
hesize is due to accumulated bias in the reward system. This bias can lead to un
reliable preference data for training the LLM. To address this issue, we first f
ormulate and analyze the generalized iterative preference fine-tuning framework
for self-rewarding language model. We then introduce the regularization to this
generalized framework to mitigate the overconfident preference labeling in the s
elf-rewarding process. Based on this theoretical insight, we propose a Consisten
cy Regularized sElf-rewarding lAnguage Model (CREAM) that leverages the consiste
ncy of rewards across different iterations to regularize the self-rewarding trai
ning, helping the model to learn from more reliable preference data. With this e
xplicit regularization, our empirical results demonstrate the superiority of CRE
AM in improving both reward consistency and alignment performance. The code is p
ublicly available at https://github.com/Raibows/CREAM.
****

Vincenzo Di Vito Francesco,Mostafa Mohammadian,Kyri Baker,Ferdinando Fioretto
Learning to Solve Differential Equation Constrained Optimization Problems
Differential equations (DE) constrained optimization plays a critical role in nu
merous scientific and engineering fields, including energy systems, aerospace en
gineering, ecology, and finance, where optimal configurations or control strateg
ies must be determined for systems governed by ordinary or stochastic differenti
al equations. Despite its significance, the computational challenges associated
with these problems have limited their practical use. To address these limitatio
ns, this paper introduces a learning-based approach to DE-constrained optimizati
on that combines techniques from proxy optimization \citep{kotary2021end} and ne
ural differential equations \citep{chen2019neural}. The proposed approach uses a
 dual-network architecture, with one approximating the control strategies, focus
ing on steady-state constraints, and another solving the associated DEs.
This combination enables the approximation of optimal strategies while accountin
g for dynamic constraints in near real-time.
Experiments across problems in energy optimization and finance modeling show tha
t this method provides full compliance with dynamic constraints and it produces
results up to 25 times more precise than other methods which do not explicitly m
odel the system's dynamic equations.
****

Xingqun Qi,Yatian Wang,Hengyuan Zhang,Jiahao Pan,Wei Xue,Shanghang Zhang,Wenhan
Luo,Qifeng Liu,Yike Guo
Co$^{\mathbf{3}}$Gesture: Towards Coherent Concurrent Co-speech 3D Gesture Gener
ation with Interactive Diffusion
Generating gestures from human speech has gained tremendous progress in animatin
g virtual avatars. While the existing methods enable synthesizing gestures coope
rated by people self-talking, they overlook the practicality of concurrent gestu
re modeling with two-person interactive conversations. Moreover, the lack of hig
h-quality datasets with concurrent co-speech gestures also limits handling this
issue. To fulfill this goal, we first construct a large-scale concurrent co-spee
ch gesture dataset that contains more than 7M frames for diverse two-person inte
ractive posture sequences, dubbed $\textbf{GES-Inter}$. Moreover, we propose Co$
^{\mathbf{3}}$Gesture, a novel framework that enables concurrent coherent co-spe
ech gesture synthesis including two-person interactive movements. Our framework
is built upon two cooperative generation branches conditioned on decomposed spea
ker audio. Specifically, to enhance the coordination of human postures w.r.t cor
responding speaker audios while interacting with the conversational partner, we
present a Temporal-Interaction Module ($\textbf{TIM}$). TIM can effectively mode
l the temporal association representation between two speakers' gesture sequence
s as interaction guidance and fuse it into the concurrent gesture generation. Th
en, we devise a mutual attention mechanism to further boost learning dependencie
s of interacted concurrent motions, thereby enabling us to generate vivid and co
herent gestures. Extensive experiments demonstrate that our method outperforms t
he state-of-the-art models on our newly collected GES-Inter dataset.
****

Dayal Singh Kalra,Tianyu He,Maissam Barkeshli

Universal Sharpness Dynamics in Neural Network Training: Fixed Point Analysis, E
dge of Stability, and Route to Chaos
In gradient descent dynamics of neural networks, the top eigenvalue of the Hessi
an of the loss (sharpness) displays a variety of robust phenomena throughout tra
ining. This includes early time regimes where the sharpness may decrease during
early periods of training (sharpness reduction), and later time behavior such as
 progressive sharpening and edge of stability. We demonstrate that a simple
-layer linear network (UV model) trained on a single training example exhibits a
ll of the essential sharpness phenomenology observed in real-world scenarios. By
 analyzing the structure of dynamical fixed points in function space and the vec
tor field of function updates, we uncover the underlying mechanisms behind these
 sharpness trends. Our analysis reveals (i) the mechanism behind early sharpness
 reduction and progressive sharpening, (ii) the required conditions for edge of
stability, and (iii) a period-doubling route to chaos on the edge of stability m
anifold as learning rate is increased. Finally, we demonstrate that various pred
ictions from this simplified model generalize to real-world scenarios and discus
s its limitations.
****

Indraneil Paul,Haoyi Yang,Goran Glavaš,Kristian Kersting,Iryna Gurevych
ObscuraCoder: Powering Efficient Code LM Pre-Training Via Obfuscation Grounding
Language models (LMs) have become a staple of the code-writing toolbox. Their pr
e-training recipe has, however, remained stagnant over recent years, barring the
 occasional changes in data sourcing and filtering strategies. In particular, re
search exploring modifications to Code-LMs' pre-training objectives, geared towa
rds improving data efficiency and better disentangling between syntax and semant
ics, has been noticeably sparse, especially compared with corresponding efforts
in natural language LMs. In this work, we examine grounding on obfuscated code a
s a means of helping Code-LMs look beyond the surface-form syntax and enhance th
eir pre-training sample efficiency. To this end, we compile ObscuraX, a dataset
of approximately 55M source and obfuscated code pairs in seven languages. Subseq
uently, we pre-train ObscuraCoder models, ranging in size from 255M to 2.8B para
meters, on a 272B-token corpus that includes ObscuraX and demonstrate that our o
bfuscation-based pre-training recipe leads to consistent improvements in Code-LM
s' abilities compared to both vanilla autoregressive pre-training as well as exi
sting de-obfuscation (DOBF) objectives. ObscuraCoder demonstrates sizeable gains
 across multiple tests of syntactic and semantic code understanding, along with
improved capabilities in multilingual code completion, multilingual code commit
summarization, and multi-purpose library-oriented code generation.
****

Julie Kallini,Shikhar Murty,Christopher D Manning,Christopher Potts,Róbert Csord
ás
MrT5: Dynamic Token Merging for Efficient Byte-level Language Models
Models that rely on subword tokenization have significant drawbacks, such as sen
sitivity to character-level noise like spelling errors and inconsistent compress
ion rates across different languages and scripts. While character- or byte-level
 models like ByT5 attempt to address these concerns, they have not gained widesp
read adoption—processing raw byte streams without tokenization results in signif
icantly longer sequence lengths, making training and inference inefficient. This
 work introduces MrT5 (MergeT5), a more efficient variant of ByT5 that integrate
s a token deletion mechanism in its encoder to dynamically shorten the input seq
uence length. After processing through a fixed number of encoder layers, a learn
ed delete gate determines which tokens are to be removed and which are to be ret
ained for subsequent layers. MrT5 effectively "merges" critical information from
 deleted tokens into a more compact sequence, leveraging contextual information
from the remaining tokens. In continued pre-training experiments, we find that M
rT5 can achieve significant gains in inference runtime with minimal effect on pe
rformance, as measured by bits-per-byte. Additionally, with multilingual trainin
g, MrT5 adapts to the orthographic characteristics of each language, learning la
nguage-specific compression rates. Furthermore, MrT5 shows comparable accuracy t
o ByT5 on downstream evaluations such as XNLI, TyDi QA, and character-level task

s while reducing sequence lengths by up to 75%. Our approach presents a solution to the practical limitations of existing byte-level models.
****
Seonghyeon Ye,Joel Jang,Byeongguk Jeon,Se June Joo,Jianwei Yang,Baolin Peng,Ajay Mandlekar,Reuben Tan,Yu-Wei Chao,Bill Yuchen Lin,Lars Liden,Kimin Lee,Jianfeng Gao,Luke Zettlemoyer,Dieter Fox,Minjoon Seo
Latent Action Pretraining from Videos
We introduce Latent Action Pretraining for general Action models (LAPA), the first unsupervised method for pretraining Vision-Language-Action (VLA) models without ground-truth robot action labels. Existing Vision-Language-Action models require action labels typically collected by human teleoperators during pretraining, which significantly limits possible data sources and scale. In this work, we propose a method to learn from internet-scale videos that do not have robot action labels. We first train an action quantization model leveraging VQ-VAE-based objective to learn discrete latent actions between image frames, then pretrain a latent VLA model to predict these latent actions from observations and task descriptions, and finally finetune the VLA on small-scale robot manipulation data to map from latent to robot actions. Experimental results demonstrate that our method significantly outperforms existing techniques that train robot manipulation policies from large-scale videos. Furthermore, it outperforms the state-of-the-art VLA model trained with robotic action labels on real-world manipulation tasks that require language conditioning, generalization to unseen objects, and semantic generalization to unseen instructions. Training only on human manipulation videos also shows positive transfer, opening up the potential for leveraging web-scale data for robotics foundation models.
****
Alihan Hüyük,Xinnuo Xu,Jacqueline R. M. A. Maasch,Aditya V. Nori,Javier Gonzalez
Reasoning Elicitation in Language Models via Counterfactual Feedback
Despite the increasing effectiveness of language models, their reasoning capabilities remain underdeveloped. In particular, causal reasoning through counterfactual question answering is lacking. This work aims to bridge this gap. We first derive novel metrics that balance accuracy in factual and counterfactual questions, capturing a more complete view of the reasoning abilities of language models than traditional factual-only based metrics. Second, we propose several fine-tuning approaches that aim to elicit better reasoning mechanisms, in the sense of the proposed metrics. Finally, we evaluate the performance of the fine-tuned language models in a variety of realistic scenarios. In particular, we investigate to what extent our fine-tuning approaches systemically achieve better generalization with respect to the base models in several problems that require, among others, inductive and deductive reasoning capabilities.
****
Marco Mistretta,Alberto Baldrati,Lorenzo Agnolucci,Marco Bertini,Andrew D. Bagdanov
Cross the Gap: Exposing the Intra-modal Misalignment in CLIP via Modality Inversion
Pre-trained multi-modal Vision-Language Models like CLIP are widely used off-the-shelf for a variety of applications. In this paper, we show that the common practice of individually exploiting the text or image encoders of these powerful multi-modal models is highly suboptimal for intra-modal tasks like image-to-image retrieval. We argue that this is inherently due to the CLIP-style inter-modal contrastive loss that does not enforce any intra-modal constraints, leading to what we call intra-modal misalignment. To demonstrate this, we leverage two optimization-based modality inversion techniques that map representations from their input modality to the complementary one without any need for auxiliary data or additional trained adapters. We empirically show that, in the intra-modal tasks of image-to-image and text-to-text retrieval, approaching these tasks inter-modally significantly improves performance with respect to intra-modal baselines on more than fifteen datasets. Additionally, we demonstrate that approaching a native inter-modal task (e.g. zero-shot image classification) intra-modally decreases performance, further validating our findings. Finally, we show that incorporating

an intra-modal term in the pre-training objective or narrowing the modality gap between the text and image feature embedding spaces helps reduce the intra-modal misalignment. The code is publicly available at: https://github.com/miccunifi/Cross-the-Gap.

****

Marco Sälzer,Eric Alsmann,Martin Lange

Transformer Encoder Satisfiability: Complexity and Impact on Formal Reasoning

We analyse the complexity of the satisfiability problem, or similarly feasibility problem, (trSAT) for transformer encoders (TE), which naturally occurs in formal verification or interpretation, collectively referred to as formal reasoning. We find that trSAT is undecidable when considering TE as they are commonly studied in the expressiveness community. Furthermore, we identify practical scenarios where trSAT is decidable and establish corresponding complexity bounds. Beyond trivial cases, we find that quantized TE, those restricted by fixed-width arithmetic, lead to the decidability of trSAT due to their limited attention capabilities. However, the problem remains difficult, as we establish scenarios where trSAT is NEXPTIME-hard and others where it is solvable in NEXPTIME for quantized TE. To complement our complexity results, we place our findings and their implications in the broader context of formal reasoning.

****

Xiao-Wen Yang,Zhi Zhou,Haiming Wang,Aoxue Li,Wen-Da Wei,Hui Jin,Zhenguo Li,Yu-Feng Li

CARTS: Advancing Neural Theorem Proving with Diversified Tactic Calibration and Bias-Resistant Tree Search

Recent advancements in neural theorem proving integrate large language models with tree search algorithms like Monte Carlo Tree Search (MCTS), where the language model suggests tactics and the tree search finds the complete proof path. However, many tactics proposed by the language model converge to semantically or strategically similar, reducing diversity and increasing search costs by expanding redundant proof paths. This issue exacerbates as computation scales and more tactics are explored per state. Furthermore, the trained value function suffers from false negatives, label imbalance, and domain gaps due to biased data construction. To address these challenges, we propose CARTS (diversified tactic CAlibration and bias-Resistant Tree Search), which balances tactic diversity and importance while calibrating model confidence. CARTS also introduce preference modeling and an adjustment term related to the ratio of valid tactics to improve the bias-resistance of the value function. Experimental results demonstrate that CARTS consistently outperforms previous methods achieving a pass@l rate of 49.6\% on the miniF2F-test benchmark. Further analysis confirms that CARTS improves tactic diversity and leads to a more balanced tree search.

****

Chukwudi Paul Obite,Zhi Chang,Keyan Wu,Shiwei Lan

Bayesian Regularization of Latent Representation

The effectiveness of statistical and machine learning methods depends on how well data features are characterized. Developing informative and interpretable latent representations with controlled complexity is essential for visualizing data structure and for facilitating efficient model building through dimensionality reduction. Latent variable models, such as Gaussian Process Latent Variable Models (GP-LVM), have become popular for learning complex, nonlinear representations as alternatives to Principal Component Analysis (PCA). In this paper, we propose a novel class of latent variable models based on the recently introduced Q-exponential process (QEP), which generalizes GP-LVM with a tunable complexity parameter, $q>0$. Our approach, the \emph{Q-exponential Process Latent Variable Model (QEP-LVM)}, subsumes GP-LVM as a special case when $q=2$, offering greater flexibility in managing representation complexity while enhancing interpretability. To ensure scalability, we incorporate sparse variational inference within a Bayesian training framework. We establish connections between QEP-LVM and probabilistic PCA, demonstrating its superior performance through experiments on datasets such as the Swiss roll, oil flow, and handwritten digits.

****

Chengke Zou,Xingang Guo,Rui Yang,Junyu Zhang,Bin Hu,Huan Zhang
DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models
The rapid advancements in Vision-Language Models (VLMs) have shown great potential in tackling mathematical reasoning tasks that involve visual context. Unlike humans who can reliably apply solution steps to similar problems with minor modifications, we found that state-of-the-art VLMs like GPT-4o can consistently fail in these scenarios, revealing limitations in their mathematical reasoning capabilities. In this paper, we investigate the **mathematical reasoning robustness** in VLMs and evaluate how well these models perform under different variants of the same question, such as changes in visual numerical values or function graphs.
While several vision-based math benchmarks have been developed to assess VLMs' problem-solving capabilities, these benchmarks contain only static sets of problems and cannot easily evaluate mathematical reasoning robustness.
To fill this gap, we introduce **DynaMath**, a dynamic visual math benchmark designed for in-depth assessment of VLMs. **DynaMath** includes 501 high-quality, multi-topic *seed* questions, *each represented as a Python program*. Those programs are carefully designed and annotated to enable the automatic generation of a much larger set of *concrete* questions, including many different types of visual and textual variations.
DynaMath allows us to evaluate the generalization ability of VLMs, by assessing their performance under varying input conditions of a seed question. We evaluated 14 state-of-the-art VLMs with 5,010 generated concrete questions (10 per seed question). Our results show that the worst-case model accuracy, defined as the percentage of correctly answered seed questions in all 10 variants, is significantly lower than the average-case accuracy. In addition, many models show high consistency in answering these questions -- the incorrectness of a certain variant of a seed question is not only due to inherent randomness. Our analysis emphasizes the need to study the robustness of VLMs' reasoning abilities, and **DynaMath** provides valuable insights to guide the development of more reliable models for mathematical reasoning.
****
Ziheng Cheng,Margalit Glasgow
Convergence of Distributed Adaptive Optimization with Local Updates
We study distributed adaptive algorithms with local updates (intermittent communication). Despite the great empirical success of adaptive methods in distributed training of modern machine learning models, the theoretical benefits of local updates within adaptive methods, particularly in terms of reducing communication complexity, have not been fully understood yet. In this paper, for the first time, we prove that \em Local SGD \em with momentum (\em Local \em SGDM) and \em Local \em Adam can outperform their minibatch counterparts in convex and weakly convex settings in certain regimes, respectively. Our analysis relies on a novel technique to prove contraction during local iterations, which is a crucial yet challenging step to show the advantages of local updates, under generalized smoothness assumption and gradient clipping strategy.
****
Yangzhen Wu,Zhiqing Sun,Shanda Li,Sean Welleck,Yiming Yang
Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving
While the scaling laws of large language models (LLMs) training have been extensively studied, optimal inference configurations of LLMs remain underexplored.
We study inference scaling laws (aka test-time scaling laws) and compute-optimal inference, focusing on the trade-offs between model sizes and generating additional tokens with different inference strategies. As a first step towards understanding and designing compute-optimal inference methods, we studied cost-performance trade-offs for inference strategies such as greedy search, majority voting, best-of-n, weighted voting, and two different tree search algorithms, using different model sizes and compute budgets. Our findings suggest that scaling inference compute with inference strategies can be more computationally efficient tha

n scaling model parameters. Additionally, smaller models combined with advanced inference algorithms offer Pareto-optimal trade-offs in cost and performance. Fo r example, the Llemma-7B model, when paired with our novel tree search algorithm , consistently outperforms the Llemma-34B model across all tested inference stra tegies on the MATH benchmark. We hope these insights contribute to a deeper unde rstanding of inference scaling laws (test-time scaling laws) for LLMs.
****

Danni Yuan,Mingda Zhang,Shaokui Wei,Li Liu,Baoyuan Wu
Activation Gradient based Poisoned Sample Detection Against Backdoor Attacks
This work studies the task of poisoned sample detection for defending against da ta poisoning based backdoor attacks. Its core challenge is finding a generalizab le and discriminative metric to distinguish between clean and various types of p oisoned samples (e.g., various triggers, various poisoning ratios). Inspired by a common phenomenon in backdoor attacks that the backdoored model tend to map si gnificantly different poisoned and clean samples within the target class to simi lar activation areas, we introduce a novel perspective of the circular distribut ion of the gradients w.r.t. sample activation, dubbed gradient circular distribu tion (GCD). And, we find two interesting observations based on GCD. One is that the GCD of samples in the target class is much more dispersed than that in the c lean class. The other is that in the GCD of target class, poisoned and clean sam ples are clearly separated. Inspired by above two observations, we develop an in novative three-stage poisoned sample detection approach, called Activation Gradi ent based Poisoned sample Detection (AGPD). First, we calculate GCDs of all clas ses from the model trained on the untrustworthy dataset. Then, we identify the t arget class(es) based on the difference on GCD dispersion between target and cle an classes. Last, we filter out poisoned samples within the identified target cl ass(es) based on the clear separation between poisoned and clean samples. Extens ive experiments under various settings of backdoor attacks demonstrate the super ior detection performance of the proposed method to existing poisoned detection approaches according to sample activation-based metrics.
****

Mehran Mirramezani,Anne S. Meeussen,Katia Bertoldi,Peter Orbanz,Ryan P Adams
Designing Mechanical Meta-Materials by Learning Equivariant Flows
Mechanical meta-materials are solids whose geometric structure results in exotic nonlinear behaviors that are not typically achievable via homogeneous materials . We show how to drastically expand the design space of a class of mechanical me ta-materials known as $\textit{cellular solids}$, by generalizing beyond transla tional symmetry. This is made possible by transforming a reference geometry acco rding to a divergence free flow that is parameterized by a neural network and eq uivariant under the relevant symmetry group. We show how to construct flows equi variant to the space groups, despite the fact that these groups are not compact. Coupling this flow with a differentiable nonlinear mechanics simulator allows u s to represent a much richer set of cellular solids than was previously possible . These materials can be optimized to exhibit desirable mechanical properties su ch as negative Poisson's ratios or to match target stress-strain curves. We vali date these new designs in simulation and by fabricating real-world prototypes. W e find that designs with higher-order symmetries can exhibit a wider range of be haviors.
****

Chayne Thrash,Reed Andreas,Ali Abbasi,Parsa Nooralinejad,Soroush Abbasi Koohpaye gani,Hamed Pirsiavash,Soheil Kolouri
MCNC: Manifold-Constrained Reparameterization for Neural Compression
The outstanding performance of large foundational models across diverse tasks, from computer vision to speech and natural language processing, has significantl y
increased their demand. However, storing and transmitting these models poses significant challenges due to their massive size (e.g., 750GB for Llama 3.1 405B).
Recent literature has focused on compressing the original weights or reducing th e

number of parameters required for fine-tuning these models. These compression methods generally constrain the parameter space, for example, through low-rank reparametrization (e.g., LoRA), pruning, or quantization (e.g., QLoRA) during or after the model training. In this paper, we present a novel model compression method, which we term Manifold-Constrained Neural Compression (MCNC). This method constrains the parameter space to low-dimensional pre-defined and frozen nonlinear manifolds, which effectively cover this space. Given the prevalence of good solutions in over-parameterized deep neural networks, we show that by constraining the parameter space to our proposed manifold, we can identify high-quality solutions while achieving unprecedented compression rates across a wide variety of tasks and architectures. Through extensive experiments in computer vision and natural language processing tasks, we demonstrate that our method significantly outperforms state-of-the-art baselines in terms of compression, accuracy, and/or model reconstruction time. Our code is publicly available at
https://github.com/mint-vu/MCNC.
****

Danqing Wang,Jianxin Ma,Fei Fang,Lei Li
TypedThinker: Diversify Large Language Model Reasoning with Typed Thinking
Large Language Models (LLMs) have demonstrated strong reasoning capabilities in solving complex problems. However, current approaches primarily enhance reasoning through the elaboration of thoughts while neglecting the diversity of reasoning types. LLMs typically employ deductive reasoning, proceeding step-by-step from given conditions, which limits their exploration during problem-solving. Our analysis reveals that certain problems are exclusively solvable through specific reasoning strategies like inductive, abductive, or analogical reasoning. However, incorporating diverse reasoning approaches presents two key challenges: identifying the appropriate reasoning type for each problem and exploiting this approach during problem-solving. Therefore, we propose the TypedThinker that predicts suitable reasoning types based on the problem and their previous effectiveness and provides relevant demonstrations to guide LLMs in applying these strategies. Experimental results show significant improvements across multiple benchmarks, with performance gains of 3.4\% for Mistral 7B, 6.5\% for LLaMA3 8B, and 7\% for Qwen 2 7B on logical and mathematical reasoning tasks. TypedThinker enhances LLM reasoning without requiring knowledge distillation from larger models. It can be integrated into more advanced systems like GPT-4o or specialized models like MetaMath to diversify their reasoning approaches and improve their problem-solving capabilities.
****

Han Shen,Pin-Yu Chen,Payel Das,Tianyi Chen
SEAL: Safety-enhanced Aligned LLM Fine-tuning via Bilevel Data Selection
Fine-tuning on task-specific data to boost downstream performance is a crucial step for leveraging Large Language Models (LLMs). However, though fine-tuning enhances the model performance for specialized applications, previous studies have demonstrated that fine-tuning the models on several adversarial samples or even benign data can greatly comprise the model's pre-equipped alignment and safety capabilities. In this work, we propose SEAL, a novel framework to enhance safety in LLM fine-tuning. SEAL learns a data ranker based on the bilevel optimization to up rank the safe and high-quality fine-tuning data and down rank the unsafe or low-quality ones. Models trained with SEAL demonstrate superior quality over multiple baselines, with 8.5\% and 9.7\% win rate increase compared to random selection respectively on Llama-3-8b-Instruct and Merlinite-7b models. Our code is available on github https://github.com/hanshen95/SEAL.
****

Haokai Hong,Wanyu Lin,KC Tan
Accelerating 3D Molecule Generation via Jointly Geometric Optimal Transport
This paper proposes a new 3D molecule generation framework, called GOAT, for fast and effective 3D molecule generation based on the flow-matching optimal transport objective. Specifically, we formulate a geometric transport formula for measuring the cost of mapping multi-modal features (e.g., continuous atom coordinate

s and categorical atom types) between a base distribution and a target data dist
ribution. Our formula is solved within a joint, equivariant, and smooth represen
tation space. This is achieved by transforming the multi-modal features into a c
ontinuous latent space with equivariant networks. In addition, we find that iden
tifying optimal distributional coupling is necessary for fast and effective tran
sport between any two distributions. We further propose a mechanism for estimati
ng and purifying optimal coupling to train the flow model with optimal transport
. By doing so, GOAT can turn arbitrary distribution couplings into new determini
stic couplings, leading to an estimated optimal transport plan for fast 3D molec
ule generation. The purification filters out the subpar molecules to ensure the
ultimate generation quality. We theoretically and empirically prove that the pro
posed optimal coupling estimation and purification yield transport plan with non
-increasing cost. Finally, extensive experiments show that GOAT enjoys the effic
iency of solving geometric optimal transport, leading to a double speedup compar
ed to the sub-optimal method while achieving the best generation quality regardi
ng validity, uniqueness, and novelty.
****

Vincent Cohen-Addad,Shaofeng H.-C. Jiang,Qiaoyuan Yang,Yubo Zhang,Samson Zhou
Fair Clustering in the Sliding Window Model
We study streaming algorithms for proportionally fair clustering, a notion origi
nally suggested by Chierichetti et al. (2017), in the sliding window model. We s
how that although there exist efficient streaming algorithms in the insertion-on
ly model, surprisingly no algorithm can achieve finite ratio without violating t
he fairness constraint in sliding window. Hence, the problem of fair clustering
is a rare separation between the insertion-only streaming model and the sliding
window model. On the other hand, we show that if the fairness constraint is rela
xed by a multiplicative $(1+\varepsilon)$ factor, there exists a $(1 + \varepsil
on)$-approximate sliding window algorithm that uses $\text{poly}(k\varepsilon^{-
1}\log n)$ space. This achieves essentially the best parameters (up to degree in
 the polynomial) provided the aforementioned lower bound. We also implement a nu
mber of empirical evaluations on real datasets to complement our theoretical res
ults.
****

Zhijing Jin,Max Kleiman-Weiner,Giorgio Piatti,Sydney Levine,Jiarui Liu,Fernando
Gonzalez Adauto,Francesco Ortu,András Strausz,Mrinmaya Sachan,Rada Mihalcea,Yeji
n Choi,Bernhard Schölkopf
Language Model Alignment in Multilingual Trolley Problems
We evaluate the moral alignment of large language models (LLMs) with human prefe
rences in multilingual trolley problems. Building on the Moral Machine experimen
t, which captures over 40 million human judgments across 200+ countries, we deve
lop a cross-lingual corpus of moral dilemma vignettes in over 100 languages call
ed MultiTP. This dataset enables the assessment of LLMs' decision-making process
es in diverse linguistic contexts. Our analysis explores the alignment of 19 dif
ferent LLMs with human judgments, capturing preferences across six moral dimensi
ons: species, gender, fitness, status, age, and the number of lives involved. By
 correlating these preferences with the demographic distribution of language spe
akers and examining the consistency of LLM responses to various prompt paraphras
ings, our findings provide insights into cross-lingual and ethical biases of LLM
s and their intersection. We discover significant variance in alignment across l
anguages, challenging the assumption of uniform moral reasoning in AI systems an
d highlighting the importance of incorporating diverse perspectives in AI ethics
. The results underscore the need for further research on the integration of mul
tilingual dimensions in responsible AI research to ensure fair and equitable AI
interactions worldwide.
****

Arun Verma,Zhongxiang Dai,Xiaoqiang Lin,Patrick Jaillet,Bryan Kian Hsiang Low
Neural Dueling Bandits: Preference-Based Optimization with Human Feedback
Contextual dueling bandit is used to model the bandit problems, where a learner'
s goal is to find the best arm for a given context using observed noisy human pr
eference feedback over the selected arms for the past contexts. However, existin

g algorithms assume the reward function is linear, which can be complex and non-linear in many real-life applications like online recommendations or ranking web search results. To overcome this challenge, we use a neural network to estimate the reward function using preference feedback for the previously selected arms. We propose upper confidence bound- and Thompson sampling-based algorithms with sub-linear regret guarantees that efficiently select arms in each round. We also extend our theoretical results to contextual bandit problems with binary feedback, which is in itself a non-trivial contribution. Experimental results on the problem instances derived from synthetic datasets corroborate our theoretical results.

****

Zixiong Yu,Songtao Tian,Guhan Chen

Divergence of Neural Tangent Kernel in Classification Problems

This paper primarily investigates the convergence of the Neural Tangent Kernel (NTK) in classification problems. This study firstly show the strictly positive definiteness of NTK of multi-layer fully connected neural networks and residual neural networks. Then, through a contradiction argument, it indicates that, during training with the cross-entropy loss function, the neural network parameters diverge due to the strictly positive definiteness of the NTK. Consequently, the empirical NTK does not consistently converge but instead diverges as time approaches infinity. This finding implies that NTK theory is not applicable in this context, highlighting significant theoretical implications for the study of neural networks in classification problems. These results can also be easily generalized to other network structures, provided that the NTK is strictly positive definite.

****

Wenhao XU,Xuefeng Gao,Xuedong He

Regret Bounds for Episodic Risk-Sensitive Linear Quadratic Regulator

Risk-sensitive linear quadratic regulator is one of the most fundamental problems in risk-sensitive optimal control.

In this paper, we study online adaptive control of risk-sensitive linear quadratic regulator in the finite horizon episodic setting.

We propose a simple least-squares greedy algorithm and show that it achieves $\widetilde{\mathcal{O}}(\log N)$ regret under a specific identifiability assumption, where N is the total number of episodes. If the identifiability assumption is not satisfied, we propose incorporating exploration noise into the least-squares-based algorithm, resulting in an algorithm with $\widetilde{\mathcal{O}}(\sqrt{N})$ regret.

To our best knowledge, this is the first set of regret bounds for episodic risk-sensitive linear quadratic regulator.

Our proof relies on perturbation analysis of less-standard Riccati equations for risk-sensitive linear quadratic control, and a delicate analysis of the loss in the risk-sensitive performance criterion due to applying the suboptimal controller in the online learning process.

****

Chenliang Li,Siliang Zeng,Zeyi Liao,Jiaxiang Li,Dongyeop Kang,Alfredo Garcia,Mingyi Hong

Joint Reward and Policy Learning with Demonstrations and Human Feedback Improves Alignment

Aligning to human preferences and/or intentions is an important requirement for contemporary foundation models. To ensure alignment, popular approaches such as reinforcement learning with human feedback (RLHF) break down the task into three stages: (i) a model is computed with supervised fine-tuning (SFT) based upon large demonstrations data, (ii) a reward model (RM) is estimated based upon human feedback data, and (iii) reinforcement learning (RL) is used to further refine the SFT model by optimizing the estimated reward model. Demonstrations and human feedback data reflect human user preferences in different ways. As a result, the reward model estimate obtained from only human feedback data is likely not as accurate as a reward model estimate obtained from both demonstration and human feedback data. A policy model that optimizes the reward model estimate obtained f

rom both demonstration and human feedback data will likely exhibit better alignm
ent performance. We introduce a tractable algorithm for finding the reward and p
olicy models and provide a finite-time performance guarantee. Additionally, we d
emonstrate the efficiency of the proposed solution with extensive experiments in
cluding alignment problems in LLMs and robotic control problems in MuJoCo. We ob
serve that the proposed solutions outperform the existing alignment algorithm by
 large margins, especially when the amounts of demonstration and preference data
 are unbalanced.
****

Nikolaos Nakis,Niels Raunkjær Holm,Andreas Lyhne Fiehn,Morten Mørup
How Low Can You Go? Searching for the Intrinsic Dimensionality of Complex Networ
ks using Metric Node Embeddings
Low-dimensional embeddings are essential for machine learning tasks involving gr
aphs, such as node classification, link prediction, community detection, network
 visualization, and network compression. Although recent studies have identified
 exact low-dimensional embeddings, the limits of the required embedding dimensio
ns remain unclear. We presently prove that lower dimensional embeddings are poss
ible when using Euclidean metric embeddings as opposed to vector-based Logistic
PCA (LPCA) embeddings. In particular, we provide an efficient logarithmic search
 procedure for identifying the exact embedding dimension and demonstrate how met
ric embeddings enable inference of the exact embedding dimensions of large-scale
 networks by exploiting that the metric properties can be used to provide linear
ithmic scaling. Empirically, we show that our approach extracts substantially lo
wer dimensional representations of networks than previously reported for small-s
ized networks. For the first time, we demonstrate that even large-scale networks
 can be effectively embedded in very low-dimensional spaces, and provide example
s of scalable, exact reconstruction for graphs with up to a million nodes. Our a
pproach highlights that the intrinsic dimensionality of networks is substantiall
y lower than previously reported and provides a computationally efficient assess
ment of the exact embedding dimension also of large-scale networks. The surprisi
ngly low dimensional representations achieved demonstrate that networks in gener
al can be losslessly represented using very low dimensional feature spaces, whic
h can be used to guide existing network analysis tasks from community detection
and node classification to structure revealing exact network visualizations.
****

Zhiqi KANG,Liyuan Wang,Xingxing Zhang,Karteek Alahari
Advancing Prompt-Based Methods for Replay-Independent General Continual Learning
General continual learning (GCL) is a broad concept to describe real-world conti
nual learning (CL) problems, which are often characterized by online data stream
s without distinct transitions between tasks, i.e., blurry task boundaries. Such
 requirements result in poor initial performance, limited generalizability, and
severe catastrophic forgetting, heavily impacting the effectiveness of mainstrea
m GCL models trained from scratch. While the use of a frozen pretrained backbone
 with appropriate prompt tuning can partially address these challenges, such pro
mpt-based methods remain suboptimal for CL of remaining tunable parameters on th
e fly. In this regard, we propose an innovative approach named MISA (Mask and In
itial Session Adaption) to advance prompt-based methods in GCL. It includes a fo
rgetting-aware initial session adaption that employs pretraining data to initial
ize prompt parameters and improve generalizability, as well as a non-parametric
logit mask of the output layers to mitigate catastrophic forgetting. Empirical r
esults demonstrate substantial performance gains of our approach compared to rec
ent competitors, especially without a replay buffer (e.g., up to 18.39, 22.06, a
nd 11.96% points performance lead on CIFAR-100, Tiny-ImageNet, and ImageNet-R, r
espectively). Moreover, our approach features the plug-in nature for prompt-base
d methods, independence of replay, ease of implementation, and avoidance of CL-r
elevant hyperparameters, serving as a strong baseline for GCL research. Our sour
ce code is publicly available at https://github.com/kangzhiq/MISA
****

Ittai Rubinstein,Samuel B. Hopkins
Robustness Auditing for Linear Regression: To Singularity and Beyond

It has recently been discovered that the conclusions of many highly influential econometrics studies can be overturned by removing a very small fraction of their samples (often less than 0.5%). These conclusions are typically based on the results of one or more Ordinary Least Squares (OLS) regressions, raising the question: given a dataset, can we certify the robustness of an OLS fit on this dataset to the removal of a given number of samples?

Brute-force techniques quickly break down even on small datasets. Existing approaches which go beyond brute force either can only find candidate small subsets to remove (but cannot certify their non-existence) [BGM20, KZC21], are computationally intractable beyond low dimensional settings [MR22], or require very strong assumptions on the data distribution and too many samples to give reasonable bounds in practice [BP21, FH23].

We present an efficient algorithm for certifying the robustness of linear regressions to removals of samples. We implement our algorithm and run it on several landmark econometrics datasets with hundreds of dimensions and tens of thousands of samples, giving the first non-trivial certificates of robustness to sample removal for datasets of dimension 4 or greater. We prove that under distributional assumptions on a dataset, the bounds produced by our algorithm are tight up to a $1 + o(1)$ multiplicative factor.
****
Hanrong Zhang,Jingyuan Huang,Kai Mei,Yifei Yao,Zhenting Wang,Chenlu Zhan,Hongwei Wang,Yongfeng Zhang
Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents
Although LLM-based agents, powered by Large Language Models (LLMs), can use external tools and memory mechanisms to solve complex real-world tasks, they may also introduce critical security vulnerabilities. However, the existing literature does not comprehensively evaluate attacks and defenses against LLM-based agents. To address this, we introduce Agent Security Bench (ASB), a comprehensive framework designed to formalize, benchmark, and evaluate the attacks and defenses of LLM-based agents, including 10 scenarios (e.g., e-commerce, autonomous driving, finance), 10 agents targeting the scenarios, over 400 tools, 27 different types of attack/defense methods, and 7 evaluation metrics. Based on ASB, we benchmark 10 prompt injection attacks, a memory poisoning attack, a novel Plan-of-Thought backdoor attack, 4 mixed attacks, and 11 corresponding defenses across 13 LLM backbones. Our benchmark results reveal critical vulnerabilities in different stages of agent operation, including system prompt, user prompt handling, tool usage, and memory retrieval, with the highest average attack success rate of 84.30\%, but limited effectiveness shown in current defenses, unveiling important works to be done in terms of agent security for the community. We also introduce a new metric to evaluate the agents' capability to balance utility and security. Our code can be found at
https://github.com/agiresearch/ASB.
****
Simon Schug,Seijin Kobayashi,Yassir Akram,Joao Sacramento,Razvan Pascanu
Attention as a Hypernetwork
Transformers can under some circumstances generalize to novel problem instances whose constituent parts might have been encountered during training, but whose compositions have not.
What mechanisms underlie this ability for compositional generalization?
By reformulating multi-head attention as a hypernetwork, we reveal that a composable, low-dimensional latent code specifies key-query specific operations.
We find empirically that this latent code is predictive of the subtasks the network performs on unseen task compositions, revealing that latent codes acquired during training are reused to solve unseen problem instances.
To further examine the hypothesis that the intrinsic hypernetwork of multi-head attention supports compositional generalization, we ablate whether making the hypernetwork-generated linear value network nonlinear strengthens compositionality

.
We find that this modification improves compositional generalization on abstract reasoning tasks.
In particular, we introduce a symbolic version of the Raven's Progressive Matrices human intelligence test, which gives us precise control over the problem compositions encountered during training and evaluation.
We demonstrate on this task how scaling model size and data enables compositional generalization in transformers and gives rise to a functionally structured latent space.
****
Youngjun Lee,Doyoung Kim,Junhyeok Kang,Jihwan Bang,Hwanjun Song,Jae-Gil Lee
RA-TTA: Retrieval-Augmented Test-Time Adaptation for Vision-Language Models
Vision-language models (VLMs) are known to be susceptible to distribution shifts between pre-training data and test data, and test-time adaptation (TTA) methods for VLMs have been proposed to mitigate the detrimental impact of the distribution shifts. However, the existing methods solely rely on the internal knowledge encoded within the model parameters, which are constrained to pre-training data. To complement the limitation of the internal knowledge, we propose **Retrieval-Augmented-TTA (RA-TTA)** for adapting VLMs to test distribution using **external** knowledge obtained from a web-scale image database. By fully exploiting the bi-modality of VLMs, RA-TTA **adaptively** retrieves proper external images for each test image to refine VLMs' predictions using the retrieved external images, where fine-grained **text descriptions** are leveraged to extend the granularity of external knowledge. Extensive experiments on 17 datasets demonstrate that the proposed RA-TTA outperforms the state-of-the-art methods by 3.01-9.63\% on average.
****
Tennison Liu,Nicolas Huynh,Mihaela van der Schaar
Decision Tree Induction Through LLMs via Semantically-Aware Evolution
Decision trees are a crucial class of models offering robust predictive performance and inherent interpretability across various domains, including healthcare, finance, and logistics. However, current tree induction methods often face limitations such as suboptimal solutions from greedy methods or prohibitive computational costs and limited applicability of exact optimization approaches.
To address these challenges, we propose an evolutionary optimization method for decision tree induction based on genetic programming (GP). Our key innovation is the integration of semantic priors and domain-specific knowledge about the search space into the optimization algorithm.
To this end, we introduce $\texttt{LLEGO}$, a framework that incorporates semantic priors into genetic search operators through the use of Large Language Models (LLMs), thereby enhancing search efficiency and targeting regions of the search space that yield decision trees with superior generalization performance. This is operationalized through novel genetic operators that work with structured natural language prompts, effectively utilizing LLMs as conditional generative models and sources of semantic knowledge. Specifically, we introduce $\textit{fitness-guided}$ crossover to exploit high-performing regions, and $\textit{diversity-guided}$ mutation for efficient global exploration of the search space. These operators are controlled by corresponding hyperparameters that enable a more nuanced balance between exploration and exploitation across the search space. Empirically, we demonstrate across various benchmarks that $\texttt{LLEGO}$ evolves superior-performing trees compared to existing tree induction methods, and exhibits significantly more efficient search performance compared to conventional GP approaches.
****
Yujian Liu,Shiyu Chang,Tommi Jaakkola,Yang Zhang
Fictitious Synthetic Data Can Improve LLM Factuality via Prerequisite Learning
Recent studies have identified one aggravating factor of LLM hallucinations as the knowledge inconsistency between pre-training and fine-tuning, where unfamiliar fine-tuning data mislead the LLM to fabricate plausible but wrong outputs. In this paper, we propose a novel fine-tuning strategy called Prereq-Tune to addres

s this knowledge inconsistency and reduce hallucinations. Fundamentally, Prereq-Tune disentangles the learning of skills and knowledge, so the model learns only the task skills without being impacted by the knowledge inconsistency. To achieve this, Prereq-Tune introduces an additional prerequisite learning stage to learn the necessary knowledge for SFT, allowing subsequent SFT to focus only on task skills. Prereq-Tune can also be combined with fictitious synthetic data to enhance the grounding of LLM outputs to their internal knowledge. Experiments show that Prereq-Tune outperforms existing baselines in improving LLM's factuality across short QA and long-form generation tasks. It also opens new possibilities for knowledge-controlled generation in LLMs. Our code is available at https://github.com/UCSB-NLP-Chang/Prereq_tune.git.
****

Wenhui Tan,Boyuan Li,Chuhao Jin,Wenbing Huang,Xiting Wang,Ruihua Song
Think Then React: Towards Unconstrained Action-to-Reaction Motion Generation
Modeling human-like action-to-reaction generation has significant real-world applications, like human-robot interaction and games.
Despite recent advancements in single-person motion generation, it is still challenging to well handle action-to-reaction generation, due to the difficulty of directly predicting reaction from action sequence without prompts, and the absence of a unified representation that effectively encodes multi-person motion.
To address these challenges, we introduce Think-Then-React (TTR), a large language-model-based framework designed to generate human-like reactions.
First, with our fine-grained multimodal training strategy, TTR is capable to unify two processes during inference: a thinking process that explicitly infers action intentions and reasons corresponding reaction description, which serve as semantic prompts, and a reacting process that predicts reactions based on input action and the inferred semantic prompts.
Second, to effectively represent multi-person motion in language models, we propose a unified motion tokenizer by decoupling egocentric pose and absolute space features, which effectively represents action and reaction motion with same encoding.
Extensive experiments demonstrate that TTR outperforms existing baselines, achieving significant improvements in evaluation metrics, such as reducing FID from 3.988 to 1.942.
****

Zeyu Gan,Yong Liu
Towards a Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective
Synthetic data has become a pivotal resource in post-training tasks for large language models (LLMs) due to the scarcity of high-quality, specific data. While various methods have been developed to generate synthetic data, there remains a discernible gap between the practical effects of synthetic data and our theoretical comprehension. To address this challenge, we commence by presenting a detailed modeling of the prevalent synthetic data generation process. Building upon this modeling, we demonstrate that the generalization capability of the post-trained model is critically determined by the information gain derived from the generative model, as analyzed from a novel reverse-bottleneck perspective. Moreover, we introduce the concept of Generalization Gain via Mutual Information (GGMI) and elucidate the relationship between generalization gain and information gain. This analysis serves as a theoretical foundation for synthetic data generation and further highlights its connection with the generalization capability of post-trained models, offering an understanding about the design of synthetic data generation techniques and the optimization of the post-training process. We open-source our code at https://github.com/ZyGan1999/Towards-a-Theoretical-Understanding-of-Synthetic-Data-in-LLM-Post-Training.
****

Zhen Yang,Ziwei Du,Minghan Zhang,Wei Du,Jie Chen,Zhen Duan,Shu Zhao
Triples as the Key: Structuring Makes Decomposition and Verification Easier in LLM-based TableQA
As the mainstream approach, LLMs have been widely applied and researched in Tabl

eQA tasks. Currently, the core of LLM-based TableQA methods typically include th
ree phases: question decomposition, sub-question TableQA reasoning, and answer v
erification. However, several challenges remain in this process: i) Sub-question
s generated by these methods often exhibit significant gaps with the original qu
estion due to critical information overlooked during the LLM's direct decomposit
ion; ii) Verification of answers is typically challenging because LLMs tend to g
enerate optimal responses during self-correct. To address these challenges, we p
ropose a Triple-Inspired Decomposition and vErification (TIDE) strategy, which l
everages the structural properties of triples to assist in decomposition and ver
ification in TableQA. The inherent structure of triples (head entity, relation,
tail entity) requires the LLM to extract as many entities and relations from the
 question as possible. Unlike direct decomposition methods that may overlook key
 information, our transformed sub-questions using triples encompass more critica
l details. Additionally, this explicit structure facilitates verification. By co
mparing the triples derived from the answers with those from the question decomp
osition, we can achieve easier and more straightforward validation than when rel
ying on the LLM's self-correct tendencies. By employing triples alongside establ
ished LLM modes, Direct Prompting and Agent modes, TIDE achieves state-of-the-ar
t performance across multiple TableQA datasets, demonstrating the effectiveness
of our method.
****
Arthur Jacot,Seok Hoan Choi,Yuxiao Wen
How DNNs break the Curse of Dimensionality: Compositionality and Symmetry Learni
ng
We show that deep neural networks (DNNs) can efficiently learn any
composition of functions with bounded F_{1}-norm, which allows
DNNs to break the curse of dimensionality in ways that shallow networks
cannot. More specifically, we derive a generalization bound that combines
a covering number argument for compositionality, and the F_{1}-norm
(or the related Barron norm) for large width adaptivity. We show that
the global minimizer of the regularized loss of DNNs can fit for example
the composition of two functions $f^{*}=h\circ g$ from a small number
of observations, assuming g is smooth/regular and reduces the dimensionality
(e.g. g could be the quotient map of the symmetries of f^{*}),
so that h can be learned in spite of its low regularity. The measures
of regularity we consider is the Sobolev norm with different levels
of differentiability, which is well adapted to the F_{1} norm.
We compute scaling laws empirically and observe phase transitions
depending on whether g or h is harder to learn, as predicted
by our theory.
****
Tianfang Zhu,Dongli Hu,Jiandong Zhou,Kai Du,Anan LI
Biologically Constrained Barrel Cortex Model Integrates Whisker Inputs and Repli
cates Key Brain Network Dynamics
The brain's ability to transform sensory inputs into motor functions is central
to neuroscience and crucial for the development of embodied intelligence. Sensor
y-motor integration involves complex neural circuits, diverse neuronal types, an
d intricate intercellular connections. Bridging the gap between biological reali
sm and behavioral functionality presents a formidable challenge. In this study,
we focus on the columnar structure of the superficial layers of mouse barrel cor
tex as a model system. We constructed a model comprising 4,218 neurons across 13
 neuronal subtypes, with neural distribution and connection strengths constraine
d by anatomical experimental findings. A key innovation of our work is the devel
opment of an effective construction and training pipeline tailored for this biol
ogically constrained model. Additionally, we converted an existing simulated whi
sker sweep dataset into a spiking-based format, enabling our network to be train
ed and tested on neural signals that more closely mimic those observed in biolog
ical systems. The results of object discrimination utilizing whisker signals dem
onstrate that our barrel cortex model, grounded in biological constraints, achie
ves a classification accuracy exceeds classical convolutional neural networks (C

NNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs), by an average of 8.6%, and is on par with recent spiking neural networks (SNNs) in performance. Interestingly, a whisker deprivation experiment, designed in accordance with neuroscience practices, further validates the perceptual capabilities of our model in behavioral tasks.
Critically, it offers significant biological interpretability: post-training analysis reveals that neurons within our model exhibit firing characteristics and distribution patterns similar to those observed in the actual neuronal systems of the barrel cortex. This study advances our understanding of neural processing in the barrel cortex and exemplifies how integrating detailed biological structures into neural network models can enhance both scientific inquiry and artificial intelligence applications. The code is available at https://github.com/fun0515/RSNN_bfd.
****
Riccardo Grazzi,Julien Siems,Arber Zela,Jörg K.H. Franke,Frank Hutter,Massimiliano Pontil
Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues
Linear Recurrent Neural Networks (LRNNs) such as Mamba, RWKV, GLA, mLSTM, and DeltaNet have emerged as efficient alternatives to Transformers for long sequences. However, both Transformers and LRNNs struggle to perform state-tracking, which may impair performance in tasks such as code evaluation. In one forward pass, current architectures are unable to solve even parity, the simplest state-tracking task, which non-linear RNNs can handle effectively. Recently, Sarrof et al. (2024) demonstrated that the failure of LRNNs like Mamba to solve parity stems from restricting the value range of their diagonal state-transition matrices to $[0, 1]$ and that incorporating negative values can resolve this issue. We extend this result to non-diagonal LRNNs such as DeltaNet. We prove that finite precision LRNNs with state-transition matrices having only positive eigenvalues cannot solve parity, while non-triangular matrices are needed to count modulo 3. Notably, we also prove that LRNNs can learn any regular language when their state-transition matrices are products of identity minus vector outer product matrices, each with eigenvalues in the range $[-1, 1]$. Our experiments confirm that extending the eigenvalue range of Mamba and DeltaNet to include negative values not only enables them to solve parity but consistently improves their performance on state-tracking tasks. We also show that state-tracking enabled LRNNs can be pretrained stably and efficiently at scale (1.3B parameters), achieving competitive performance on language modeling and showing promise on code and math tasks.
****
Cade W Gordon,Amy X. Lu,Pieter Abbeel
Protein Language Model Fitness is a Matter of Preference
Leveraging billions of years of evolution, scientists have trained protein language models (pLMs) to understand the sequence and structure space of proteins aiding in the design of more functional proteins. Although they have shown ability to improve efficiency in engineering, it remains unclear under what conditions they will succeed or fail. We aim to predict the circumstances in which pLMs can successfully perform zero-shot fitness estimation. Our work demonstrates the trends observed over hundreds of deep mutational scans across multiple different fitness objectives. We find that the likelihood, or abstractly, implicit preference of a certain protein sequence imbued during pretraining is predictive fitness prediction capabilities. Both over-preferred and under-preferred wild type sequences harm performance. Generating a causal link between training data and likelihood, we show a power law tail over what data increases protein likelihood which is tied to training sequence homology. Lastly, proteins of low likelihood can be remedied by unsupervised finetuning. In sum, the zero-shot fitness estimation abilities of pLMs can be predicted by the likelihood of the engineered sequence, thus suggesting when pLMs should be deployed in protein maturation campaigns and a way to improve their performance under circumstances of low likelihood.
****
Shaobo Wang,Hongxuan Tang,Mingyang Wang,Hongrui Zhang,Xuyang Liu,Weiya Li,Xuming Hu,Linfeng Zhang

Gnothi Seauton: Empowering Faithful Self-Interpretability in Black-Box Transformers

The debate between self-interpretable models and post-hoc explanations for black-box models is central to Explainable AI (XAI). Self-interpretable models, such as concept-based networks, offer insights by connecting decisions to human-understandable concepts but often struggle with performance and scalability. Conversely, post-hoc methods like Shapley values, while theoretically robust, are computationally expensive and resource-intensive. To bridge the gap between these two lines of research, we propose a novel method that combines their strengths, providing theoretically guaranteed self-interpretability for black-box models without compromising prediction accuracy. Specifically, we introduce a parameter-efficient pipeline, AutoGnothi, which integrates a small side network into the black-box model, allowing it to generate Shapley value explanations without changing the original network parameters. This side-tuning approach significantly reduces memory, training, and inference costs, outperforming traditional parameter-efficient methods, where full fine-tuning serves as the optimal baseline. AutoGnothi enables the black-box model to predict and explain its predictions with minimal overhead. Extensive experiments show that AutoGnothi offers accurate explanations for both vision and language tasks, delivering superior computational efficiency with comparable interpretability.

****

Linzheng Chai,Shukai Liu,Jian Yang,Yuwei Yin,JinKe,Jiaheng Liu,Tao Sun,Ge Zhang,Changyu Ren,Hongcheng Guo,Noah Wang,Boyang Wang,Xianjie Wu,Bing Wang,Tongliang Li,Liqun Yang,Sufeng Duan,Zhaoxiang Zhang,Zhoujun Li

McEval: Massively Multilingual Code Evaluation

Code large language models (LLMs) have shown remarkable advances in code understanding, completion, and generation tasks. Programming benchmarks, comprised of a selection of code challenges and corresponding test cases, serve as a standard to evaluate the capability of different LLMs in such tasks. However, most existing benchmarks primarily focus on Python and are still restricted to a limited number of languages, where other languages are translated from the Python samples degrading the data diversity. To further facilitate the research of code LLMs, we propose a massively multilingual code benchmark covering 40 programming languages (McEval) with 16K test samples, which substantially pushes the limits of code LLMs in multilingual scenarios. The benchmark contains challenging code completion, understanding, and generation evaluation tasks with finely curated massively multilingual instruction corpora McEval-Instruct. In addition, we introduce an effective multilingual coder mCoder trained on McEval-Instruct to support multilingual programming language generation. Extensive experimental results on McEval show that there is still a difficult journey between open-source models and closed-source LLMs in numerous languages. The instruction corpora and evaluation benchmark are available at https://github.com/MCEVAL/McEval.

****

Ozan Ozdenizci,Elmar Rueckert,Robert Legenstein

Privacy-Aware Lifelong Learning

Lifelong learning algorithms enable models to incrementally acquire new knowledge without forgetting previously learned information. Contrarily, the field of machine unlearning focuses on explicitly forgetting certain previous knowledge from pretrained models when requested, in order to comply with data privacy regulations on the right-to-be-forgotten. Enabling efficient lifelong learning with the capability to selectively unlearn sensitive information from models presents a critical and largely unaddressed challenge with contradicting objectives. We address this problem from the perspective of simultaneously preventing catastrophic forgetting and allowing forward knowledge transfer during task-incremental learning, while ensuring exact task unlearning and minimizing memory requirements, based on a single neural network model to be adapted. Our proposed solution, privacy-aware lifelong learning (PALL), involves optimization of task-specific sparse subnetworks with parameter sharing within a single architecture. We additionally utilize an episodic memory rehearsal mechanism to facilitate exact unlearning without performance degradations. We empirically demonstrate the scalability of

PALL across various architectures in image classification, and provide a state-of-the-art solution that uniquely integrates lifelong learning and privacy-aware unlearning mechanisms for responsible AI applications.
****
Wenbo Hu,Jia-Chen Gu,Zi-Yi Dou,Mohsen Fayyaz,Pan Lu,Kai-Wei Chang,Nanyun Peng
MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models
Existing multimodal retrieval benchmarks primarily focus on evaluating whether models can retrieve and utilize external textual knowledge for question answering. However, there are scenarios where retrieving visual information is either more beneficial or easier to access than textual data.
In this paper, we introduce a multimodal retrieval-augmented generation benchmark, MRAG-Bench, in which we systematically identify and categorize scenarios where visually augmented knowledge is better than textual knowledge, for instance, more images from varying viewpoints.
MRAG-Bench consists of 16,130 images and 1,353 human-annotated multiple-choice questions across 9 distinct scenarios. With MRAG-Bench, we conduct an evaluation of 10 open-source and 4 proprietary large vision-language models (LVLMs). Our results show that all LVLMs exhibit greater improvements when augmented with images compared to textual knowledge, confirming that MRAG-Bench is vision-centric. Additionally, we conduct extensive analysis with MRAG-Bench, which offers valuable insights into retrieval-augmented LVLMs. Notably, the top-performing model, GPT-4o, faces challenges in effectively leveraging retrieved knowledge, achieving only a 5.82\% improvement with ground-truth information, in contrast to a 33.16\% improvement observed in human participants. These findings highlight the importance of MRAG-Bench in encouraging the community to enhance LVLMs' ability to utilize retrieved visual knowledge more effectively.
****
Tianjun Yao,Yongqiang Chen,Kai Hu,Tongliang Liu,Kun Zhang,Zhiqiang Shen
Learning Graph Invariance by Harnessing Spuriosity
Recently, graph invariant learning has become the _de facto_ approach to tackle the Out-of-Distribution (OOD) generalization failure in graph representation learning. They generically follow the framework of invariant risk minimization to capture the invariance of graph data from different environments. Despite some success, it remains unclear to what extent existing approaches have captured invariant features for OOD generalization on graphs. In this work, we find that representative OOD methods such as IRM and VRex, and their variants on graph invariant learning may have captured a limited set of invariant features. To tackle this challenge, we propose $\texttt{LIRS}$, a novel learning framework designed to **L**earn graph **I**nvariance by **R**emoving **S**purious features. Different from most existing approaches that _directly_ learn the invariant features, $\texttt{LIRS}$ takes an _indirect_ approach by first learning the spurious features and then removing them from the ERM-learned features, which contains both spurious and invariant features. We demonstrate that learning the invariant graph features in an _indirect_ way can learn a more comprehensive set of invariant features. Moreover, our proposed method outperforms the second-best method by as much as 25.50% across all competitive baseline methods, highlighting its effectiveness in learning graph invariant features.
****
Shenyu Lu,Junyi Chai,Xiaoqian Wang
Mitigating Spurious Correlations in Zero-Shot Multimodal Models
Multimodal models or Vision Language Models (VLMs) have reshaped the paradigm in machine learning, offering zero-shot capabilities that require no additional training when adapted to new classification tasks. However, despite their advancements, spurious correlations still exist in VLMs. Existing approaches to tackle this issue often require target label annotations, contradicting the principle of zero-shot classification, or they primarily focus on a single modality, risking misalignment between text and image modalities. Others rely on extensive domain knowledge or large language models (LLMs) to characterize spurious features, making the performance sensitive to the generated prompts and undermining zero-shot capability. In response, we propose a new solution that tackles spurious corre

lations in VLMs within the zero-shot setting. Our approach utilizes a translatio
n operation that preserves the latent space distribution to address issues of sp
urious correlations. In particular, our method is grounded in and inspired by a
theoretical analysis, which identifies that the optimal translation directions a
re along the spurious vector. As VLMs unify two modalities, we compute spurious
vectors from the text prompts and guide the translation for image embeddings, al
igning the requirements for the fusion of different modalities in VLMs. We condu
cted experiments on benchmark datasets, which have shown significant improvement
s in worst-group accuracy. Additionally, our visualizations of VLMs further demo
nstrate the effectiveness of this intervention.
****
Xiaoyi MAI,Zhenyu Liao
The Breakdown of Gaussian Universality in Classification of High-dimensional Lin
ear Factor Mixtures
The assumption of Gaussian or Gaussian mixture data has been extensively exploit
ed in a long series of precise performance analyses of machine learning (ML) met
hods, on large datasets having comparably numerous samples and features.
To relax this restrictive assumption, subsequent efforts have been devoted to es
tablish "Gaussian equivalent principles" by studying scenarios of Gaussian unive
rsality where the asymptotic performance of ML methods on non-Gaussian data rema
ins unchanged when replaced with Gaussian data having the *same mean and covaria
nce*.
Beyond the realm of Gaussian universality, there are few exact results on how th
e data distribution affects the learning performance.

In this article, we provide a precise high-dimensional characterization of empir
ical risk minimization, for classification under a general mixture data setting
of *linear factor models* that extends Gaussian mixtures.
The Gaussian universality is shown to break down under this setting, in the sens
e that the asymptotic learning performance depends on the data distribution *bey
ond* the class means and covariances.
To clarify the limitations of Gaussian universality in the classification of mix
ture data and to understand the impact of its breakdown, we specify conditions f
or Gaussian universality and discuss their implications for the choice of loss f
unction.
****
Zhihao Shi,Jie Wang,Zhiwei Zhuang,Xize Liang,Bin Li,Feng Wu
Accurate and Scalable Graph Neural Networks via Message Invariance
Message passing-based graph neural networks (GNNs) have achieved great success i
n many real-world applications. For a sampled mini-batch of target nodes, the me
ssage passing process is divided into two parts: message passing between nodes w
ithin the batch (MP-IB) and message passing from nodes outside the batch to thos
e within it (MP-OB). However, MP-OB recursively relies on higher-order out-of-ba
tch neighbors, leading to an exponentially growing computational cost with respe
ct to the number of layers. Due to the neighbor explosion, the whole message pas
sing stores most nodes and edges on the GPU such that many GNNs are infeasible t
o large-scale graphs. To address this challenge, we propose an accurate and fast
 mini-batch approach for large graph transductive learning, namely topological c
ompensation (TOP), which obtains the outputs of the whole message passing solely
 through MP-IB, without the costly MP-OB. The major pillar of TOP is a novel con
cept of message invariance, which defines message-invariant transformations to c
onvert costly MP-OB into fast MP-IB. This ensures that the modified MP-IB has th
e same output as the whole message passing. Experiments demonstrate that TOP is
significantly faster than existing mini-batch methods by order of magnitude on v
ast graphs (millions of nodes and billions of edges) with limited accuracy degra
dation.
****
Ryuichi Kanoh,Mahito Sugiyama
Linear Mode Connectivity in Differentiable Tree Ensembles
Linear Mode Connectivity (LMC) refers to the phenomenon that performance remains

consistent for linearly interpolated models in the parameter space. For independently optimized model pairs from different random initializations, achieving LMC is considered crucial for understanding the stable success of the non-convex optimization in modern machine learning models and for facilitating practical parameter-based operations such as model merging. While LMC has been achieved for neural networks by considering the permutation invariance of neurons in each hidden layer, its attainment for other models remains an open question. In this paper, we first achieve LMC for soft tree ensembles, which are tree-based differentiable models extensively used in practice. We show the necessity of incorporating two invariances: subtree flip invariance and splitting order invariance, which do not exist in neural networks but are inherent to tree architectures, in addition to permutation invariance of trees. Moreover, we demonstrate that it is even possible to exclude such additional invariances while keeping LMC by designing decision list-based tree architectures, where such invariances do not exist by definition. Our findings indicate the significance of accounting for architecture-specific invariances in achieving LMC.
****

Ziwei Xu,Mohan Kankanhalli
Strong Preferences Affect the Robustness of Preference Models and Value Alignment
Value alignment, which aims to ensure that large language models (LLMs) and other AI agents behave in accordance with human values, is critical for ensuring safety and trustworthiness of these systems. A key component of value alignment is the modeling of human preferences as a representation of human values. In this paper, we investigate the robustness of value alignment by examining the sensitivity of preference models. Specifically, we ask: how do changes in the probabilities of some preferences affect the predictions of these models for other preferences? To answer this question, we theoretically analyze the robustness of widely used preference models by examining their sensitivities to minor changes in preferences they model. Our findings reveal that, in the Bradley-Terry and the Plackett-Luce model, the probability of a preference can change significantly as other preferences change, especially when these preferences are dominant (i.e., with probabilities near zero or one). We identify specific conditions where this sensitivity becomes significant for these models and discuss the practical implications for the robustness and safety of value alignment in AI systems.
****

Kuo-Han Hung,Pang-Chi Lo,Jia-Fong Yeh,Han-Yuan Hsu,Yi-Ting Chen,Winston H. Hsu
VICtoR: Learning Hierarchical Vision-Instruction Correlation Rewards for Long-horizon Manipulation
We study reward models for long-horizon manipulation by learning from action-free videos and language instructions, which we term the visual-instruction correlation (VIC) problem. Existing VIC methods face challenges in learning rewards for long-horizon tasks due to their lack of sub-stage awareness, difficulty in modeling task complexities, and inadequate object state estimation. To address these challenges,
we introduce VICtoR, a novel hierarchical VIC reward model capable of providing effective reward signals for long-horizon manipulation tasks. Trained solely on primitive motion demonstrations, VICtoR effectively provides precise reward signals for long-horizon tasks by assessing task progress at various stages using a novel stage detector and motion progress evaluator. We conducted extensive experiments in both simulated and real-world datasets. The results suggest that VICtoR outperformed the best existing methods, achieving a 43% improvement in success rates for long-horizon tasks. Our project page can be found at https://cmlab-victor.github.io/cmlab-vicotor.github.io/.
****

Adarsh Kumarappan,Mo Tiwari,Peiyang Song,Robert Joseph George,Chaowei Xiao,Anima Anandkumar
LeanAgent: Lifelong Learning for Formal Theorem Proving
Large Language Models (LLMs) have been successful in mathematical reasoning tasks such as formal theorem proving when integrated with interactive proof assistan

ts like Lean. Existing approaches involve training or fine-tuning an LLM on a sp
ecific dataset to perform well on particular domains, such as undergraduate-leve
l mathematics. These methods struggle with generalizability to advanced mathemat
ics. A fundamental limitation is that these approaches operate on static domains
, failing to capture how mathematicians often work across multiple domains and p
rojects simultaneously or cyclically. We present LeanAgent, a novel lifelong lea
rning framework for formal theorem proving that continuously generalizes to and
improves on ever-expanding mathematical knowledge without forgetting previously
learned knowledge. LeanAgent introduces several key innovations, including a cur
riculum learning strategy that optimizes the learning trajectory in terms of mat
hematical difficulty, a dynamic database for efficient management of evolving ma
thematical knowledge, and progressive training to balance stability and plastici
ty. LeanAgent successfully generates formal proofs for 155 theorems across 23 di
verse Lean repositories where formal proofs were previously missing, many from a
dvanced mathematics. It performs significantly better than the static LLM baseli
ne, proving challenging theorems in domains like abstract algebra and algebraic
topology while showcasing a clear progression of learning from basic concepts to
 advanced topics. In addition, we analyze LeanAgent's superior performance on ke
y lifelong learning metrics. LeanAgent achieves exceptional scores in stability
and backward transfer, where learning new tasks improves performance on previous
ly learned tasks. This emphasizes LeanAgent's continuous generalizability and im
provement, explaining its superior theorem-proving performance.
****

Yu Heng Hung,Kai-Jie Lin,Yu-Heng Lin,Chien-Yi Wang,Cheng Sun,Ping-Chun Hsieh
BOFormer: Learning to Solve Multi-Objective Bayesian Optimization via Non-Markov
ian RL
Bayesian optimization (BO) offers an efficient pipeline for optimizing black-box
 functions with the help of a Gaussian process prior and an acquisition function
 (AF). Recently, in the context of single-objective BO, learning-based AFs witne
ssed promising empirical results given its favorable non-myopic nature. Despite
this, the direct extension of these approaches to multi-objective Bayesian optim
ization (MOBO) suffer from the hypervolume identifiability issue, which results
from the non-Markovian nature of MOBO problems. To tackle this, inspired by the
non-Markovian RL literature and the success of Transformers in language modeling
, we present a generalized deep Q-learning framework and propose BOFormer, which
 substantiates this framework for MOBO via sequence modeling. Through extensive
evaluation, we demonstrate that BOFormer constantly achieves better performance
than the benchmark rule-based and learning-based algorithms in various synthetic
 MOBO and real-world multi-objective hyperparameter optimization problems.
****

Wen-Ding Li,Keya Hu,Carter Larsen,Yuqing Wu,Simon Alford,Caleb Woo,Spencer M. Du
nn,Hao Tang,Wei-Long Zheng,Yewen Pu,Kevin Ellis
Combining Induction and Transduction for Abstract Reasoning
When learning an input-output mapping from very few examples, is it better to fi
rst infer a latent function that explains the examples, or is it better to direc
tly predict new test outputs, e.g. using a neural network? We study this questio
n on ARC by training neural models for \emph{induction} (inferring latent functi
ons) and \emph{transduction} (directly predicting the test output for a given te
st input). We train
on synthetically generated variations of Python programs that solve ARC training
 tasks. We find inductive and transductive models solve different kinds of test
problems, despite having the same training problems and sharing the same neural
architecture: Inductive program synthesis excels at precise computations, and at
 composing multiple concepts, while transduction succeeds on fuzzier perceptual
concepts. Ensembling them approaches human-level performance on ARC.
****

Melis Ilayda Bal,Volkan Cevher,Michael Muehlebach
Adversarial Training for Defense Against Label Poisoning Attacks
As machine learning models grow in complexity and increasingly rely on publicly
sourced data, such as the human-annotated labels used in training large language

models, they become more vulnerable to label poisoning attacks.
These attacks, in which adversaries subtly alter the labels within a training dataset, can severely degrade model performance, posing significant risks in critical applications.
In this paper, we propose $\textbf{Floral}$, a novel adversarial training defense strategy based on support vector machines (SVMs) to counter these threats.
Utilizing a bilevel optimization framework, we cast the training process as a non-zero-sum Stackelberg game between an $\textit{attacker}$, who strategically poisons critical training labels, and the $\textit{model}$, which seeks to recover from such attacks.
Our approach accommodates various model architectures and employs a projected gradient descent algorithm with kernel SVMs for adversarial training.
We provide a theoretical analysis of our algorithm’s convergence properties and empirically evaluate $\textbf{Floral}$'s effectiveness across diverse classification tasks.
Compared to robust baselines and foundation models such as RoBERTa, $\textbf{Floral}$ consistently achieves higher robust accuracy under increasing attacker budgets.
These results underscore the potential of $\textbf{Floral}$ to enhance the resilience of machine learning models against label poisoning threats,
thereby ensuring robust classification in adversarial settings.
****
Pengfei Chen,Lingxi Xie,Xinyue Huo,Xuehui Yu,XIAOPENG ZHANG,Yingfei Sun,Zhenjun Han,Qi Tian
SAM-CP: Marrying SAM with Composable Prompts for Versatile Segmentation
The Segment Anything model (SAM) has shown a generalized ability to group image pixels into patches, but applying it to semantic-aware segmentation still faces major challenges. This paper presents SAM-CP, a simple approach that establishes two types of composable prompts beyond SAM and composes them for versatile segmentation. Specifically, given a set of classes (in texts) and a set of SAM patches, the Type-I prompt judges whether a SAM patch aligns with a text label, and the Type-II prompt judges whether two SAM patches with the same text label also belong to the same instance. To decrease the complexity in dealing with a large number of semantic classes and patches, we establish a unified framework that calculates the affinity between (semantic and instance) queries and SAM patches, and then merges patches with high affinity to the query. Experiments show that SAM-CP achieves semantic, instance, and panoptic segmentation in both open and closed domains. In particular, it achieves state-of-the-art performance in open-vocabulary segmentation. Our research offers a novel and generalized methodology for equipping vision foundation models like SAM with multi-grained semantic perception abilities. Codes are released on https://github.com/ucas-vg/SAM-CP.
****
Mu Cai,Jianwei Yang,Jianfeng Gao,Yong Jae Lee
Matryoshka Multimodal Models
Large Multimodal Models (LMMs) such as LLaVA have shown strong performance in visual-linguistic reasoning. These models first embed images into a fixed large number of visual tokens and then feed them into a Large Language Model (LLM). However, this design causes an excessive number of tokens for dense visual scenarios such as high-resolution images and videos, leading to great inefficiency. While token pruning/merging methods do exist, they produce a single length output for each image and do not afford flexibility in trading off information density v.s. efficiency. Inspired by the concept of Matryoshka Dolls, we propose : Matryoshka Multimodal Models, which learns to represent visual content as nested sets of visual tokens that capture information across multiple coarse-to-fine granularities. Our approach offers several unique benefits for LMMs: (1) One can explicitly control the visual granularity per test instance during inference, e.g. , adjusting the number of tokens used to represent an image based on the anticipated complexity or simplicity of the content; (2) provides a framework for analyzing the granularity needed for existing datasets, where we find that COCO-style benchmarks only need around 9 visual tokens to obtain accuracy similar to that of u

sing all 576 tokens; (3) Our approach provides a foundation to explore the best trade-off between performance and visual token length at sample level, where our investigation reveals that a large gap exists between the oracle upper bound and current fixed-scale representations.
****
Bernd Frauenknecht,Devdutt Subhasish,Friedrich Solowjow,Sebastian Trimpe
On Rollouts in Model-Based Reinforcement Learning
Model-based reinforcement learning (MBRL) seeks to enhance data efficiency by learning a model of the environment and generating synthetic rollouts from it. However, accumulated model errors during these rollouts can distort the data distribution, negatively impacting policy learning and hindering long-term planning. Thus, the accumulation of model errors is a key bottleneck in current MBRL methods. We propose Infoprop, a model-based rollout mechanism that separates aleatoric from epistemic model uncertainty and reduces the influence of the latter on the data distribution. Further, Infoprop keeps track of accumulated model errors along a model rollout and provides termination criteria to limit data corruption. We demonstrate the capabilities of Infoprop in the Infoprop-Dyna algorithm, reporting state-of-the-art performance in Dyna-style MBRL on common MuJoCo benchmark tasks while substantially increasing rollout length and data quality.
****
CHAO WANG,Giulio Franzese,Alessandro Finamore,Massimo Gallo,Pietro Michiardi
Information Theoretic Text-to-Image Alignment
Diffusion models for Text-to-Image (T2I) conditional generation have recently achieved
tremendous success. Yet, aligning these models with user’s intentions still involves a
laborious trial-and-error process, and this challenging alignment problem has attracted
considerable attention from the research community. In this work, instead of relying on
fine-grained linguistic analyses of prompts, human annotation, or auxiliary vision-language
models, we use Mutual Information (MI) to guide model alignment. In brief, our method
uses self-supervised fine-tuning and relies on a point-wise MI estimation between prompts
and images to create a synthetic fine-tuning set for improving model alignment. Our
analysis indicates that our method is superior to the state-of-the-art, yet it only requires
the pre-trained denoising network of the T2I model itself to estimate MI, and a simple
fine-tuning strategy that improves alignment while maintaining image quality. Code available at https://github.com/Chao0511/mitune.
****
Wonho Bae,Danica J. Sutherland,Gabriel L. Oliveira
Uncertainty Herding: One Active Learning Method for All Label Budgets
Most active learning research has focused on methods which perform well when many labels are available, but can be dramatically worse than random selection when label budgets are small.
Other methods have focused on the low-budget regime, but do poorly as label budgets increase.
As the line between "low" and "high" budgets varies by problem,
this is a serious issue in practice.
We propose *uncertainty coverage*,
an objective which generalizes a variety of low- and high-budget objectives,
as well as natural, hyperparameter-light methods to smoothly interpolate between low- and high-budget regimes.
We call greedy optimization of the estimate Uncertainty Herding;
this simple method is computationally fast,

and we prove that it nearly optimizes the distribution-level coverage.
In experimental validation across a variety of active learning tasks,
our proposal matches or beats state-of-the-art performance in essentially all cases;
it is the only method of which we are aware that reliably works well in both low- and high-budget settings.
****
Yifei Ming,Senthil Purushwalkam,Shrey Pandit,Zixuan Ke,Xuan-Phi Nguyen,Caiming Xiong,Shafiq Joty
FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows"
Ensuring faithfulness to context in large language models (LLMs) and retrieval-augmented generation (RAG) systems is crucial for reliable deployment in real-world applications, as incorrect or unsupported information can erode user trust. Despite advancements on standard benchmarks, faithfulness hallucination—where models generate responses misaligned with the provided context—remains a significant challenge. In this work, we introduce FaithEval, a novel and comprehensive benchmark tailored to evaluate the faithfulness of LLMs in contextual scenarios across three diverse tasks: unanswerable, inconsistent, and counterfactual contexts. These tasks simulate real-world challenges where retrieval mechanisms may surface incomplete, contradictory, or fabricated information. FaithEval comprises 4.9K high-quality problems in total, validated through a rigorous four-stage context construction and validation framework, employing both LLM-based auto-evaluation and human validation. Our extensive study across a wide range of open-source and proprietary models reveals that even state-of-the-art models often struggle to remain faithful to the given context, and that larger models do not necessarily exhibit improved faithfulness. Code is available at: https://github.com/SalesforceAIResearch/FaithEval.
****
Yuepeng Hu,Zhengyuan Jiang,Moyang Guo,Neil Zhenqiang Gong
A Transfer Attack to Image Watermarks
Watermark has been widely deployed by industry to detect AI-generated images. The robustness of such watermark-based detector against evasion attacks in the white-box and black-box settings is well understood in the literature. However, the robustness in the no-box setting is much less understood. In this work, we propose a new transfer evasion attack to image watermark in the no-box setting. Our transfer attack adds a perturbation to a watermarked image to evade multiple surrogate watermarking models trained by the attacker itself, and the perturbed watermarked image also evades the target watermarking model. Our major contribution is to show that, both theoretically and empirically, watermark-based AI-generated image detector based on existing watermarking methods is not robust to evasion attacks even if the attacker does not have access to the watermarking model nor the detection API. Our code is available at: https://github.com/hifi-hyp/Watermark-Transfer-Attack.
****
Kairui Yang,Zihao Guo,Gengjie Lin,Haotian Dong,Zhao Huang,Yipeng Wu,Die Zuo,Jibin Peng,Ziyuan Zhong,Xin WANG,Qing Guo,Xiaosong Jia,Junchi Yan,Di Lin
Trajectory-LLM: A Language-based Data Generator for Trajectory Prediction in Autonomous Driving
Vehicle trajectory prediction is a crucial aspect of autonomous driving, which requires extensive trajectory data to train prediction models to understand the complex, varied, and unpredictable patterns of vehicular interactions. However, acquiring real-world data is expensive, so we advocate using Large Language Models (LLMs) to generate abundant and realistic trajectories of interacting vehicles efficiently. These models rely on textual descriptions of vehicle-to-vehicle interactions on a map to produce the trajectories. We introduce Trajectory-LLM (Traj-LLM), a new approach that takes brief descriptions of vehicular interactions as input and generates corresponding trajectories. Unlike language-based approaches that translate text directly to trajectories, Traj-LLM uses reasonable driving behaviors to align the vehicle trajectories with the text. This results in an

"interaction-behavior-trajectory" translation process. We have also created a new dataset, Language-to-Trajectory (L2T), which includes 240K textual descriptions of vehicle interactions and behaviors, each paired with corresponding map topologies and vehicle trajectory segments. By leveraging the L2T dataset, Traj-LLM can adapt interactive trajectories to diverse map topologies. Furthermore, Traj-LLM generates additional data that enhances downstream prediction models, leading to consistent performance improvements across public benchmarks. The source code is released at https://github.com/TJU-IDVLab/Traj-LLM.
****
Hengyu Fu,Zihao Wang,Eshaan Nichani,Jason D. Lee
Learning Hierarchical Polynomials of Multiple Nonlinear Features
In deep learning theory, a critical question is to understand how neural networks learn hierarchical features. In this work, we study the learning of hierarchical polynomials of multiple nonlinear features using three-layer neural networks. We examine a broad class of functions of the form $f^{\star}=g^{\star}\circ \mathbf{p}$, where $\mathbf{p}:\mathbb{R}^{d} \rightarrow \mathbb{R}^{r}$ represents multiple quadratic features with $r \ll d$ and $g^{\star}:\mathbb{R}^{r}\rightarrow \mathbb{R}$ is a polynomial of degree p. This can be viewed as a nonlinear generalization of the multi-index model, and also an expansion upon previous work on nonlinear feature learning that focused only on a single feature (i.e. $r = 1$).
Our primary contribution shows that a three-layer neural network trained via layerwise gradient descent suffices for
- complete recovery of the space spanned by the nonlinear features
- efficient learning of the target function $f^{\star}=g^{\star}\circ \mathbf{p}$ or transfer learning of $f=g\circ \mathbf{p}$ with a different link function

within $\widetilde{\mathcal{O}}(d^4)$ samples and polynomial time.
For such hierarchical targets, our result substantially improves the sample complexity ${\Theta}(d^{2p})$ of the kernel methods, demonstrating the power of efficient feature learning. It is important to highlight that our results leverage novel techniques and thus manage to go beyond all prior settings such as single-index and multi-index models as well as models depending just on one nonlinear feature, contributing to a more comprehensive understanding of feature learning in deep learning.
****
Sharath Matada,Luke Bhan,Yuanyuan Shi,Nikolay Atanasov
Generalizable Motion Planning via Operator Learning
In this work, we introduce a planning neural operator (PNO) for predicting the value function of a motion planning problem. We recast value function approximation as learning a single operator from the cost function space to the value function
space, which is defined by an Eikonal partial differential equation (PDE). Therefore, our PNO model, despite being trained with a finite number of samples at coarse resolution, inherits the zero-shot super-resolution property of neural operators. We demonstrate accurate value function approximation at 16× the training resolution on the MovingAI lab’s 2D city dataset, compare with state-of-the-art neural value
function predictors on 3D scenes from the iGibson building dataset and showcase optimal planning with 4-joint robotic manipulators. Lastly, we investigate employing the value function output of PNO as a heuristic function to accelerate motion planning. We show theoretically that the PNO heuristic is ϵ-consistent by introducing an inductive bias layer that guarantees our value functions satisfy the triangle inequality. With our heuristic, we achieve a 30% decrease in nodes visited while obtaining near optimal path lengths on the MovingAI lab 2D city dataset, compared to classical planning methods (A*, RRT*).
****
Jianing Chu,Shu Yang,Wenbin Lu,PULAK GHOSH
Efficient Causal Decision Making with One-sided Feedback
We study a class of decision-making problems with one-sided feedback, where outc

omes are only observable for specific actions. A typical example is bank loans, where the repayment status is known only if a loan is approved and remains undefined if rejected. In such scenarios, conventional approaches to causal decision evaluation and learning from observational data are not directly applicable. In this paper, we introduce a novel value function to evaluate decision rules that addresses the issue of undefined counterfactual outcomes. Without assuming no unmeasured confounders, we establish the identification of the value function using shadow variables. Furthermore, leveraging semiparametric theory, we derive the efficiency bound for the proposed value function and develop efficient methods for decision evaluation and learning. Numerical experiments and a real-world data application demonstrate the empirical performance of our proposed methods.
****

Yi Zeng,Yu Yang,Andy Zhou,Jeffrey Ziwei Tan,Yuheng Tu,Yifan Mai,Kevin Klyman,Minzhou Pan,Ruoxi Jia,Dawn Song,Percy Liang,Bo Li
AIR-BENCH 2024: A Safety Benchmark based on Regulation and Policies Specified Risk Categories
Foundation models (FMs) provide societal benefits but also amplify risks. Governments, companies, and researchers have proposed regulatory frameworks, acceptable use policies, and safety benchmarks in response. However, existing public benchmarks often define safety categories based on previous literature, intuitions, or common sense, leading to disjointed sets of categories for risks specified in recent regulations and policies, which makes it challenging to evaluate and compare FMs across these benchmarks. To bridge this gap, we introduce AIR-BENCH 2024, the first AI safety benchmark aligned with emerging government regulations and company policies, following the regulation-based safety categories grounded in the AI Risks taxonomy, AIR 2024. AIR 2024 decomposes 8 government regulations and 16 company policies into a four-tiered safety taxonomy with 314 granular risk categories in the lowest tier. AIR-BENCH 2024 contains 5,694 diverse prompts spanning these categories, with manual curation and human auditing to ensure quality. We evaluate leading language models on AIR-BENCH 2024 uncovering insights into their alignment with specified safety concerns. By bridging the gap between public benchmarks and practical AI risks, AIR-BENCH 2024 provides a foundation for assessing model safety across jurisdictions, fostering the development of safer and more responsible AI systems.
****

Johannes Von Oswald,Seijin Kobayashi,Yassir Akram,Angelika Steger
Learning Randomized Algorithms with Transformers
Randomization is a powerful tool that endows algorithms with remarkable properties. For instance, randomized algorithms excel in adversarial settings, often surpassing the worst-case performance of deterministic algorithms with large margins. Furthermore, their success probability can be amplified by simple strategies such as repetition and majority voting. In this paper, we enhance deep neural networks, in particular transformer models, with randomization. We demonstrate for the first time that randomized algorithms can be instilled in transformers through learning, in a purely data- and objective-driven manner. First, we analyze known adversarial objectives for which randomized algorithms offer a distinct advantage over deterministic ones. We then show that common optimization techniques, such as gradient descent or evolutionary strategies, can effectively learn transformer parameters that make use of the randomness provided to the model. To illustrate the broad applicability of randomization in empowering neural networks, we study three conceptual tasks: associative recall, graph coloring, and agents that explore grid worlds. In addition to demonstrating increased robustness against oblivious adversaries through learned randomization, our experiments reveal remarkable performance improvements due to the inherently random nature of the neural networks' computation and predictions.
****

Bowei He,Lihao Yin,Huiling Zhen,Jianping Zhang,Lanqing HONG,Mingxuan Yuan,Chen Ma
Certifying Language Model Robustness with Fuzzed Randomized Smoothing: An Efficient Defense Against Backdoor Attacks

The widespread deployment of pre-trained language models (PLMs) has exposed them to textual backdoor attacks, particularly those planted during the pre-training stage. These attacks pose significant risks to high-reliability applications, as they can stealthily affect multiple downstream tasks. While certifying robustness against such threats is crucial, existing defenses struggle with the high-dimensional, interdependent nature of textual data and the lack of access to original poisoned pre-training data. To address these challenges, we introduce **F**uzzed **R**andomized **S**moothing (**FRS**), a novel approach for efficiently certifying language model robustness against backdoor attacks. FRS integrates software robustness certification techniques with biphased model parameter smoothing, employing Monte Carlo tree search for proactive fuzzing to identify vulnerable textual segments within the Damerau-Levenshtein space. This allows for targeted and efficient text randomization, while eliminating the need for access to poisoned training data during model smoothing. Our theoretical analysis demonstrates that FRS achieves a broader certified robustness radius compared to existing methods. Extensive experiments across various datasets, model configurations, and attack strategies validate FRS's superiority in terms of defense efficiency, accuracy, and robustness.
****

Haoyuan WU,Haisheng Zheng,Yuan Pu,Bei Yu
Circuit Representation Learning with Masked Gate Modeling and Verilog-AIG Alignment
Understanding the structure and function of circuits is crucial for electronic design automation (EDA).
Circuits can be formulated as And-Inverter graphs (AIGs), enabling efficient implementation of representation learning through graph neural networks (GNNs).
Masked modeling paradigms have been proven effective in graph representation learning.
However, masking augmentation to original circuits will destroy their logical equivalence, which is unsuitable for circuit representation learning.
Moreover, existing masked modeling paradigms often prioritize structural information at the expense of abstract information such as circuit function.
To address these limitations, we introduce MGVGA, a novel constrained masked modeling paradigm incorporating masked gate modeling (MGM) and Verilog-AIG alignment (VGA).
Specifically, MGM preserves logical equivalence by masking gates in the latent space rather than in the original circuits, subsequently reconstructing the attributes of these masked gates.
Meanwhile, large language models (LLMs) have demonstrated an excellent understanding of the Verilog code functionality.
Building upon this capability, VGA performs masking operations on original circuits and reconstructs masked gates under the constraints of equivalent Verilog codes, enabling GNNs to learn circuit functions from LLMs.
We evaluate MGVGA on various logic synthesis tasks for EDA and show the superior performance of MGVGA compared to previous state-of-the-art methods.
Our code is available at https://github.com/wuhy68/MGVGA.
****

Jasper Dekoninck,Maximilian Baader,Martin Vechev
Polyrating: A Cost-Effective and Bias-Aware Rating System for LLM Evaluation
Rating-based human evaluation has become an essential tool to accurately evaluate the impressive performance of large language models (LLMs). However, current rating systems suffer from several important limitations: first, they fail to account for biases that significantly influence evaluation results, second, they require large and expensive preference datasets to obtain accurate ratings, and third, they do not facilitate meaningful comparisons of model ratings across different tasks. To address these issues, we introduce Polyrating, an expressive and flexible rating system based on maximum a posteriori estimation that enables a more nuanced and thorough analysis of model performance at lower costs. Polyrating can detect and quantify biases affecting human preferences, ensuring fairer model comparisons. Further, Polyrating can reduce the cost of human evaluations by

up to 41% for new models and up to 77% for new tasks by leveraging existing benchmark scores. Lastly, Polyrating enables direct comparisons of ratings acro ss different tasks, providing a comprehensive understanding of an LLMs' strength s, weaknesses, and relative performance across different applications.
****

Shaolei Zhang,Qingkai Fang,Zhe Yang,Yang Feng
LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision To ken
The advent of real-time large multimodal models (LMMs) like GPT-4o has sparked c onsiderable interest in efficient LMMs. LMM frameworks typically encode visual i nputs into vision tokens (continuous representations) and integrate them and tex tual instructions into the context of large language models (LLMs), where large-scale parameters and numerous context tokens (predominantly vision tokens) resul t in substantial computational overhead. Previous efforts towards efficient LMMs always focus on replacing the LLM backbone with smaller models, while neglectin g the crucial issue of token quantity. In this paper, we introduce LLaVA-Mini, a n efficient LMM with minimal vision tokens. To achieve a high compression ratio of vision tokens while preserving visual information, we first analyze how LMMs understand vision tokens and find that most vision tokens only play a crucial ro le in the early layers of LLM backbone, where they mainly fuse visual informatio n into text tokens. Building on this finding, LLaVA-Mini introduces modality pre -fusion to fuse visual information into text tokens in advance, thereby facilita ting the extreme compression of vision tokens fed to LLM backbone into one token . LLaVA-Mini is a unified large multimodal model that can support the understand ing of images, high-resolution images, and videos in an efficient manner. Experi ments across 11 image-based and 7 video-based benchmarks demonstrate that LLaVA-Mini outperforms LLaVA-v1.5 with just 1 vision token instead of 576. Efficiency analyses reveal that LLaVA-Mini can reduce FLOPs by 77%, deliver low-latency res ponses within 40 milliseconds, and process over 10,000 frames of video on the GP U hardware with 24GB of memory.
****

Gyoung S. Na,Chanyoung Park
Self-Supervised Diffusion Models for Electron-Aware Molecular Representation Lea rning
Physical properties derived from electronic distributions are essential informat ion that determines molecular properties. However, the electron-level informatio n is not accessible in most real-world complex molecules due to the extensive co mputational costs of determining uncertain electronic distributions. For this re ason, existing methods for molecular property prediction have remained in regres sion models on simplified atom-level molecular descriptors, such as atomic struc tures and fingerprints. This paper proposes an efficient knowledge transfer meth od for electron-aware molecular representation learning. To this end, we devised a self-supervised diffusion method that estimates the electron-level informatio n of real-world complex molecules without expensive quantum mechanical calculati ons. The proposed method achieved state-of-the-art prediction accuracy in the ta sks of predicting molecular properties on extensive real-world molecular dataset s.
****

Yifan Du,Yuqi Huo,Kun Zhou,Zijia Zhao,Haoyu Lu,Han Huang,Xin Zhao,Bingning Wang, weipeng chen,Ji-Rong Wen
Exploring the Design Space of Visual Context Representation in Video MLLMs
Video Multimodal Large Language Models~(MLLMs) have shown remarkable capability of understanding the video semantics on various downstream tasks. Despite the ad vancements, there is still a lack of systematic research on visual context repre sentation, which refers to the scheme to select frames from a video and further select the tokens from a frame. In this paper, we explore the design space for v isual context representation, and aim to improve the performance of video MLLMs by finding more effective representation schemes. Firstly, we formulate the task of visual context representation as a constrained optimization problem, and mod el the language modeling loss as a function of the number of frames and the numb

er of embeddings (or tokens) per frame, given the maximum visual context window size. Then, we explore the scaling effects in frame selection and token selection respectively, and fit the corresponding function curve by conducting extensive empirical experiments. We examine the effectiveness of typical selection strategies and present empirical findings to determine the two factors. Furthermore, we study the joint effect of frame selection and token selection, and derive the optimal formula for determining the two factors. We demonstrate that the derived optimal settings show alignment with the best-performed results of empirical experiments. The data and code are available at: https://github.com/RUCAIBox/Opt-Visor.
****

Wenjing Chen,Shuo Xing,Samson Zhou,Victoria G. Crawford
Fair Submodular Cover
Machine learning algorithms are becoming increasing prevalent in the modern world, and as a result there has been significant recent study into algorithmic fairness in order to minimize the possibility of unintentional bias or discrimination in these algorithms. Submodular optimization problems also arise in many machine learning applications, including those such as data summarization and clustering where fairness is an important concern. In this paper, we initiate the study of the Fair Submodular Cover Problem (FSC). Given a ground set U, a monotone submodular function $f:2^U\to\mathbb{R}_{\ge 0}$, and a threshold τ, the goal of FSC is to find a balanced subset of U with minimum cardinality such that $f(S)\ge\tau$. We first introduce discrete algorithms for FSC that achieve a bicriteria approximation ratio of $(\frac{1}{\varepsilon}, 1-O(\varepsilon))$. We then present a continuous algorithm that achieves a $(\ln\frac{1}{\varepsilon}, 1-O(\varepsilon))$-bicriteria approximation ratio, which matches the best approximation guarantee of submodular cover without a fairness constraint. Finally, we complement our theoretical results with a number of empirical evaluations that demonstrate the efficiency of our algorithms on instances of maximum coverage.
****

Guneet S. Dhillon,Xingjian Shi,Yee Whye Teh,Alex Smola
L3Ms — Lagrange Large Language Models
Supervised fine-tuning (SFT) and alignment of large language models (LLMs) are key steps in providing a good user experience. However, the concept of an appropriate alignment is inherently application-dependent, and current methods often rely on heuristic choices to drive optimization. In this work, we formulate SFT and alignment as a constrained optimization problem: the LLM is fine-tuned on a task while being required to meet application-specific requirements, without resorting to heuristics. To solve this, we propose Lagrange Large Language Models (L3Ms), which employ logarithmic barriers to enforce the constraints. This approach allows for the customization of L3Ms across diverse applications while avoiding heuristic-driven processes. We experimentally demonstrate the versatility and efficacy of L3Ms in achieving tailored alignments for various applications.
****

Yining Hong,Beide Liu,Maxine Wu,Yuanhao Zhai,Kai-Wei Chang,Linjie Li,Kevin Lin,Chung-Ching Lin,Jianfeng Wang,Zhengyuan Yang,Ying Nian Wu,Lijuan Wang
SlowFast-VGen: Slow-Fast Learning for Action-Driven Long Video Generation
Human beings are endowed with a complementary learning system, which bridges the slow learning of general world dynamics with fast storage of episodic memory from a new experience. Previous video generation models, however, primarily focus on slow learning by pre-training on vast amounts of data, overlooking the fast learning phase crucial for episodic memory storage. This oversight leads to inconsistencies across temporally distant frames when generating longer videos, as these frames fall beyond the model's context window. To this end, we introduce SlowFast-VGen, a novel dual-speed learning system for action-driven long video generation. Our approach incorporates a masked conditional video diffusion model for the slow learning of world dynamics, alongside an inference-time fast learning strategy based on a temporal LoRA module. Specifically, the fast learning process updates its temporal LoRA parameters based on local inputs and outputs, thereby efficiently storing episodic memory in its parameters. We further propose a sl

ow-fast learning loop algorithm that seamlessly integrates the inner fast learni
ng loop into the outer slow learning loop, enabling the recall of prior multi-ep
isode experiences for context-aware skill learning. To facilitate the slow learn
ing of an approximate world model, we collect a large-scale dataset of 200k vide
os with language action annotations, covering a wide range of scenarios. Extensi
ve experiments show that SlowFast-VGen outperforms baselines across various metr
ics for action-driven video generation, achieving an FVD score of 514 compared t
o 782, and maintaining consistency in longer videos, with an average of 0.37 sce
ne cuts versus 0.89. The slow-fast learning loop algorithm significantly enhance
s performances on long-horizon planning tasks as well.
****

Drew Linsley,Peisen Zhou,Alekh Karkada Ashok,Akash Nagaraj,Gaurav Gaonkar,Franci
s E Lewis,Zygmunt Pizlo,Thomas Serre
The 3D-PC: a benchmark for visual perspective taking in humans and machines
Visual perspective taking (VPT) is the ability to perceive and reason about the
perspectives of others. It is an essential feature of human intelligence, which
develops over the first decade of life and requires an ability to process the 3D
 structure of visual scenes. A growing number of reports have indicated that dee
p neural networks (DNNs) become capable of analyzing 3D scenes after training on
 large image datasets. We investigated if this emergent ability for 3D analysis
in DNNs is sufficient for VPT with the 3D perception challenge (3D-PC): a novel
benchmark for 3D perception in humans and DNNs. The 3D-PC is comprised of three
3D-analysis tasks posed within natural scene images: (i.) a simple test of objec
t depth order, (ii.) a basic VPT task (VPT-basic), and (iii.) a more challenging
 version of VPT (VPT-perturb) designed to limit the effectiveness of "shortcut"
visual strategies. We tested human participants (N=33) and linearly probed or te
xt-prompted over 300 DNNs on the challenge and found that nearly all of the DNNs
 approached or exceeded human accuracy in analyzing object depth order. Surprisi
ngly, DNN accuracy on this task correlated with their object recognition perform
ance. In contrast, there was an extraordinary gap between DNNs and humans on VPT
-basic. Humans were nearly perfect, whereas most DNNs were near chance. Fine-tun
ing DNNs on VPT-basic brought them close to human performance, but they, unlike
humans, dropped back to chance when tested on VPT-perturb. Our challenge demonst
rates that the training routines and architectures of today's DNNs are well-suit
ed for learning basic 3D properties of scenes and objects but are ill-suited for
 reasoning about these properties like humans do. We release our 3D-PC datasets
and code to help bridge this gap in 3D perception between humans and machines.
****

Jaehun Jung,Faeze Brahman,Yejin Choi
Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement
We present a principled approach to provide LLM-based evaluation with a rigorous
 guarantee of human agreement. We first propose that a reliable evaluation metho
d should not uncritically rely on model preferences for pairwise evaluation, but
 rather assess the confidence of judge models and selectively decide when to tru
st its judgement. We then show that under this *selective evaluation* framework,
 human agreement can be provably guaranteed---such that the model evaluation ali
gns with that of humans to a user-specified agreement level. As part of our fram
ework, we also introduce *Simulated Annotators*, a novel confidence estimation m
ethod that significantly improves judge calibration and thus enables high covera
ge of evaluated instances. Finally, we propose *Cascaded Selective Evaluation*,
where we use cheaper models as initial judges and escalate to stronger models on
ly when necessary---again, while still providing a provable guarantee of human a
greement. Experimental results show that Cascaded Selective Evaluation guarantee
s strong alignment with humans, far beyond what LLM judges could achieve without
 selective evaluation. For example, on a subset of Chatbot Arena where GPT-4 alm
ost never achieves 80% human agreement, our method, even while employing substan
tially cost-effective models such as Mistral-7B, *guarantees* over 80% human agr
eement with almost 80% test coverage.
****

Tokio Kajitsuka,Issei Sato

On the Optimal Memorization Capacity of Transformers
Recent research in the field of machine learning has increasingly focused on the memorization capacity of Transformers, but how efficient they are is not yet well understood.
We demonstrate that Transformers can memorize labels with $\tilde{O}(\sqrt{N})$ parameters in a next-token prediction setting for N input sequences of length n, which is proved to be optimal up to logarithmic factors.
This indicates that Transformers can efficiently perform memorization with little influence from the input length n owing to the benefit of parameter sharing.
We also analyze the memorization capacity in the sequence-to-sequence setting, and find that $\tilde{O}(\sqrt{nN})$ parameters are not only sufficient, but also necessary at least for Transformers with hardmax.
These results suggest that while self-attention mechanisms can efficiently identify input sequences, the feed-forward network becomes a bottleneck when associating a label to each token.
****
Chenyang Zhu,Kai Li,Yue Ma,Longxiang Tang,Chengyu Fang,Chubin Chen,Qifeng Chen,Xiu Li
InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences
Recent advances in Customized Concept Swapping (CCS) enable a text-to-image model to swap a concept in the source image with a customized target concept. However, the existing methods still face the challenges of $\textit{\textbf{inconsistency}}$ and $\textit{\textbf{inefficiency}}$. They struggle to maintain consistency in both the foreground and background during concept swapping, especially when the shape difference is large between objects. Additionally, they either require time-consuming training processes or involve redundant calculations during inference. To tackle these issues, we introduce InstantSwap, a new CCS method that aims to handle sharp shape disparity at speed. Specifically, we first extract the bbox of the object in the source image $\textit{automatically}$ based on attention map analysis and leverage the bbox to achieve both foreground and background consistency. For background consistency, we remove the gradient outside the bbox during the swapping process so that the background is free from being modified. For foreground consistency, we employ a cross-attention mechanism to inject semantic information into both source and target concepts inside the box. This helps learn semantic-enhanced representations that encourage the swapping process to focus on the foreground objects. To improve swapping speed, we avoid computing gradients at each timestep but instead calculate them periodically to reduce the number of forward passes, which improves efficiency a lot with a little sacrifice on performance. Finally, we establish a benchmark dataset to facilitate comprehensive evaluation. Extensive evaluations demonstrate the superiority and versatility of InstantSwap.
****
Gresa Shala,André Biedenkapp,Pierre Krack,Florian Walter,Josif Grabocka
Efficient Cross-Episode Meta-RL
We introduce Efficient Cross-Episodic Transformers (ECET), a new algorithm for online Meta-Reinforcement Learning that addresses the challenge of enabling reinforcement learning agents to perform effectively in previously unseen tasks. We demonstrate how past episodes serve as a rich source of in-context information, which our model effectively distills and applies to new contexts. Our learned algorithm is capable of outperforming the previous state-of-the-art and provides more efficient meta-training while significantly improving generalization capabilities. Experimental results, obtained across various simulated tasks of the MuJoCo, Meta-World and ManiSkill benchmarks, indicate a significant improvement in learning efficiency and adaptability compared to the state-of-the-art. Our approach enhances the agent's ability to generalize from limited data and paves the way for more robust and versatile AI systems.
****
Masakazu Yoshimura,Teruaki Hayashi,Yota Maeda
MambaPEFT: Exploring Parameter-Efficient Fine-Tuning for Mamba
An ecosystem of Transformer-based models has been established by building large

models with extensive data.
Parameter-efficient fine-tuning (PEFT) is a crucial technology for deploying the se models to downstream tasks with minimal cost while achieving effective perfor mance. Recently, Mamba, a State Space Model (SSM)-based model, has attracted att ention as a potential alternative to Transformers. While many large-scale Mamba-based models have been proposed, efficiently adapting pre-trained Mamba-based mo dels to downstream tasks remains unexplored.
In this paper, we conduct an exploratory analysis of PEFT methods for Mamba. We investigate the effectiveness of existing PEFT methods for Transformers when app lied to Mamba. We also modify these methods to better align with the Mamba archi tecture. Additionally, we propose new Mamba-specific PEFT methods that leverage the distinctive structure of Mamba. Our experiments indicate that PEFT performs more effectively for Mamba than Transformers. Lastly, we demonstrate how to effe ctively combine multiple PEFT methods and provide a framework that outperforms p revious works. To ensure reproducibility, we will release the code after publica tion.
****

Maayan Ehrenberg,Roy Ganz,Nir Rosenfeld
Adversaries With Incentives: A Strategic Alternative to Adversarial Robustness
Adversarial training aims to defend against *adversaries*: malicious opponents w hose sole aim is to harm predictive performance in any way possible. This presen ts a rather harsh perspective, which we assert results in unnecessarily conserva tive training. As an alternative, we propose to model opponents as simply pursui ng their own goals—rather than working directly against the classifier. Employin g tools from strategic modeling, our approach enables knowledge or beliefs regar ding the opponent's possible incentives to be used as inductive bias for learnin g. Accordingly, our method of *strategic training* is designed to defend against all opponents within an `incentive uncertainty set'. This resorts to adversaria l learning when the set is maximal, but offers potential gains when the set can be appropriately reduced. We conduct a series of experiments that show how even mild knowledge regarding the opponent's incentives can be useful, and that the d egree of potential gains depends on how these incentives relate to the structure of the learning task.
****

Jonathan Yuyang Zhou,Yao Xie
Nonlinear Sequence Embedding by Monotone Variational Inequality
In the wild, we often encounter collections of sequential data such as electroca rdiograms, motion capture, genomes, and natural language, and sequences may be m ultichannel or symbolic with nonlinear dynamics. We introduce a method to learn low-dimensional representations of nonlinear sequence and time-series data witho ut supervision which has provable recovery guarantees. The learned representatio n can be used for downstream machine-learning tasks such as clustering and class ification. The method assumes that the observed sequences arise from a common do main, with each sequence following its own autoregressive model, and these model s are related through low-rank regularization. We cast the problem as a convex m atrix parameter recovery problem using monotone variational inequalities (VIs) a nd encode the common domain assumption via low-rank constraint across the learne d representations, which can learn a subspace approximately spanning the entire domain as well as faithful representations for the dynamics of each individual s equence incorporating the domain information in totality. We show the competitiv e performance of our method on real-world time-series data with baselines and de monstrate its effectiveness for symbolic text modeling and RNA sequence clusteri ng.
****

Han Li,Shaohui Li,Wenrui Dai,Maida Cao,Nuowen Kan,Chenglin Li,Junni Zou,Hongkai Xiong
On Disentangled Training for Nonlinear Transform in Learned Image Compression
Learned image compression (LIC) has demonstrated superior rate-distortion (R-D) performance compared to traditional codecs, but is challenged by training ineffi ciency that could incur more than two weeks to train a state-of-the-art model fr

om scratch. Existing LIC methods overlook the slow convergence caused by compact ing energy in learning nonlinear transforms. In this paper, we first reveal that such energy compaction consists of two components, \emph{i.e.}, feature decorre lation and uneven energy modulation. On such basis, we propose a linear auxiliar y transform (AuxT) to disentangle energy compaction in training nonlinear transf orms. The proposed AuxT obtains coarse approximation to achieve efficient energy compaction such that distribution fitting with the nonlinear transforms can be simplified to fine details. We then develop wavelet-based linear shortcuts (WLSs) for AuxT that leverages wavelet-based downsampling and orthogonal linear proje ction for feature decorrelation and subband-aware scaling for uneven energy modu lation. AuxT is lightweight and plug-and-play to be integrated into diverse LIC models to address the slow convergence issue. Experimental results demonstrate t hat the proposed approach can accelerate training of LIC models by 2 times and simultaneously achieves an average 1\% BD-rate reduction. To our best knowledge, this is one of the first successful attempt that can significantly improve the convergence of LIC with comparable or superior rate-distortion performance.
****

Xinting Huang,Andy Yang,Satwik Bhattamishra,Yash Sarrof,Andreas Krebs,Hattie Zho u,Preetum Nakkiran,Michael Hahn

A Formal Framework for Understanding Length Generalization in Transformers

A major challenge for transformers is generalizing to sequences longer than thos e observed during training. While previous works have empirically shown that tra nsformers can either succeed or fail at length generalization depending on the t ask, theoretical understanding of this phenomenon remains limited. In this work, we introduce a rigorous theoretical framework to analyze length generalization in causal transformers with learnable absolute positional encodings. In particul ar, we characterize those functions that are identifiable in the limit from suff iciently long inputs with absolute positional encodings under an idealized infer ence scheme using a norm-based regularizer. This enables us to prove the possibi lity of length generalization for a rich family of problems. We experimentally v alidate the theory as a predictor of success and failure of length generalizatio n across a range of algorithmic and formal language tasks. Our theory not only e xplains a broad set of empirical observations but also opens the way to provably predicting length generalization capabilities in transformers.
****

Tom Jacobs,Rebekka Burkholz

Mask in the Mirror: Implicit Sparsification

Continuous sparsification strategies are among the most effective methods for re ducing the inference costs and memory demands of large-scale neural networks. A key factor in their success is the implicit L_1 regularization induced by join tly learning both mask and weight variables, which has been shown experimentally to outperform explicit L_1 regularization. We provide a theoretical explanati on for this observation by analyzing the learning dynamics, revealing that early continuous sparsification is governed by an implicit L_2 regularization that gradually transitions to an L_1 penalty over time. Leveraging this insight, we propose a method to dynamically control the strength of this implicit bias. Thr ough an extension of the mirror flow framework, we establish convergence and opt imality guarantees in the context of underdetermined linear regression. Our theo retical findings may be of independent interest, as we demonstrate how to enter the rich regime and show that the implicit bias can be controlled via a time-dep endent Bregman potential. To validate these insights, we introduce PILoT, a cont inuous sparsification approach with novel initialization and dynamic regularizat ion, which consistently outperforms baselines in standard experiments.
****

Chen Chen,Yuchen Hu,Siyin Wang,Helin Wang,Zhehuai Chen,Chao Zhang,Chao-Han Huck Yang,EngSiong Chng

Audio Large Language Models Can Be Descriptive Speech Quality Evaluators

An ideal multimodal agent should be aware of the quality of its input modalities . Recent advances have enabled large language models (LLMs) to incorporate audit ory systems for handling various speech-related tasks. However, most audio LLMs

remain unaware of the quality of the speech they process. This limitation arises because speech quality evaluation is typically excluded from multi-task training due to the lack of suitable datasets. To address this, we introduce the first natural language-based speech evaluation corpus, generated from authentic human ratings. In addition to the overall Mean Opinion Score (MOS), this corpus offers detailed analysis across multiple dimensions and identifies causes of quality degradation. It also enables descriptive comparisons between two speech samples (A/B tests) with human-like judgment. Leveraging this corpus, we propose an alignment approach with LLM distillation (ALLD) to guide the audio LLM in extracting relevant information from raw speech and generating meaningful responses. Experimental results demonstrate that ALLD outperforms the previous state-of-the-art regression model in MOS prediction, with a mean square error of 0.17 and an A/B test accuracy of 98.6%. Additionally, the generated responses achieve BLEU scores of 25.8 and 30.2 on two tasks, surpassing the capabilities of task-specific models. This work advances the comprehensive perception of speech signals by audio LLMs, contributing to the development of real-world auditory and sensory intelligent agents.
****
Hongkai Zheng,Wenda Chu,Bingliang Zhang,Zihui Wu,Austin Wang,Berthy Feng,Caifeng Zou,Yu Sun,Nikola Borislavov Kovachki,Zachary E Ross,Katherine Bouman,Yisong Yue
InverseBench: Benchmarking Plug-and-Play Diffusion Priors for Inverse Problems in Physical Sciences
Plug-and-play diffusion priors (PnPDP) have emerged as a promising research direction for solving inverse problems. However, current studies primarily focus on natural image restoration, leaving the performance of these algorithms in scientific inverse problems largely unexplored. To address this gap, we introduce \textsc{InverseBench}, a framework that evaluates diffusion models across five distinct scientific inverse problems. These problems present unique structural challenges that differ from existing benchmarks, arising from critical scientific applications such as optical tomography, medical imaging, black hole imaging, seismology, and fluid dynamics. With \textsc{InverseBench}, we benchmark 14 inverse problem algorithms that use plug-and-play diffusion priors against strong, domain-specific baselines, offering valuable new insights into the strengths and weaknesses of existing algorithms. To facilitate further research and development, we open-source the codebase, along with datasets and pre-trained models, at https://devzhk.github.io/InverseBench/.
****
Haoyu Wang,Sunhao Dai,Haiyuan Zhao,Liang Pang,Xiao Zhang,Gang Wang,Zhenhua Dong,Jun Xu,Ji-Rong Wen
Perplexity Trap: PLM-Based Retrievers Overrate Low Perplexity Documents
Previous studies have found that PLM-based retrieval models exhibit a preference for LLM-generated content, assigning higher relevance scores to these documents even when their semantic quality is comparable to human-written ones. This phenomenon, known as source bias, threatens the sustainable development of the information access ecosystem. However, the underlying causes of source bias remain unexplored. In this paper, we explain the process of information retrieval with a causal graph and discover that PLM-based retrievers learn perplexity features for relevance estimation, causing source bias by ranking the documents with low perplexity higher. Theoretical analysis further reveals that the phenomenon stems from the positive correlation between the gradients of the loss functions in language modeling task and retrieval task. Based on the analysis, a causal-inspired inference-time debiasing method is proposed, called **C**ausal **D**iagnosis and **C**orrection (CDC). CDC first diagnoses the bias effect of the perplexity and then separates the bias effect from the overall estimated relevance score. Experimental results across three domains demonstrate the superior debiasing effectiveness of CDC, emphasizing the validity of our proposed explanatory framework. Source codes are available at https://github.com/WhyDwelledOnAi/Perplexity-Trap.
****

Razmik Arman Khosrovian,Takaharu Yaguchi,Hiroaki Yoshimura,Takashi Matsubara
Poisson-Dirac Neural Networks for Modeling Coupled Dynamical Systems across Domains
Deep learning has achieved great success in modeling dynamical systems, providing data-driven simulators to predict complex phenomena, even without known governing equations. However, existing models have two major limitations: their narrow focus on mechanical systems and their tendency to treat systems as monolithic. These limitations reduce their applicability to dynamical systems in other domains, such as electrical and hydraulic systems, and to coupled systems. To address these limitations, we propose Poisson-Dirac Neural Networks (PoDiNNs), a novel framework based on the Dirac structure that unifies the port-Hamiltonian and Poisson formulations from geometric mechanics. This framework enables a unified representation of various dynamical systems across multiple domains as well as their interactions and degeneracies arising from couplings. Our experiments demonstrate that PoDiNNs offer improved accuracy and interpretability in modeling unknown coupled dynamical systems from data.
****
Ziming Zhang,Fangzhou Lin,Haotian Liu,Jose Morales,Haichong Zhang,Kazunori Yamada,Vijaya B Kolachalama,Venkatesh Saligrama
GPS: A Probabilistic Distributional Similarity with Gumbel Priors for Set-to-Set Matching
Set-to-set matching aims to identify correspondences between two sets of unordered items by minimizing a distance metric or maximizing a similarity measure. Traditional metrics, such as Chamfer Distance (CD) and Earth Mover’s Distance (EMD), are widely used for this purpose but often suffer from limitations like suboptimal performance in terms of accuracy and robustness, or high computational costs - or both. In this paper, we propose a novel, simple yet effective set-to-set matching similarity measure, GPS, based on Gumbel prior distributions. These distributions are typically used to model the extrema of samples drawn from various distributions. Our approach is motivated by the observation that the distributions of minimum distances from CD, as encountered in real world applications such as point cloud completion, can be accurately modeled using Gumbel distributions. We validate our method on tasks like few-shot image classification and 3D point cloud completion, demonstrating significant improvements over state of-the-art loss functions across several benchmark datasets. Our demo code is publicly available at https://github.com/Zhang-VISLab/ICLR2025-GPS
****
Yuqi Yang,Peng-Tao Jiang,Qibin Hou,Hao Zhang,Jinwei Chen,Bo Li
Multi-Task Dense Predictions via Unleashing the Power of Diffusion
Diffusion models have exhibited extraordinary performance in dense prediction tasks. However, there are few works exploring the diffusion pipeline for multi-task dense predictions. In this paper, we unlock the potential of diffusion models in solving multi-task dense predictions and propose a novel diffusion-based method, called TaskDiffusion, which leverages the conditional diffusion process in the decoder. Instead of denoising the noisy labels for different tasks separately, we propose a novel joint denoising diffusion process to capture the task relations during denoising. To be specific, our method first encodes the task-specific labels into a task-integration feature space to unify the encoding strategy. This allows us to get rid of the cumbersome task-specific encoding process. In addition, we also propose a cross-task diffusion decoder conditioned on task-specific multi-level features, which can model the interactions among different tasks and levels explicitly while preserving efficiency. Experiments show that our TaskDiffusion outperforms previous state-of-the-art methods for all dense prediction tasks on the widely-used PASCAL-Context and NYUD-v2 datasets. Our code is available at https://github.com/YuqiYang213/TaskDiffusion.
****
Dongsung Huh
A Differentiable Metric for Discovering Groups and Unitary Representations
Discovering group structures in data poses a fundamental challenge across diverse scientific domains. The primary obstacle lies in the non-differentiable nature

of group axioms, impeding their integration into deep learning framework. To ov
ercome this, we present a novel differentiable approach leveraging the represent
ation theory of finite groups. Our method features a unique neural network archi
tecture that models interactions between group elements as multiplications of th
eir matrix representations, coupled with a regularizer that promotes unitarity o
f these matrices. Crucially, our model implicitly defines a complexity metric th
at favors the discovery of group structures. Evaluations demonstrate our method'
s ability to accurately recover group operations and learn their unitary represe
ntations from partial observation. Our work lays the foundation for a promising
new paradigm in automated algebraic structure discovery, with far-reaching appli
cations across diverse domains, particularly in enabling automatic symmetry disc
overy for geometric deep learning.
****
Elad Romanov,Fangzhao Zhang,Mert Pilanci
Newton Meets Marchenko-Pastur: Massively Parallel Second-Order Optimization with
Hessian Sketching and Debiasing
Motivated by recent advances in serverless cloud computing, in particular the ``
function as a service'' (FaaS) model,
we consider the problem of minimizing a convex function in a massively parallel
fashion, where communication between workers is limited.
Focusing on the case of a twice-differentiable objective subject to an L2 penalt
y, we propose a scheme where the central node (server) effectively runs a Newton
method,
offloading its high per-iteration cost---stemming from the need to invert the He
ssian---to the workers.
In our solution, workers produce independently coarse but low-bias estimates of
the inverse Hessian, using an adaptive sketching scheme. The server then average
s the descent directions produced by the workers, yielding a good approximation
for the exact Newton step. The main component of our adaptive sketching scheme i
s a low-complexity procedure for selecting the sketching dimension, an issue tha
t was left largely unaddressed in the existing literature on Hessian sketching f
or distributed optimization. Our solution is based on ideas from asymptotic rand
om matrix theory, specifically the Marchenko-Pastur law. For Gaussian sketching
matrices, we derive non asymptotic guarantees for our algorithm which do not dep
end on the condition number of the Hessian nor a priori require the sketching di
mension to be proportional to the dimension, as is often the case in asymptotic
random matrix theory. Lastly, when the objective is self-concordant, we provide
convergence guarantees for the approximate Newton's method with noisy Hessians,
which may be of independent interest beyond the setting considered in this paper
.
****
Qiushi Huang,Tom Ko,Zhan Zhuang,Lilian Tang,Yu Zhang
HiRA: Parameter-Efficient Hadamard High-Rank Adaptation for Large Language Model
s
We propose Hadamard High-Rank Adaptation (HiRA), a parameter-efficient fine-tuni
ng (PEFT) method that enhances the adaptability of Large Language Models (LLMs).
While Low-rank Adaptation (LoRA) is widely used to reduce resource demands, its
low-rank updates may limit its expressiveness for new tasks. HiRA addresses thi
s by using a Hadamard product to retain high-rank update parameters, improving t
he model capacity. Empirically, HiRA outperforms LoRA and its variants on severa
l tasks, with extensive ablation studies validating its effectiveness. Our code
is available at https://github.com/hqsiswiliam/hira.
****
Lucas Prieto,Melih Barsbey,Pedro A. M. Mediano,Tolga Birdal
Grokking at the Edge of Numerical Stability
Grokking, or sudden generalization that occurs after prolonged overfitting, is a
surprising phenomenon that has challenged our understanding of deep learning. W
hile a lot of progress has been made in understanding grokking, it is still not
clear why generalization is delayed and why grokking often does not happen witho
ut regularization. In this work we argue that without regularization, grokking t

asks push models to the edge of numerical stability, introducing floating point errors in the Softmax that we refer to as _Softmax Collapse_ (SC). We show that SC prevents grokking and that mitigating SC leads to grokking _without_ regularization. Investigating the root cause of SC, we find that beyond the point of overfitting, the gradients strongly align with what we call the _naïve loss minimization_ (NLM) direction. This component of the gradient does not change the predictions of the model but decreases the loss by scaling the logits, usually through the scaling of the weights along their current direction. We show that this scaling of the logits explains the delay in generalization characteristic of grokking, and eventually leads to SC, stopping learning altogether. To validate these hypotheses, we introduce two key contributions that mitigate the issues faced in grokking tasks: (i) $\mathrm{StableMax}$, a new activation function that prevents SC and enables grokking without regularization, and (ii) $\perp\mathrm{Grad}$, a training algorithm that leads to quick generalization in grokking tasks by preventing NLM altogether. These contributions provide new insights into grokking, shedding light on its delayed generalization, reliance on regularization, and the effectiveness of known grokking-inducing methods.
****

Yinpeng Chen,DeLesley Hutchins,Aren Jansen,Andrey Zhmoginov,David Racz,Jesper Sparre Andersen
MELODI: Exploring Memory Compression for Long Contexts
We present MELODI, a novel memory architecture designed to efficiently process long documents using short context windows. The key principle behind MELODI is to represent short-term and long-term memory as a hierarchical compression scheme across both transformer layers and context windows. Specifically, the short-term memory is achieved through recurrent compression of context windows across multiple layers, ensuring smooth transitions between windows. In contrast, the long-term memory performs further compression within a single middle layer and aggregates information across context windows, effectively consolidating crucial information from the entire history. Compared to a strong baseline - the Memorizing Transformer employing dense attention over a large long-term memory (64K key-value pairs) - our method demonstrates superior performance on various long-context datasets while remarkably reducing the memory footprint by a factor of 8.
****

Eric Lei,Hamed Hassani,Shirin Saeedi Bidokhti
Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding
Neural compression has brought tremendous progress in designing lossy compressors with good rate-distortion (RD) performance at low complexity. Thus far, neural compression design involves transforming the source to a latent vector, which is then rounded to integers and entropy coded. While this approach has been shown to be optimal on a few specific sources, we show that it can be highly sub-optimal on synthetic sources whose intrinsic dimensionality is greater than one. With integer rounding in the latent space, the quantization regions induced by neural transformations, remain square-like and fail to match those of optimal vector quantization. We demonstrate that this phenomenon is due to the choice of scalar quantization in the latent space, and not the transform design. By employing lattice quantization instead, we propose Lattice Transform Coding (LTC) and show that it approximately recovers optimal vector quantization at reasonable complexity. On real-world sources, LTC improves upon standard neural compressors. LTC also provides a framework that can integrate structurally (near) optimal information-theoretic designs into lossy compression; examples include block coding, which yields coding gain over optimal one-shot coding and approaches the asymptotically-achievable rate-distortion function, as well as nested lattice quantization for low complexity fixed-rate coding.
****

Syeda Nahida Akter,Shrimai Prabhumoye,John Kamalu,Sanjeev Satheesh,Eric Nyberg,Mostofa Patwary,Mohammad Shoeybi,Bryan Catanzaro
MIND: Math Informed syNthetic Dialogues for Pretraining LLMs
The utility of synthetic data to enhance pretraining data quality and hence to i

mprove downstream task accuracy has been widely explored in recent large languag
e models (LLMs). Yet, these approaches fall inadequate in complex, multi-hop and
 mathematical reasoning tasks as the synthetic data typically fails to add compl
ementary knowledge to the existing raw corpus. In this work, we propose a novel
large-scale and diverse Math Informed syNthetic Dialogue (MIND) generation metho
d that improves the mathematical reasoning ability of LLMs. Specifically, using
MIND, we generate synthetic conversations based on OpenWebMath (OWM), resulting
in a new math corpus, MIND-OWM. Our experiments with different conversational se
ttings reveal that incorporating knowledge gaps between dialog participants is e
ssential for generating high-quality math data. We further identify an effective
 way to format and integrate synthetic and raw data during pretraining to maximi
ze the gain in mathematical reasoning, emphasizing the need to restructure raw d
ata rather than use it as-is. Compared to pretraining just on raw data, a model
pretrained on MIND-OWM shows significant boost in mathematical reasoning (GSM8K:
 +13.42%, MATH: +2.30%), including superior performance in specialized knowledge
 (MMLU: +4.55%, MMLU-STEM: +4.28%) and general
purpose reasoning tasks (GENERAL REASONING: +2.51%).
****

Jeongsol Kim,Geon Yeong Park,Hyungjin Chung,Jong Chul Ye
Regularization by Texts for Latent Diffusion Inverse Solvers
The recent development of diffusion models has led to significant progress in so
lving inverse problems by leveraging these models as powerful generative priors.
 However, challenges persist due to the ill-posed nature of such problems, often
 arising from ambiguities in measurements or intrinsic system symmetries. To add
ress this, we introduce a novel latent diffusion inverse solver, regularization
by text (TReg), inspired by the human ability to resolve visual ambiguities thro
ugh perceptual biases. TReg integrates textual descriptions of preconceptions ab
out the solution during reverse diffusion sampling, dynamically reinforcing thes
e descriptions through null-text optimization, which we refer to as adaptive neg
ation. Our comprehensive experimental results demonstrate that TReg effectively
mitigates ambiguity in inverse problems, improving both accuracy and efficiency.
****

Nabarun Goswami,Hanqin Wang,Tatsuya Harada
T2V2: A Unified Non-Autoregressive Model for Speech Recognition and Synthesis vi
a Multitask Learning
We introduce T2V2 (**T**ext to **V**oice and **V**oice to **T**ext), a unified n
on-autoregressive model capable of performing both automatic speech recognition
(ASR) and text-to-speech (TTS) synthesis within the same framework. T2V2 uses a
shared Conformer backbone with rotary positional embeddings to efficiently handl
e these core tasks, with ASR trained using Connectionist Temporal Classification
 (CTC) loss and TTS using masked language modeling (MLM) loss. The model operate
s on discrete tokens, where speech tokens are generated by clustering features f
rom a self-supervised learning model. To further enhance performance, we introdu
ce auxiliary tasks: CTC error correction to refine raw ASR outputs using context
ual information from speech embeddings, and unconditional speech MLM, enabling c
lassifier free guidance to improve TTS. Our method is self-contained, leveraging
 intermediate CTC outputs to align text and speech using Monotonic Alignment Sea
rch, without relying on external aligners. We perform extensive experimental eva
luation to verify the efficacy of the T2V2 framework, achieving state-of-the-art
 performance on TTS task and competitive performance in discrete ASR.
****

Zhengqiang ZHANG,Ruihuang Li,Lei Zhang
FreCaS: Efficient Higher-Resolution Image Generation via Frequency-aware Cascade
d Sampling
While image generation with diffusion models has achieved a great success, gener
ating images of higher resolution than the training size remains a challenging t
ask due to the high computational cost. Current methods typically perform the en
tire sampling process at full resolution and process all frequency components si
multaneously, contradicting with the inherent coarse-to-fine nature of latent di
ffusion models and wasting computations on processing premature high-frequency d

etails at early diffusion stages. To address this issue, we introduce an efficie nt $\textbf{Fre}$quency-aware $\textbf{Ca}$scaded $\textbf{S}$ampling framework, $\textbf{FreCaS}$ in short, for higher-resolution image generation. FreCaS deco mposes the sampling process into cascaded stages with gradually increased resolu tions, progressively expanding frequency bands and refining the corresponding de tails. We propose an innovative frequency-aware classifier-free guidance (FA-CFG) strategy to assign different guidance strengths for different frequency compon ents, directing the diffusion model to add new details in the expanded frequency domain of each stage. Additionally, we fuse the cross-attention maps of previou s and current stages to avoid synthesizing unfaithful layouts. Experiments demon strate that FreCaS significantly outperforms state-of-the-art methods in image q uality and generation speed. In particular, FreCaS is about 2.86$\times$ and 6.0 7$\times$ faster than ScaleCrafter and DemoFusion in generating a 2048$\times$20 48 image using a pretrained SDXL model and achieves an FID_b improvemen t of 11.6 and 3.7, respectively. FreCaS can be easily extended to more complex m odels such as SD3. The source code of FreCaS can be found at https://github.com/ xtudbxk/FreCaS.
****

Yukang Cao,Liang Pan,Kai Han,Kwan-Yee K. Wong,Ziwei Liu
AvatarGO: Zero-shot 4D Human-Object Interaction Generation and Animation
Recent advancements in diffusion models have led to significant improvements in the generation and animation of 4D full-body human-object interactions (HOI). Ne vertheless, existing methods primarily focus on SMPL-based motion generation, wh ich is limited by the scarcity of realistic large-scale interaction data. This c onstraint affects their ability to create everyday HOI scenes. This paper addres ses this challenge using a zero-shot approach with a pre-trained diffusion model . Despite this potential, achieving our goals is difficult due to the diffusion model's lack of understanding of ''where'' and ''how'' objects interact with the human body. To tackle these issues, we introduce **AvatarGO**, a novel framewor k designed to generate animatable 4D HOI scenes directly from textual inputs. Sp ecifically, **1)** for the ''where'' challenge, we propose **LLM-guided contact retargeting**, which employs Lang-SAM to identify the contact body part from tex t prompts, ensuring precise representation of human-object spatial relations. ** 2)** For the ''how'' challenge, we introduce **correspondence-aware motion optim ization** that constructs motion fields for both human and object models using t he linear blend skinning function from SMPL-X. Our framework not only generates coherent compositional motions, but also exhibits greater robustness in handlin g penetration issues. Extensive experiments with existing methods validate Avata rGO's superior generation and animation capabilities on a variety of human-objec t pairs and diverse poses. As the first attempt to synthesize 4D avatars with ob ject interactions, we hope AvatarGO could open new doors for human-centric 4D co ntent creation.
****

Fei Wang,Xingyu Fu,James Y. Huang,Zekun Li,Qin Liu,Xiaogeng Liu,Mingyu Derek Ma, Nan Xu,Wenxuan Zhou,Kai Zhang,Tianyi Lorena Yan,Wenjie Jacky Mo,Hsiang-Hui Liu,P an Lu,Chunyuan Li,Chaowei Xiao,Kai-Wei Chang,Dan Roth,Sheng Zhang,Hoifung Poon,M uhao Chen
MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding
We introduce MuirBench, a comprehensive benchmark that focuses on robust multi-i mage understanding capabilities of multimodal LLMs. MuirBench consists of 12 div erse multi-image tasks (e.g., scene understanding, ordering) that involve 10 cat egories of multi-image relations (e.g., multiview, temporal relations). Comprisi ng 11,264 images and 2,600 multiple-choice questions, MuirBench is created in a pairwise manner, where each standard instance is paired with an unanswerable var iant that has minimal semantic differences, in order for a reliable assessment. Evaluated upon 20 recent multi-modal LLMs, our results reveal that even the best -performing models like GPT-4o and Gemini Pro find it challenging to solve MuirB ench, achieving 68.0% and 49.3% in accuracy. Open-source multimodal LLMs trained on single images can hardly generalize to multi-image questions, hovering below 33.3% in accuracy. These results highlight the importance of MuirBench in encou

raging the community to develop multimodal LLMs that can look beyond a single image, suggesting potential pathways for future improvements.
****
Suyu Ge,Xihui Lin,Yunan Zhang,Jiawei Han,Hao Peng
A Little Goes a Long Way: Efficient Long Context Training and Inference with Partial Contexts
Training and serving long-context large language models (LLMs) incurs substantial overhead.
To address this, two critical steps are often required: a pretrained LLM typically undergoes a separate stage for context length extension by training on long-context data, followed by architectural modifications to reduce the overhead of KV cache during serving.
This paper argues that integrating length extension with a GPU-friendly KV cache reduction architecture not only reduces training overhead during length extension, but also achieves better long-context performance.
This leads to our proposed LongGen, which finetunes a pretrained LLM into an efficient architecture during length extension.
LongGen builds on three key insights:
(1) Sparse attention patterns, such as window attention (attending to recent tokens), attention sink (initial ones), and blockwise sparse attention (strided token blocks) are well-suited for building efficient long-context models, primarily due to their GPU-friendly memory access patterns, enabling efficiency gains not just theoretically but in practice as well.
(2) It is essential for the model to have direct access to all tokens.
A hybrid architecture with 1/3 full attention layers and 2/3 efficient ones achieves a balanced trade-off between efficiency and long-context performance.
(3) Lightweight training on 5B long-context data is sufficient to extend the hybrid model's context length from 4K to 128K.

We evaluate LongGen on both Llama-2 7B and Llama-2 70B, demonstrating its effectiveness across different scales.
During training with 128K-long contexts, LongGen achieves 1.55x training speedup and reduces wall-clock time by 36%, compared to a full-attention baseline.
During inference, LongGen reduces KV cache memory by 62%, achieving 1.67x prefilling speedup and 1.41x decoding speedup.
Compared to baselines that apply KV-cache reduction techniques to full-attention long-context LLMs, LongGen achieves substantially stronger performance not only on the Needle-in-a-Haystack retrieval task, but also on more challenging long-context reasoning tasks, including BABILong and RULER.
****
Yury Demidovich,Petr Ostroukhov,Grigory Malinovsky,Samuel Horváth,Martin Taká■,Peter Richtárik,Eduard Gorbunov
Methods with Local Steps and Random Reshuffling for Generally Smooth Non-Convex Federated Optimization
Non-convex Machine Learning problems typically do not adhere to the standard smoothness assumption. Based on empirical findings, Zhang et al. (2020b) proposed a more realistic generalized (L_0,L_1)-smoothness assumption, though it remains largely unexplored. Many existing algorithms designed for standard smooth problems need to be revised. However, in the context of Federated Learning, only a few works address this problem but rely on additional limiting assumptions. In this paper, we address this gap in the literature: we propose and analyze new methods with local steps, partial participation of clients, and Random Reshuffling without extra restrictive assumptions beyond generalized smoothness. The proposed methods are based on the proper interplay between clients' and server's stepsizes and gradient clipping. Furthermore, we perform the first analysis of these methods under the Polyak-■ojasiewicz condition. Our theory is consistent with the known results for standard smooth problems, and our experimental results support the theoretical insights.
****
Stathis Megas,Daniel G. Chen,Krzysztof Polanski,Moshe Eliasof,Carola-Bibiane Sch

önlieb,Sarah A Teichmann
Estimation of single-cell and tissue perturbation effect in spatial transcriptomics via Spatial Causal Disentanglement
Models of Virtual Cells and Virtual Tissues at single-cell resolution would allow us to test perturbations in silico and accelerate progress in tissue and cell engineering.
However, most such models are not rooted in causal inference and as a result, could mistake correlation for causation.
We introduce Celcomen, a novel generative graph neural network grounded in mathematical causality to disentangle intra- and inter-cellular gene regulation in spatial transcriptomics and single-cell data.
Celcomen can also be prompted by perturbations to generate spatial counterfactuals, thus offering insights into experimentally inaccessible states, with potential applications in human health.
We validate the model's disentanglement and identifiability through simulations, and demonstrate its counterfactual predictions in clinically relevant settings, including human glioblastoma and fetal spleen, recovering inflammation-related gene programs post immune system perturbation.
Moreover, it supports mechanistic interpretability, as its parameters can be reverse-engineered from observed behavior, making it an accessible model for understanding both neural networks and complex biological systems.
****
Yuanzhao Zhang,William Gilpin
Zero-shot forecasting of chaotic systems
Time-series forecasting is a challenging problem that traditionally requires specialized models custom-trained for the specific task at hand. Recently, inspired by the success of large language models, foundation models pre-trained on vast amounts of time-series data from diverse domains have emerged as a promising candidate for general-purpose time-series forecasting. The defining characteristic of these foundation models is their ability to perform zero-shot learning, that is, forecasting a new system from limited context data without explicit re-training or fine-tuning. Here, we evaluate whether the zero-shot learning paradigm extends to the challenging task of forecasting chaotic systems. Across 135 distinct chaotic dynamical systems and 10^8 timepoints, we find that foundation models produce competitive forecasts compared to custom-trained models (including NBEATS, TiDE, etc.), particularly when training data is limited. Interestingly, even after point forecasts fail, large foundation models are able to preserve the geometric and statistical properties of the chaotic attractors.
We attribute this success to foundation models' ability to perform in-context learning and identify context parroting as a simple mechanism used by these models to capture the long-term behavior of chaotic dynamical systems. Our results highlight the potential of foundation models as a tool for probing nonlinear and complex systems.
****
Hao Luo,Zongqing Lu
Learning Video-Conditioned Policy on Unlabelled Data with Joint Embedding Predictive Transformer
The video-conditioned policy takes prompt videos of the desired tasks as a condition and is regarded for its prospective generalizability. Despite its promise, training a video-conditioned policy is non-trivial due to the need for abundant demonstrations. In some tasks, the expert rollouts are merely available as videos, and costly and time-consuming efforts are required to annotate action labels. To address this, we explore training video-conditioned policy on a mixture of demonstrations and unlabeled expert videos to reduce reliance on extensive manual annotation. We introduce the Joint Embedding Predictive Transformer (JEPT) to learn a video-conditioned policy through sequence modeling. JEPT is designed to jointly learn visual transition prediction and inverse dynamics. The visual transition is captured from both demonstrations and expert videos, on the basis of which the inverse dynamics learned from demonstrations is generalizable to the tasks without action labels. Experiments on a series of simulated visual control ta

sks evaluate that JEPT can effectively leverage the mixture dataset to learn a g
eneralizable policy. JEPT outperforms baselines in the tasks without action-labe
led data and unseen tasks. We also experimentally reveal the potential of JEPT a
s a simple visual priors injection approach to enhance the video-conditioned pol
icy.
****
Rafael Pastrana,Eder Medina,Isabel M. de Oliveira,Sigrid Adriaenssens,Ryan P Ada
ms
Real-time design of architectural structures with differentiable mechanics and n
eural networks
Designing mechanically efficient geometry for architectural structures like shel
ls, towers, and bridges, is an expensive iterative process.
Existing techniques for solving such inverse problems rely on traditional optimi
zation methods, which are slow and computationally expensive, limiting iteration
 speed and design exploration.
Neural networks would seem to offer a solution via data-driven amortized optimiz
ation, but they often require extensive fine-tuning and cannot ensure that impor
tant design criteria, such as mechanical integrity, are met.
In this work, we combine neural networks with a differentiable mechanics simulat
or to develop a model that accelerates the solution of shape approximation probl
ems for architectural structures represented as bar systems.
This model explicitly guarantees compliance with mechanical constraints while ge
nerating designs that closely match target geometries.
We validate our approach in two tasks, the design of masonry shells and cable-ne
t towers.
Our model achieves better accuracy and generalization than fully neural alternat
ives, and comparable accuracy to direct optimization but in real time, enabling
fast and reliable design exploration.
We further demonstrate its advantages by integrating it into 3D modeling softwar
e and fabricating a physical prototype.
Our work opens up new opportunities for accelerated mechanical design enhanced b
y neural networks for the built environment.
****
Jian-Qiao Zhu,Haijiang Yan,Thomas L. Griffiths
Language Models Trained to do Arithmetic Predict Human Risky and Intertemporal C
hoice
The observed similarities in the behavior of humans and Large Language Models (L
LMs) have prompted researchers to consider the potential of using LLMs as models
 of human cognition. However, several significant challenges must be addressed b
efore LLMs can be legitimately regarded as cognitive models. For instance, LLMs
are trained on far more data than humans typically encounter, and may have been
directly trained on human data in specific cognitive tasks or aligned with human
 preferences. Consequently, the origins of these behavioral similarities are not
 well understood. In this paper, we propose a novel way to enhance the utility o
f language models as cognitive models. This approach involves (i) leveraging com
putationally equivalent tasks that both a language model and a rational agent ne
ed to master for solving a cognitive problem and (ii) examining the specific tas
k distributions required for a language model to exhibit human-like behaviors. W
e apply this approach to decision-making -- specifically risky and intertemporal
 choice -- where the key computationally equivalent task is the arithmetic of ex
pected value calculations. We show that a small language model pretrained on an
ecologically valid arithmetic dataset, which we call Arithmetic-GPT, predicts hu
man behavior better than many traditional cognitive models. Pretraining language
 models on ecologically valid arithmetic datasets is sufficient to produce a str
ong correspondence between these models and human decision-making. Our results a
lso suggest that language models used as cognitive models should be carefully in
vestigated via ablation studies of the pretraining data.
****
Zechen Bai,Tianjun Xiao,Tong He,Pichao WANG,Zheng Zhang,Thomas Brox,Mike Zheng S
hou

Bridging Information Asymmetry in Text-video Retrieval: A Data-centric Approach
As online video content rapidly grows, the task of text-video retrieval (TVR) be comes increasingly important. A key challenge in TVR is the information asymmetr y between video and text: videos are inherently richer in information, while the ir textual descriptions often capture only fragments of this complexity. This pa per introduces a novel, data-centric framework to bridge this gap by enriching t extual representations to better match the richness of video content. During tra ining, videos are segmented into event-level clips and captioned to ensure compr ehensive coverage. During retrieval, a large language model (LLM) generates sema ntically diverse queries to capture a broader range of possible matches. To enha nce retrieval efficiency, we propose a query selection mechanism that identifies the most relevant and diverse queries, reducing computational cost while improv ing accuracy. Our method achieves state-of-the-art results across multiple bench marks, demonstrating the power of data-centric approaches in addressing informat ion asymmetry in TVR. This work paves the way for new research focused on levera ging data to improve cross-modal retrieval.
****
Tian Ye,Zicheng Xu,Yuanzhi Li,Zeyuan Allen-Zhu
Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process
Recent advances in language models have demonstrated their capability to solve m athematical reasoning problems, achieving near-perfect accuracy on grade-school level math benchmarks like GSM8K. In this paper, we formally study how language models solve these problems. We design a series of controlled experiments to add ress several fundamental questions: (1) Can language models truly develop reason ing skills, or do they simply memorize templates? (2) What is the model's hidden (mental) reasoning process? (3) Do models solve math questions using skills sim ilar to or different from humans? (4) Do models trained on GSM8K-like datasets d evelop reasoning skills beyond those necessary for solving GSM8K problems? (5) W hat mental process causes models to make reasoning mistakes? (6) How large or de ep must a model be to effectively solve GSM8K-level math questions?

Our study uncovers many hidden mechanisms by which language models solve mathema tical questions, providing insights that extend beyond current understandings of LLMs.
****
Varun Khurana,Yaman Kumar Singla,Jayakumar Subramanian,Changyou Chen,Rajiv Ratn Shah,zhiqiang xu,Balaji Krishnamurthy
Measuring And Improving Engagement of Text-to-Image Generation Models
Recent advances in text-to-image generation have achieved impressive aesthetic q uality, making these models usable for both personal and commercial purposes. Ho wever, in the fields of marketing and advertising, images are often created to b e more engaging, as reflected in user behaviors such as increasing clicks, likes , and purchases, in addition to being aesthetically pleasing. To this end, we in troduce the challenge of optimizing the image generation process for improved vi ewer engagement. In order to study image engagement and utility in real-world ma rketing scenarios, we collect *EngagingImageNet*, the first large-scale dataset of images, along with associated user engagement metrics. Further, we find that existing image evaluation metrics like aesthetics, CLIPScore, PickScore, ImageRe ward, *etc.* are unable to capture viewer engagement. To address the lack of rel iable metrics for assessing image utility, we use the *EngagingImageNet* dataset to train *EngageNet*, an engagement-aware Vision Language Model (VLM) that pred icts viewer engagement of images by leveraging contextual information about the tweet content, enterprise details, and posting time. We then explore methods to enhance the engagement of text-to-image models, making initial strides in this d irection. These include conditioning image generation on improved prompts, super vised fine-tuning of stable diffusion on high-performing images, and reinforceme nt learning to align stable diffusion with *EngageNet*-based reward signals, all of which lead to the generation of images with higher viewer engagement. Finall y, we propose the *Engagement Arena*, to benchmark text-to-image models based on

their ability to generate engaging images, using *EngageNet* as the evaluator, thereby encouraging the research community to measure further advances in the engagement of text-to-image modeling. These contributions provide a new pathway for advancing utility-driven image generation, with significant implications for the commercial application of image generation. We have released our code and dataset on behavior-in-the-wild.github.io/image-engagement.
****
Ming Zhong,Aston Zhang,Xuewei Wang,Rui Hou,Wenhan Xiong,Chenguang Zhu,Zhengxing Chen,Liang Tan,Chloe Bi,Mike Lewis,Sravya Popuri,Sharan Narang,Melanie Kambadur, Dhruv Mahajan,Sergey Edunov,Jiawei Han,Laurens van der Maaten
Law of the Weakest Link: Cross Capabilities of Large Language Models
The development and evaluation of Large Language Models (LLMs) have largely focused on individual capabilities. However, this overlooks the intersection of multiple abilities across different types of expertise that are often required for real-world tasks, which we term **cross capabilities**. To systematically explore this concept, we first define seven core individual capabilities and then pair them to form seven common cross capabilities, each supported by a manually constructed taxonomy. Building on these definitions, we introduce *CrossEval*, a benchmark comprising 1,400 human-annotated prompts, with 100 prompts for each individual and cross capability. To ensure reliable evaluation, we involve expert annotators to assess 4,200 model responses, gathering 8,400 human ratings with detailed explanations to serve as reference examples. Our findings reveal that current LLMs consistently exhibit the ``Law of the Weakest Link,'' where cross-capability performance is significantly constrained by the weakest component. Across 58 cross-capability scores from 17 models, 38 scores are lower than all individual capabilities, while 20 fall between strong and weak, but closer to the weaker ability. These results highlight LLMs' underperformance in cross-capability tasks, emphasizing the need to identify and improve their weakest capabilities as a key research priority. The code, benchmarks, and evaluations are available on our [project website](https://www.llm-cross-capabilities.org).
****
Tyna Eloundou,Alex Beutel,David G. Robinson,Keren Gu,Anna-Luisa Brakman,Pamela Mishkin,Meghan Shah,Johannes Heidecke,Lilian Weng,Adam Tauman Kalai
First-Person Fairness in Chatbots
Evaluating chatbot fairness is crucial given their rapid proliferation, yet typical chatbot tasks (e.g., resume writing, entertainment) diverge from the institutional decision-making tasks (e.g., resume screening) which have traditionally been central to discussion of algorithmic fairness. The open-ended nature and diverse use-cases of chatbots necessitate novel methods for bias assessment. This paper addresses these challenges by introducing a scalable counterfactual approach to evaluate "first-person fairness," meaning fairness toward chatbot users based on demographic characteristics. Our method employs a Language Model as a Research Assistant (LMRA) to yield quantitative measures of harmful stereotypes and qualitative analyses of demographic differences in chatbot responses. We apply this approach to assess biases in six of our language models across millions of interactions, covering sixty-six tasks in nine domains and spanning two genders and four races. Independent human annotations corroborate the LMRA-generated bias evaluations. This study represents the first large-scale fairness evaluation based on real-world chat data. We highlight that post-training reinforcement learning techniques significantly mitigate these biases. This evaluation provides a practical methodology for ongoing bias monitoring and mitigation.
****
Jie Chen
Graph Neural Preconditioners for Iterative Solutions of Sparse Linear Systems
Preconditioning is at the heart of iterative solutions of large, sparse linear systems of equations in scientific disciplines. Several algebraic approaches, which access no information beyond the matrix itself, are widely studied and used, but ill-conditioned matrices remain very challenging. We take a machine learning approach and propose using graph neural networks as a general-purpose precondit

ioner. They show attractive performance for many problems and can be used when t
he mainstream preconditioners perform poorly. Empirical evaluation on over 800 m
atrices suggests that the construction time of these graph neural preconditioner
s (GNPs) is more predictable and can be much shorter than that of other widely u
sed ones, such as ILU and AMG, while the execution time is faster than using a K
rylov method as the preconditioner, such as in inner-outer GMRES. GNPs have a st
rong potential for solving large-scale, challenging algebraic problems arising f
rom not only partial differential equations, but also economics, statistics, gra
ph, and optimization, to name a few.
****

Shicheng Liu,Minghui Zhu
UTILITY: Utilizing Explainable Reinforcement Learning to Improve Reinforcement L
earning
Reinforcement learning (RL) faces two challenges: (1) The RL agent lacks explain
ability. (2) The trained RL agent is, in many cases, non-optimal and even far fr
om optimal. To address the first challenge, explainable reinforcement learning (
XRL) is proposed to explain the decision-making of the RL agent. In this paper,
we demonstrate that XRL can also be used to address the second challenge, i.e.,
improve RL performance. Our method has two parts. The first part provides a two-
level explanation for why the RL agent is not optimal by identifying the mistake
s made by the RL agent. Since this explanation includes the mistakes of the RL a
gent, it has the potential to help correct the mistakes and thus improve RL perf
ormance. The second part formulates a constrained bi-level optimization problem
to learn how to best utilize the two-level explanation to improve RL performance
. In specific, the upper level learns how to use the high-level explanation to s
hape the reward so that the corresponding policy can maximize the cumulative gro
und truth reward, and the lower level learns the corresponding policy by solving
 a constrained RL problem formulated using the low-level explanation. We propose
 a novel algorithm to solve this constrained bi-level optimization problem, and
theoretically guarantee that the algorithm attains global optimality. We use MuJ
oCo experiments to show that our method outperforms state-of-the-art baselines.
****

Haofu Qian,Chenjia Bai,Jiatao Zhang,Fei Wu,Wei Song,Xuelong Li
Discriminator-Guided Embodied Planning for LLM Agent
Large Language Models (LLMs) have showcased remarkable reasoning capabilities in
 various domains, yet face challenges in complex embodied tasks due to the need
for a coherent long-term policy and context-sensitive environmental understandin
g. Previous work performed LLM refinement relying on outcome-supervised feedback
, which can be costly and ineffective. In this work, we introduce a novel framew
ork, Discriminator-Guided Action Optimization (DGAP), for facilitating the optim
ization of LLM action plans via step-wise signals. Specifically, we employ a lim
ited set of demonstrations to enable the discriminator to learn a score function
, which assesses the alignment between LLM-generated actions and the underlying
optimal ones at every step. Based on the discriminator, LLMs are prompted to gen
erate actions that maximize the score, utilizing historical action-score pair tr
ajectories as guidance. Under mild conditions, DGAP resembles critic-regularized
 optimization and has been demonstrated to achieve a stronger policy than the LL
M planner. In experiments across different LLMs (GPT-4, Llama3-70B) in ScienceWo
rld and VirtualHome, our method achieves superior performance and better efficie
ncy than previous methods.
****

ChuXin Wang,Wenfei Yang,Xiang Liu,Tianzhu Zhang
State Space Model Meets Transformer: A New Paradigm for 3D Object Detection
DETR-based methods, which use multi-layer transformer decoders to refine object
queries iteratively, have shown promising performance in 3D indoor object detect
ion. However, the scene point features in the transformer decoder remain fixed,
leading to minimal contributions from later decoder layers, thereby limiting per
formance improvement. Recently, State Space Models (SSM) have shown efficient co
ntext modeling ability with linear complexity through iterative interactions bet
ween system states and inputs. Inspired by SSMs, we propose a new 3D object DEte

ction paradigm with an interactive STate space model (DEST). In the interactive SSM, we design a novel state-dependent SSM parameterization method that enables system states to effectively serve as queries in 3D indoor detection tasks. In addition, we introduce four key designs tailored to the characteristics of point cloud and SSM: The serialization and bidirectional scanning strategies enable bidirectional feature interaction among scene points within the SSM. The inter-state attention mechanism models the relationships between state points, while the gated feed-forward network enhances inter-channel correlations. To the best of our knowledge, this is the first method to model queries as system states and scene points as system inputs, which can simultaneously update scene point features and query features with linear complexity. Extensive experiments on two challenging datasets demonstrate the effectiveness of our DEST-based method. Our method improves the GroupFree baseline in terms of AP_{50} on ScanNet V2 (+5.3) and SUN RGB-D (+3.2) datasets. Based on the VDETR baseline, Our method sets a new state-of-the-art on the ScanNetV2 and SUN RGB-D datasets.
****

Yan Xie,Minrui Lv,Chen-Song Zhang
MGCFNN: A Neural MultiGrid Solver with Novel Fourier Neural Network for High Wave Number Helmholtz Equations
Solving high wavenumber Helmholtz equations is notoriously challenging. Traditional solvers have yet to yield satisfactory results, and most neural network methods struggle to accurately solve cases with extremely high wavenumbers within heterogeneous media. This paper presents an advanced multigrid-hierarchical AI solver, tailored specifically for high wavenumber Helmholtz equations. We adapt the MGCNN architecture to align with the problem setting and incorporate a novel Fourier neural network (FNN) to match the characteristics of Helmholtz equations. FNN, mathematically akin to the convolutional neural network (CNN), enables faster propagation of source influence during the solve phase, making it particularly suitable for handling large size, high wavenumber problems. We conduct supervised learning tests against numerous neural operator learning methods to demonstrate the superior learning capabilities of our solvers. Additionally, we perform scalability tests using an unsupervised strategy to highlight our solvers' significant speedup over the most recent specialized AI solver and AI-enhanced traditional solver for high wavenumber Helmholtz equations. We also carry out an ablation study to underscore the effectiveness of the multigrid hierarchy and the benefits of introducing FNN. Notably, our solvers exhibit optimal convergence of $\mathcal{O}(k)$ up to $k \approx 2000$.
****

Edward S. Hu,Kwangjun Ahn,Qinghua Liu,Haoran Xu,Manan Tomar,Ada Langford,Dinesh Jayaraman,Alex Lamb,John Langford
The Belief State Transformer
We introduce the "Belief State Transformer", a next-token predictor that takes both a prefix and suffix as inputs, with a novel objective of predicting both the next token for the prefix and the previous token for the suffix. The Belief State Transformer effectively learns to solve challenging problems that conventional forward-only transformers struggle with, in a domain-independent fashion. Key to this success is learning a compact belief state that captures all relevant information necessary for accurate predictions.
Empirical ablations show that each component of the model is essential in difficult scenarios where standard Transformers fall short.
For the task of story writing with known prefixes and suffixes, our approach outperforms the Fill-in-the-Middle method for reaching known goals and demonstrates improved performance even when the goals are unknown. Altogether, the Belief State Transformer enables more efficient goal-conditioned decoding, better test-time inference, and high-quality text representations on small scale problems. Website: https://sites.google.com/view/belief-state-transformer/
****

Jamie Hayes,Ilia Shumailov,William P. Porter,Aneesh Pappu
Measuring memorization in RLHF for code completion
Reinforcement learning with human feedback (RLHF) has become the dominant method

to align large models to user preferences.
Unlike fine-tuning, for which there are many studies regarding training data memorization, it is not clear how memorization is affected by or introduced in the RLHF alignment process.
Understanding this relationship is important as real user data may be collected and used to align large models; if user data is memorized during RLHF and later regurgitated, this could raise privacy concerns. In addition to RLHF, other methods such as Direct Preference Optimization (DPO) and ΨPO have gained popularity for learning directly from human preferences, removing the need for optimizing intermediary reward models with reinforcement learning.
In this work, we analyze how training data memorization can surface and propagate through each phase of RLHF and direct preference learning.
We focus our study on code completion models, as code completion is one of the most popular use cases for large language models. We find that RLHF significantly decreases the chance that data used for reward modeling and reinforcement learning is memorized in comparison to directly fine-tuning on this data, but that examples already memorized during the fine-tuning stage of RLHF, will, in the majority of cases, remain memorized after RLHF. In contrast, we find that aligning by learning directly from human preference data via a special case of ΨPO, Identity Preference Optimization (IPO), increases the likelihood that training data is regurgitated compared to RLHF. Our work suggests that RLHF, as opposed to direct preference learning, is a safer way to mitigate the risk of regurgitating sensitive preference data when aligning large language models. We find our conclusions are robust across multiple code completion datasets, tasks, and model scales.
****
S Sakshi,Utkarsh Tyagi,Sonal Kumar,Ashish Seth,Ramaneswaran Selvakumar,Oriol Nieto,Ramani Duraiswami,Sreyan Ghosh,Dinesh Manocha
MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark
The ability to comprehend audio—which includes speech, non-speech sounds, and music—is crucial for AI agents to interact effectively with the world. We present MMAU, a novel benchmark designed to evaluate multimodal audio understanding models on tasks requiring expert-level knowledge and complex reasoning. MMAU comprises 10k carefully curated audio clips paired with human-annotated natural language questions and answers spanning speech, environmental sounds, and music. It includes information extraction and reasoning questions, requiring models to demonstrate 27 distinct skills across unique and challenging tasks. Unlike existing benchmarks, MMAU emphasizes advanced perception and reasoning with domain-specific knowledge, challenging models to tackle tasks akin to those faced by experts. We assess 18 open-source and proprietary (Large) Audio-Language Models, demonstrating the significant challenges
posed by MMAU. Notably, even the most advanced Gemini 2.0 Flash achieves only 59.93% accuracy, and the state-of-the-art open-source Qwen2-Audio achieves only 52.50%, highlighting considerable room for improvement. We believe MMAU will drive the audio and multimodal research community to develop more advanced audio understanding models capable of solving complex audio tasks.
****
Elies Gil-Fuster,Casper Gyurik,Adrian Perez-Salinas,Vedran Dunjko
On the Relation between Trainability and Dequantization of Variational Quantum Learning Models
Quantum machine learning (QML) explores the potential advantages of quantum computers for machine learning tasks, with variational QML among the main current approaches.
While quantum computers promise to solve problems that are classically intractable, it has been recently shown that a particular quantum algorithm which outperforms all pre-existing classical algorithms can be matched by a newly developed classical approach (often inspired by the quantum algorithm).
We say such algorithms have been dequantized.
For QML models to be effective, they must be trainable and non-dequantizable.
The relationship between these properties is still not fully understood and rece

nt works raised into question to what extent we could ever have QML models which are both trainable and non-dequantizable.
This challenges the potential of QML altogether.
In this work we answer open questions regarding when trainability and non-dequantization are compatible.
We first formalize the key concepts and put them in the context of prior research.
We introduce the role of "variationalness" of QML models using well-known quantum circuit architectures as leading examples.
Our results provide recipes for variational QML models that are trainable and non-dequantizable.
By ensuring that variational QML models are both trainable and non-dequantizable, we pave the way toward practical relevance.
****

Fu Luo,Xi Lin,Yaoxin Wu,Zhenkun Wang,Tong Xialiang,Mingxuan Yuan,Qingfu Zhang
Boosting Neural Combinatorial Optimization for Large-Scale Vehicle Routing Problems
Neural Combinatorial Optimization (NCO) methods have exhibited promising performance in solving Vehicle Routing Problems (VRPs). However, most NCO methods rely on the conventional self-attention mechanism that induces excessive computational complexity, thereby struggling to contend with large-scale VRPs and hindering their practical applicability. In this paper, we propose a lightweight cross-attention mechanism with linear complexity, by which a Transformer network is developed to learn efficient and favorable solutions for large-scale VRPs. We also propose a Self-Improved Training (SIT) algorithm that enables direct model training on large-scale VRP instances, bypassing extensive computational overhead for attaining labels. By iterating solution reconstruction, the Transformer network itself can generate improved partial solutions as pseudo-labels to guide the model training. Experimental results on the Travelling Salesman Problem (TSP) and the Capacitated Vehicle Routing Problem (CVRP) with up to 100K nodes indicate that our method consistently achieves superior performance for synthetic and real-world benchmarks, significantly boosting the scalability of NCO methods.
****

Yaniv Oren,Viliam Vadocz,Matthijs T. J. Spaan,Wendelin Boehmer
Epistemic Monte Carlo Tree Search
The AlphaZero/MuZero (A/MZ) family of algorithms has achieved remarkable success across various challenging domains by integrating Monte Carlo Tree Search (MCTS) with learned models. Learned models introduce epistemic uncertainty, which is caused by learning from limited data and is useful for exploration in sparse reward environments. MCTS does not account for the propagation of this uncertainty however. To address this, we introduce Epistemic MCTS (EMCTS): a theoretically motivated approach to account for the epistemic uncertainty in search and harness the search for deep exploration. In the challenging sparse-reward task of writing code in the Assembly language SUBLEQ, AZ paired with our method achieves significantly higher sample efficiency over baseline AZ. Search with EMCTS solves variations of the commonly used hard-exploration benchmark Deep Sea - which baseline A/MZ are practically unable to solve - much faster than an otherwise equivalent method that does not use search for uncertainty estimation, demonstrating significant benefits from search for epistemic uncertainty estimation.
****

Huaijin Wu,Lianqiang Li,Hantao Huang,Tu Yi,Jihang Zhang,Minghui Yu,Junchi Yan
HShare: Fast LLM Decoding by Hierarchical Key-Value Sharing
The frequent retrieval of Key-Value (KV) cache data has emerged as a significant factor contributing to the inefficiency of the inference process in large language models. Previous research has demonstrated that a small subset of critical KV cache tokens largely influences attention outcomes, leading to methods that either employ fixed sparsity patterns or dynamically select critical tokens based on the query. While dynamic sparse patterns have proven to be more effective, they introduce significant computational overhead, as critical tokens must be reselected for each self-attention computation. In this paper, we reveal substantial

similarities in KV cache token criticality across neighboring queries, layers, and heads. Motivated by this insight, we propose HShare, a hierarchical KV sharing framework. HShare facilitates the sharing of critical KV cache token indices across layers, heads, and queries, which significantly reduces the computational overhead associated with query-aware dynamic token sparsity. In addition, we introduce a greedy algorithm that dynamically determines the optimal layer-level and head-level sharing configuration for the decoding phase. We evaluate the effectiveness and efficiency of HShare across various tasks using three models: LLaMA2-7b, LLaMA3-70b, and Mistral-7b. Experimental results demonstrate that HShare achieves competitive accuracy with different sharing ratios, while delivering up to an $8.6\times$ speedup in self-attention operations and a $2.7\times$ improvement in end-to-end throughput compared with FlashAttention2 and GPT-fast respectively. The source code is publicly available at ~\url{https://github.com/wuhuaijin/HShare}.
****
Weihuang Wen,Tianshu Yu
HyperPLR: Hypergraph Generation through Projection, Learning, and Reconstruction
Hypergraphs are essential in modeling higher-order complex networks, excelling in representing group interactions within real-world contexts. This is particularly evident in collaboration networks, where they facilitate the capture of group wise polyadic patterns, extending beyond traditional pairwise dyadic interactions. The use of hypergraph generators, or generative models, is a crucial method for promoting and validating our understanding of these structures. If such generators accurately replicate observed hypergraph patterns, it reinforces the validity of our interpretations. In this context, we introduce a novel hypergraph generative paradigm, **HyperPLR**, encompassing three phases: Projection, Learning, and Reconstruction. Initially, the hypergraph is projected onto a weighted graph. Subsequently, the model learns this graph's structure within a latent space, while simultaneously computing a distribution between the hyperedge and the projected graph. Finally, leveraging the learned model and distribution, HyperPLR generates new weighted graphs and samples cliques from them. These cliques are then used to reconstruct new hypergraphs by solving a specific clique cover problem.
We have evaluated HyperPLR on existing real-world hypergraph datasets, which consistently demonstrate superior performance and validate the effectiveness of our approach.
****
Richard Bergna,Sergio Calvo Ordoñez,Felix Opolka,Pietro Lio,José Miguel Hernández-Lobato
Uncertainty Modeling in Graph Neural Networks via Stochastic Differential Equations
We propose a novel Stochastic Differential Equation (SDE) framework to address the problem of learning uncertainty-aware representations for graph-structured data. While Graph Neural Ordinary Differential Equations (GNODEs) have shown promise in learning node representations, they lack the ability to quantify uncertainty. To address this, we introduce Latent Graph Neural Stochastic Differential Equations (LGNSDE), which enhance GNODE by embedding randomness through a Bayesian prior-posterior mechanism for epistemic uncertainty and Brownian motion for aleatoric uncertainty. By leveraging the existence and uniqueness of solutions to graph-based SDEs, we prove that the variance of the latent space bounds the variance of model outputs, thereby providing theoretically sensible guarantees for the uncertainty estimates. Furthermore, we show mathematically that LGNSDEs are robust to small perturbations in the input, maintaining stability over time. Empirical results across several benchmarks demonstrate that our framework is competitive in out-of-distribution detection, robustness to noise perturbations, and active learning, underscoring the ability of LGNSDEs to quantify uncertainty reliably.
****
Yuxuan Zhou,Xien Liu,Chen Ning,Xiao Zhang,Ji Wu
Reliable and Diverse Evaluation of LLM Medical Knowledge Mastery

Mastering medical knowledge is crucial for medical-specific LLMs. However, despi
te the existence of medical benchmarks like MedQA, a unified framework that full
y leverages existing knowledge bases to evaluate LLMs' mastery of medical knowle
dge is still lacking. We propose PretexEval, a novel framework that dynamically
generates reliable and diverse test samples to evaluate LLMs for any given medic
al knowledge base. We notice that test samples produced directly from knowledge
bases by templates or LLMs may introduce factual errors and also lack diversity.
To address these issues, our framework employs predicate equivalence transforma
tions to produce a series of variants for any given medical knowledge point. Fin
ally, these produced predicate variants are converted into textual language, res
ulting in a series of reliable and diverse test samples. Here, we use our propos
ed framework to systematically investigate the mastery of medical factual knowle
dge of 12 well-known LLMs, based on two knowledge bases that are crucial for cli
nical diagnosis and treatment. The evaluation results illustrate that current LL
Ms still exhibit significant deficiencies in fully mastering medical knowledge,
despite achieving considerable success on some famous public benchmarks. These n
ew findings provide valuable insights for developing medical-specific LLMs, high
lighting that current LLMs urgently need to strengthen their comprehensive and i
n-depth mastery of medical knowledge before being applied to real-world medical
scenarios.
****
Zhengfeng Lai,Vasileios Saveris,Chen Chen,Hong-You Chen,Haotian Zhang,Bowen Zhan
g,Wenze Hu,Juan Lao Tebar,Zhe Gan,Peter Grasch,Meng Cao,Yinfei Yang
Revisit Large-Scale Image-Caption Data in Pre-training Multimodal Foundation Mod
els
Recent advancements in multimodal models highlight the value of rewritten captio
ns for improving performance, yet key challenges remain. For example, while synt
hetic captions often provide superior quality and image-text alignment, it is no
t clear whether they can fully replace AltTexts: the role of synthetic captions
and their interaction with original web-crawled AltTexts in pre-training is stil
l not well understood. Moreover, different multimodal foundation models may have
unique preferences for specific caption formats, but efforts to identify the op
timal captions for each model remain limited. In this work, we propose a novel,
controllable, and scalable captioning pipeline designed to generate diverse capt
ion formats tailored to various multimodal models. By examining short synthetic
captions (SSC) and descriptive synthetic captions (DSC) as case studies, we syst
ematically explore their effects and interactions with AltTexts across models su
ch as CLIP, multimodal LLMs, and diffusion models. Our findings reveal that a hy
brid approach that keeps both synthetic captions and AltTexts can outperform the
use of synthetic captions alone, improving both alignment and performance, with
each model demonstrating preferences for particular caption formats. This compr
ehensive analysis provides valuable insights into optimizing captioning strategi
es, thereby advancing the pre-training of multimodal foundation models.
****
Xinghao Chen,Siwei Li,Yijing Yang,Yunhe Wang
DECO: Unleashing the Potential of ConvNets for Query-based Detection and Segment
ation
Transformer and its variants have shown great potential for various vision tasks
in recent years, including image classification, object detection and segmentat
ion.
Meanwhile, recent studies also reveal that with proper architecture design, conv
olutional networks (ConvNets) also achieve competitive performance with transfor
mers.
However, no prior methods have explored to utilize pure convolution to build a T
ransformer-style Decoder module, which is essential for Encoder-Decoder architec
ture like Detection Transformer (DETR).
To this end, in this paper we explore whether we could build query-based detect
ion and segmentation framework with ConvNets instead of sophisticated transforme
r architecture.
We propose a novel mechanism dubbed InterConv to perform interaction between obj

ect queries and image features via convolutional layers.
Equipped with the proposed InterConv, we build Detection ConvNet (DECO), which is composed of a backbone and convolutional encoder-decoder architecture. We compare the proposed DECO against prior detectors on the challenging COCO benchmark. Despite its simplicity, our DECO achieves competitive performance in terms of detection accuracy and running speed. Specifically,
with the ResNet-18 and ResNet-50 backbone, our DECO achieves 40.5% and 47.8% AP with 66 and 34 FPS, respectively. The proposed method is also evaluated on the segment anything task, demonstrating similar performance and higher efficiency.
We hope the proposed method brings another perspective for designing architectures for vision tasks.
Codes are available at \url{https://github.com/xinghaochen/DECO} and \url{https://github.com/mindspore-lab/models/tree/master/research/huawei-noah/DECO}.
****
Tomas Geffner,Kieran Didi,Zuobai Zhang,Danny Reidenbach,Zhonglin Cao,Jason Yim,Mario Geiger,Christian Dallago,Emine Kucukbenli,Arash Vahdat,Karsten Kreis
Proteina: Scaling Flow-based Protein Structure Generative Models
Recently, diffusion- and flow-based generative models of protein structures have emerged as a powerful tool for de novo protein design. Here, we develop *Proteina*, a new large-scale flow-based protein backbone generator that utilizes hierarchical fold class labels for conditioning and relies on a tailored scalable transformer architecture with up to $5\times$ as many parameters as previous models. To meaningfully quantify performance, we introduce a new set of metrics that directly measure the distributional similarity of generated proteins with reference sets, complementing existing metrics. We further explore scaling training data to millions of synthetic protein structures and explore improved training and sampling recipes adapted to protein backbone generation. This includes fine-tuning strategies like LoRA for protein backbones, new guidance methods like classifier-free guidance and autoguidance for protein backbones, and new adjusted training objectives. Proteina achieves state-of-the-art performance on de novo protein backbone design and produces diverse and designable proteins at unprecedented length, up to 800 residues. The hierarchical conditioning offers novel control, enabling high-level secondary-structure guidance as well as low-level fold-specific generation.
****
Jannis Chemseddine,Christian Wald,Richard Duong,Gabriele Steidl
Neural Sampling from Boltzmann Densities: Fisher-Rao Curves in the Wasserstein Geometry
We deal with the task of sampling from an unnormalized Boltzmann density ρ_D
by learning a Boltzmann curve given by energies f_t starting in a simple density ρ_Z.
First, we examine conditions under which Fisher-Rao flows are absolutely continuous in the Wasserstein geometry.
Second, we address specific interpolations f_t and the learning of the related density/velocity pairs (ρ_t,v_t).
It was numerically observed that the linear interpolation,
which requires only a parametrization of the velocity field v_t,
suffers from a "teleportation-of-mass" issue.
Using tools from the Wasserstein geometry,
we give an analytical example,
where we can precisely measure the explosion of the velocity field.
Inspired by Máté and Fleuret, who
parametrize both f_t and v_t, we propose an
interpolation which parametrizes only f_t and fixes an appropriate v_t.
This corresponds to
the Wasserstein gradient flow of the Kullback-Leibler divergence related to Langevin dynamics.
We demonstrate by numerical examples that our model provides a well-behaved flow

field which successfully solves the above sampling task.
****
Jacob Beck,Shikha Surana,Manus McAuliffe,Oliver Bent,Thomas D Barrett,Juan Jose Garau-Luis,Paul Duckworth
Metalic: Meta-Learning In-Context with Protein Language Models
Predicting the biophysical and functional properties of proteins is essential for in silico protein design. Machine learning has emerged as a promising technique for such prediction tasks. However, the relative scarcity of in vitro annotations means that these models often have little, or no, specific data on the desired fitness prediction task. As a result of limited data, protein language models (PLMs) are typically trained on general protein sequence modeling tasks, and then fine-tuned, or applied zero-shot, to protein fitness prediction. When no task data is available, the models make strong assumptions about the correlation between the protein sequence likelihood and fitness scores. In contrast, we propose meta-learning over a distribution of standard fitness prediction tasks, and demonstrate positive transfer to unseen fitness prediction tasks. Our method, called Metalic (Meta-Learning In-Context), uses in-context learning and fine-tuning, when data is available, to adapt to new tasks. Crucially, fine-tuning enables considerable generalization, even though it is not accounted for during meta-training. Our fine-tuned models achieve strong results with 18 times fewer parameters than state-of-the-art models. Moreover, our method sets a new state-of-the-art in low-data settings on ProteinGym, an established fitness-prediction benchmark. Due to data scarcity, we believe meta-learning will play a pivotal role in advancing protein engineering.
****
Shauli Ravfogel,Anej Svete,Vésteinn Snæbjarnarson,Ryan Cotterell
Gumbel Counterfactual Generation From Language Models
Understanding and manipulating the causal generation mechanisms in language models is essential for controlling their behavior. Previous work has primarily relied on techniques such as representation surgery---e.g., model ablations or manipulation of linear subspaces tied to specific concepts---to intervene on these models. To understand the impact of interventions precisely, it is useful to examine counterfactuals---e.g., how a given sentence would have appeared had it been generated by the model following a specific intervention. We highlight that counterfactual reasoning is conceptually distinct from interventions, as articulated in Pearl's causal hierarchy. Based on this observation, we propose a framework for generating true string counterfactuals by reformulating language models as a structural equation model using the Gumbel-max trick, which we called Gumbel counterfactual generation.
This reformulation allows us to model the joint distribution over original strings and their counterfactuals resulting from the same instantiation of the sampling noise. We develop an algorithm based on hindsight Gumbel sampling that allows us to infer the latent noise variables and generate counterfactuals of observed strings. Our experiments demonstrate that the approach produces meaningful counterfactuals while at the same time showing that commonly used intervention techniques have considerable undesired side effects.
****
Taesung Kwon,Jong Chul Ye
Solving Video Inverse Problems Using Image Diffusion Models
Recently, diffusion model-based inverse problem solvers (DIS) have emerged as state-of-the-art approaches for addressing inverse problems, including image super-resolution, deblurring, inpainting, etc.
However, their application to video inverse problems arising from spatio-temporal degradation remains largely unexplored due to the challenges in training video diffusion models.
To address this issue, here we introduce an innovative video inverse solver that leverages only image diffusion models.
Specifically, by
drawing inspiration from the success of the recent decomposed diffusion sampler (DDS),

our method treats the time dimension of a video as the batch dimension of image diffusion models and solves spatio-temporal optimization problems within denoised spatio-temporal batches derived from each image diffusion model.
Moreover, we introduce a batch-consistent diffusion sampling strategy that encourages consistency across batches by synchronizing the stochastic noise components in image diffusion models.
Our approach synergistically combines batch-consistent sampling with simultaneous optimization of denoised spatio-temporal batches at each reverse diffusion step, resulting in a novel and efficient diffusion sampling strategy for video inverse problems.
Experimental results demonstrate that our method effectively addresses various spatio-temporal degradations in video inverse problems, achieving state-of-the-art reconstructions.
Project page: https://svi-diffusion.github.io/
****
Mingkun Zhang,Keping Bi,Wei Chen,Jiafeng Guo,Xueqi Cheng
CLIPure: Purification in Latent Space via CLIP for Adversarially Robust Zero-Shot Classification
In this paper, we aim to build an adversarially robust zero-shot image classifier that can accurately and efficiently classify unseen examples while defending against unforeseen adversarial attacks, addressing critical challenges in real-world safety-sensitive scenarios. To achieve this, we focus on two key challenges: zero-shot classification and defense against unforeseen attacks. We ground our work on CLIP, a vision-language pre-trained model to perform zero-shot classification.
To defend against unforeseen attacks, we adopt a purification approach, as it is independent of specific attack types.
We then define a purification risk as the KL divergence between the joint distributions of the purification and attack process.
The derived lower bound of purification risk inspires us to explore purification in CLIP's multi-modal latent space.
We propose a CLIP-based purification method called CLIPure, which has two variants: _CLIPure-Diff_, which models image likelihood with a generative process of its latent vector, and _CLIPure-Cos_, which models the likelihood based on the similarity between embeddings of the image and a blank template. As far as we know, CLIPure is the first purification method in latent space and _CLIPure-Cos_ is the first purification method not relying on generative models, substantially improving defense efficiency. Extensive experimental results show that the robustness achieved by CLIPure is within a small gap of clean accuracy, outperforming SOTA robustness by a large margin, e.g., from 71.7\% to **91.1\%** on CIFAR10, from 59.6\% to **72.6\%** on ImageNet, and **108\%** relative improvements of average robustness on the 13 datasets over previous SOTA, with only 14\% extra inference cost and no additional training.
****
Daeho Um,Yoonji Lee,Jiwoong Park,Seulki Park,Yuneil Yeo,Seong Jin Ahn
Relation-Aware Diffusion for Heterogeneous Graphs with Partially Observed Features
Diffusion-based imputation methods, which impute missing features through the iterative propagation of observed features, have shown impressive performance in homogeneous graphs. However, these methods are not directly applicable to heterogeneous graphs, which have multiple types of nodes and edges, due to two key issues: (1) the presence of nodes with undefined features hinders diffusion-based imputation; (2) treating various edge types equally during diffusion does not fully utilize information contained in heterogeneous graphs. To address these challenges, this paper presents a novel imputation scheme that enables diffusion-based imputation in heterogeneous graphs. Our key idea involves (1) assigning a {\it virtual feature} to an undefined node feature and (2) determining the importance of each edge type during diffusion according to a new criterion. Through experiments, we demonstrate that our virtual feature scheme effectively serves as a bridge between existing diffusion-based methods and heterogeneous graphs, maintain

ing the advantages of these methods. Furthermore, we confirm that adjusting the importance of each edge type leads to significant performance gains on heterogeneous graphs. Extensive experimental results demonstrate the superiority of our scheme in both semi-supervised node classification and link prediction tasks on heterogeneous graphs with missing rates ranging from low to exceedingly high. The source code is available at https://github.com/daehoum1/hetgfd.
****

Tianren Ma,Lingxi Xie,Yunjie Tian,Boyu Yang,Qixiang Ye
ClawMachine: Learning to Fetch Visual Tokens for Referential Comprehension
Aligning vision and language concepts at a finer level remains an essential topic of multimodal large language models (MLLMs), particularly for tasks such as referring and grounding. Existing methods, such as *proxy encoding* and *geometry encoding* genres, incorporate additional syntax to encode spatial information, imposing extra burdens when communicating between language with vision modules. In this study, we propose ClawMachine, offering a new methodology that explicitly notates each entity using **token collectives**—groups of visual tokens that collaboratively represent higher-level semantics. A hybrid perception mechanism is also explored to perceive and understand scenes from both discrete and continuous spaces. Our method unifies the prompt and answer of visual referential tasks without using additional syntax. By leveraging a joint vision-language vocabulary, ClawMachine integrates referring and grounding in an auto-regressive manner, demonstrating great potential with scaled up pre-training data. Experiments show that ClawMachine achieves superior performance on scene-level and referential understanding tasks with higher efficiency. It also exhibits the potential to integrate multi-source information for complex visual reasoning, which is beyond the capability of many MLLMs. Our code is available at https://github.com/martian422/ClawMachine.
****

Ruiqi Ni,zherong pan,Ahmed H Qureshi
Physics-informed Temporal Difference Metric Learning for Robot Motion Planning
The motion planning problem involves finding a collision-free path from a robot's starting to its target configuration. Recently, self-supervised learning methods have emerged to tackle motion planning problems without requiring expensive expert demonstrations. They solve the Eikonal equation for training neural networks and lead to efficient solutions. However, these methods struggle in complex environments because they fail to maintain key properties of the Eikonal equation, such as optimal value functions and geodesic distances. To overcome these limitations, we propose a novel self-supervised temporal difference metric learning approach that solves the Eikonal equation more accurately and enhances performance in solving complex and unseen planning tasks. Our method enforces Bellman's principle of optimality over finite regions, using temporal difference learning to avoid spurious local minima while incorporating metric learning to preserve the Eikonal equation's essential geodesic properties. We demonstrate that our approach significantly outperforms existing self-supervised learning methods in handling complex environments and generalizing to unseen environments, with robot configurations ranging from 2 to 12 degrees of freedom (DOF).
****

Xiaorui Peng,Yuheng Jia,Fuchao Yang,Ran Wang,Min-Ling Zhang
Noise Separation guided Candidate Label Reconstruction for Noisy Partial Label Learning
Partial label learning is a weakly supervised learning problem in which an instance is annotated with a set of candidate labels, among which only one is the correct label. However, in practice the correct label is not always in the candidate label set, leading to the noisy partial label learning (NPLL) problem. In this paper, we theoretically prove that the generalization error of the classifier constructed under NPLL paradigm is bounded by the noise rate and the average length of the candidate label set. Motivated by the theoretical guide, we propose a novel NPLL framework that can separate the noisy samples from the normal samples to reduce the noise rate and reconstruct the shorter candidate label sets for both of them. Extensive experiments on multiple benchmark datasets confirm the ef

ficacy of the proposed method in addressing NPLL. For example, on CIFAR100 datas et with severe noise, our method improves the classification accuracy of the sta te-of-the-art one by 11.57%. The code is available at: https://github.com/pruiru i/PLRC.
****
André Hottung,Mridul Mahajan,Kevin Tierney
PolyNet: Learning Diverse Solution Strategies for Neural Combinatorial Optimizat ion
Reinforcement learning-based methods for constructing solutions to combinatorial optimization problems are rapidly approaching the performance of human-designed algorithms. To further narrow the gap, learning-based approaches must efficient ly explore the solution space during the search process. Recent approaches artif icially increase exploration by enforcing diverse solution generation through ha ndcrafted rules, however, these rules can impair solution quality and are diffic ult to design for more complex problems. In this paper, we introduce PolyNet, a n approach for improving exploration of the solution space by learning complemen tary solution strategies. In contrast to other works, PolyNet uses only a single -decoder and a training schema that does not enforce diverse solution generation through handcrafted rules. We evaluate PolyNet on four combinatorial optimizati on problems and observe that the implicit diversity mechanism allows PolyNet to find better solutions than approaches that explicitly enforce diverse solution g eneration.
****
Xinmeng Xu,Yiqun Zhang,Jizhen Li,Yuhong Yang,Yong Luo,Weiping Tu
FIRING-Net: A filtered feature recycling network for speech enhancement
Current deep neural networks for speech enhancement (SE) aim to minimize the dis tance between the output signal and the clean target by filtering out noise feat ures from input features. However, when noise and speech components are highly s imilar, SE models struggle to learn effective discrimination patterns. To addres s this challenge, we propose a Filter-Recycle-Interguide framework termed Filter -Recycle-INterGuide NETwork (FIRING-Net) for SE, which filters the input feature s to extract target features and recycles the filtered-out features as non-targe t features. These two feature sets then guide each other to refine the features, leading to the aggregation of speech information within the target features and noise information within the non-target features. The proposed FIRING-Net mainl y consists of a Local Module (LM) and a Global Module (GM). The LM uses outputs of the speech extraction network as target features and the residual between inp ut and output as non-target features. The GM leverages the energy distribution o f self-attention map to extract target and non-target features guided by highest and lowest energy regions. Both LM and GM include interaction modules to levera ge the two feature sets in an inter-guided manner for collecting speech from non -target features and filtering out noise from target features. Experiments confi rm the effectiveness of the Filter-Recycle-Interguide framework, with FIRING-Net achieving a strong balance between SE performance and computational efficiency, surpassing comparable models across various SNR levels and noise environments.
****
Ayush Kaushal,Tejas Vaidhya,Arnab Kumar Mondal,Tejas Pandey,Aaryan Bhagat,Irina Rish
Surprising Effectiveness of pretraining Ternary Language Model at Scale
Rapid advancements in GPU computational power has outpaced memory capacity and b andwidth growth, creating bottlenecks in Large Language Model (LLM) inference. P ost-training quantization is the leading method for addressing memory-related bo ttlenecks in LLM inference, but it suffers from significant performance degradat ion below 4-bit precision. This paper addresses these challenges by investigatin g the pretraining of low-bitwidth models specifically Ternary Language Models (T riLMs) as an alternative to traditional floating-point models (FloatLMs) and the ir post-training quantized versions (QuantLMs). We present Spectra LLM suite, th e first open suite of LLMs spanning multiple bit-widths, including FloatLMs, Qua ntLMs, and TriLMs, ranging from 99M to 3.9B parameters trained on 300B tokens. O ur comprehensive evaluation demonstrates that TriLMs offer superior scaling beha

vior in terms of model size (in bits). Surprisingly, at scales exceeding one bil
lion parameters, TriLMs consistently outperform their QuantLM and FloatLM counte
rparts for a given bit size across various benchmarks. Notably, the 3.9B paramet
er TriLM matches the performance of the FloatLM 3.9B across all benchmarks, desp
ite having fewer bits than FloatLM 830M. Overall, this research provides valuabl
e insights into the feasibility and scalability of low-bitwidth language models,
 paving the way for the development of more efficient LLMs.
****
Gustaf Ahdritz,Aravind Gollakota,Parikshit Gopalan,Charlotte Peale,Udi Wieder
Provable Uncertainty Decomposition via Higher-Order Calibration
We give a principled method for decomposing the predictive uncertainty of a mode
l into aleatoric and epistemic components with explicit semantics relating them
to the real-world data distribution. While many works in the literature have pro
posed such decompositions, they lack the type of formal guarantees we provide. O
ur method is based on the new notion of higher-order calibration, which generali
zes ordinary calibration to the setting of higher-order predictors that predict
mixtures over label distributions at every point. We show how to measure as we
ll as achieve higher-order calibration using access to k-snapshots, namely exa
mples where each point has k independent conditional labels. Under higher-orde
r calibration, the estimated aleatoric uncertainty at a point is guaranteed to m
atch the real-world aleatoric uncertainty averaged over all points where the pre
diction is made. To our knowledge, this is the first formal guarantee of this ty
pe that places no assumptions whatsoever on the real-world data distribution. Im
portantly, higher-order calibration is also applicable to existing higher-order
predictors such as Bayesian and ensemble models and provides a natural evaluatio
n metric for such models. We demonstrate through experiments that our method pro
duces meaningful uncertainty decompositions in tasks such as image classificatio
n.
****
Mayukh Deb,Mainak Deb,Apurva Ratan Murty
TopoNets: High performing vision and language models with brain-like topography
Neurons in the brain are organized such that nearby cells tend to share similar
functions. AI models lack this organization, and past efforts to introduce topog
raphy have often led to trade-offs between topography and task performance. In t
his work, we present *TopoLoss*, a new loss function that promotes spatially org
anized topographic representations in AI models without significantly sacrificin
g task performance. TopoLoss is highly adaptable and can be seamlessly integrate
d into the training of leading model architectures. We validate our method on bo
th vision (ResNet-18, ResNet-50, ViT) and language models (GPT-Neo-125M, NanoGPT
), collectively *TopoNets*. TopoNets are the highest performing supervised topog
raphic models to date, exhibiting brain-like properties such as localized featur
e processing, lower dimensionality, and increased efficiency. TopoNets also pred
ict responses in the brain and replicate the key topographic signatures observed
 in the brain’s visual and language cortices, further bridging the gap between b
iological and artificial systems. This work establishes a robust and generalizab
le framework for integrating topography into AI, advancing the development of hi
gh performing models that more closely emulate the computational strategies of t
he human brain. Our project page: https://toponets.github.io
****
Benjamin Vandersmissen,Lucas Deckers,Jose Oramas
Improving Neural Network Accuracy by Concurrently Training with a Twin Network
Recently within Spiking Neural Networks, a method called Twin Network Augmentati
on (TNA) has been introduced. This technique claims to improve the validation ac
curacy of a Spiking Neural Network simply by training two networks in conjunctio
n and matching the logits via the Mean Squared Error loss. In this paper, we val
idate the viability of this method on a wide range of popular Convolutional Neur
al Network (CNN) benchmarks and compare this approach to existing Knowledge Dist
illation schemes. Next, we conduct a in-depth study of the different components
that make up TNA and determine that its effectiveness is not solely situated in
an increase of trainable parameters, but rather the effect of the training metho

dology. Finally, we analyse the representations learned by networks trained with TNA and highlight their superiority in a number of tasks, thus proving empirically the applicability of Twin Network Augmentation on CNN models.
****

Zhenwei Wang,Tengfei Wang,Zexin He,Gerhard Petrus Hancke,Ziwei Liu,Rynson W. H. Lau

Phidias: A Generative Model for Creating 3D Content from Text, Image, and 3D Conditions with Reference-Augmented Diffusion

Generative 3D modeling has made significant advances recently, but it remains constrained by its inherently ill-posed nature, leading to challenges in quality and controllability. Inspired by the real-world workflow that designers typically refer to existing 3D models when creating new ones, we propose Phidias, a novel generative model that uses diffusion for reference-augmented 3D generation. Given an image, our method leverages a retrieved or user-provided 3D reference model to guide the generation process, thereby enhancing the generation quality, generalization ability, and controllability. Phidias integrates three key components: 1) meta-ControlNet to dynamically modulate the conditioning strength, 2) dynamic reference routing to mitigate misalignment between the input image and 3D reference, and 3) self-reference augmentations to enable self-supervised training with a progressive curriculum. Collectively, these designs result in significant generative improvements over existing methods. Phidias forms a unified framework for 3D generation using text, image, and 3D conditions, offering versatile applications.
****

Ziyan Jiang,Rui Meng,Xinyi Yang,Semih Yavuz,Yingbo Zhou,Wenhu Chen

VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks

Embedding models play a crucial role in a variety of downstream tasks, including semantic similarity, information retrieval, and clustering. While there has been a surge of interest in developing universal text embedding models that generalize across tasks (e.g., MTEB), progress in learning universal multimodal embedding models has been comparatively slow, despite their importance and practical applications. In this work, we explore the potential of building universal multimodal embeddings capable of handling a broad range of downstream tasks.

Our contributions are twofold: (1) we propose MMEB (Massive Multimodal Embedding Benchmark), which covers four meta-tasks (classification, visual question answering, multimodal retrieval, and visual grounding) and 36 datasets, including 20 training datasets and 16 evaluation datasets spanning both in-distribution and out-of-distribution tasks, and (2) VLM2Vec (Vision-Language Model → Vector), a contrastive training framework that transforms any vision-language model into an embedding model through contrastive training on MMEB.

Unlike previous models such as CLIP and BLIP, which encode text and images independently without task-specific guidance, VLM2Vec can process any combination of images and text while incorporating task instructions to generate a fixed-dimensional vector. We develop a series of VLM2Vec models based on state-of-the-art VLMs, including Phi-3.5-V, LLaVA-1.6, and Qwen2-VL, and evaluate them on MMEB’s benchmark. With LoRA tuning, VLM2Vec achieves a 10% to 20% improvement over existing multimodal embedding models on MMEB’s evaluation sets. Our findings reveal that VLMs are surprisingly strong embedding models.
****

Leonardo Iurada,Marco Ciccone,Tatiana Tommasi

Efficient Model Editing with Task-Localized Sparse Fine-tuning

Task arithmetic has emerged as a promising approach for editing models by representing task-specific knowledge as composable task vectors. However, existing methods rely on network linearization to derive task vectors, leading to computational bottlenecks during training and inference. Moreover, linearization alone does not ensure weight disentanglement, the key property that enables conflict-free composition of task vectors. To address this, we propose TaLoS which allows to build sparse task vectors with minimal interference without requiring explicit l

inearization and sharing information across tasks. We find that pre-trained mode
ls contain a subset of parameters with consistently low gradient sensitivity acr
oss tasks, and that sparsely updating only these parameters allows for promoting
 weight disentanglement during fine-tuning. Our experiments prove that TaLoS imp
roves training and inference efficiency while outperforming current methods in t
ask addition and negation. By enabling modular parameter editing, our approach f
osters practical deployment of adaptable foundation models in real-world applica
tions.
****
Nadezhda Chirkova,Thibault Formal,Vassilina Nikoulina,Stéphane CLINCHANT
Provence: efficient and robust context pruning for retrieval-augmented generatio
n
Retrieval-Augmented Generation improves various aspects of large language models
 (LLMs) generation, but suffers from computational overhead caused by long cont
exts, and the propagation of irrelevant retrieved information into generated res
ponses. Context pruning deals with both aspects, by removing irrelevant parts of
 retrieved contexts before LLM generation. Existing context pruning approaches a
re limited, and do not present a universal model that would be both _efficient_
and _robust_ in a wide range of scenarios, e.g., when contexts contain a variabl
e amount of relevant information or vary in length, or when evaluated on various
 domains. In this work, we close this gap and introduce Provence (Pruning and Re
ranking Of retrieVEd relevaNt ContExts), an efficient and robust context pruner
for Question Answering, which dynamically detects the needed amount of pruning f
or a given context and can be used out-of-the-box for various domains. The three
 key ingredients of Provence are formulating the context pruning task as sequen
ce labeling, unifying context pruning capabilities with context reranking, and t
raining on diverse data. Our experimental results show that Provence enables con
text pruning with negligible to no drop in performance, in various domains and s
ettings, at almost no cost in a standard RAG pipeline. We also conduct a deeper
analysis alongside various ablations to provide insights into training context p
runers for future work.
****
Zhixiang Chi,Li Gu,Huan Liu,Ziqiang Wang,Yanan Wu,Yang Wang,Konstantinos N Plata
niotis
Learning to Adapt Frozen CLIP for Few-Shot Test-Time Domain Adaptation
Few-shot Test-Time Domain Adaptation focuses on adapting a model at test time to
 a specific domain using only a few unlabeled examples, addressing domain shift.
 Prior methods leverage CLIP's strong out-of-distribution (OOD) abilities by gen
erating domain-specific prompts to guide its generalized, frozen features. Howev
er, since downstream datasets are not explicitly seen by CLIP, solely depending
on the feature space knowledge is constrained by CLIP's prior knowledge. Notably
, when using a less robust backbone like ViT-B/16, performance significantly dro
ps on challenging real-world benchmarks. Departing from the state-of-the-art of
inheriting the intrinsic OOD capability of CLIP, this work introduces learning d
irectly on the input space to complement the dataset-specific knowledge for froz
en CLIP. Specifically, an independent side branch is attached in parallel with C
LIP and enforced to learn exclusive knowledge via revert attention. To better c
apture the dataset-specific label semantics for downstream adaptation, we propos
e to enhance the inter-dispersion among text features via greedy text ensemble a
nd refinement. The text and visual features are then progressively fused in a do
main-aware manner by a generated domain prompt to adapt toward a specific domain
. Extensive experiments show our method's superiority on 5 large-scale benchmark
s (WILDS and DomainNet), notably improving over smaller networks like ViT-B/16 w
ith gains of \textbf{+5.1} in F1 for iWildCam and \textbf{+3.1\%} in WC Acc for
FMoW. \href{https://github.com/chi-chi-zx/L2C}{Our Code: L2C}
****
Weijia Shi,Jaechan Lee,Yangsibo Huang,Sadhika Malladi,Jieyu Zhao,Ari Holtzman,Da
ogao Liu,Luke Zettlemoyer,Noah A. Smith,Chiyuan Zhang
MUSE: Machine Unlearning Six-Way Evaluation for Language Models
Language models (LMs) are trained on vast amounts of text data, which may includ

e private and copyrighted content. Data owners may request the removal of their data from a trained model due to privacy or copyright concerns. However, exactly unlearning only these datapoints (i.e., retraining with the data removed) is intractable in modern-day models. This has led to the development of many approximate unlearning algorithms. The evaluation of the efficacy of these algorithms has traditionally been narrow in scope, failing to precisely quantify the success and practicality of the algorithm from the perspectives of both the model deployers and the data owners. We address this issue by proposing MUSE, a comprehensive machine unlearning evaluation benchmark that enumerates six diverse desirable properties for unlearned models: (1) no verbatim memorization, (2) no knowledge memorization, (3) no privacy leakage, (4) utility preservation on data not intended for removal, (5) scalability with respect to the size of removal requests, and (6) sustainability over sequential unlearning requests. Using these criteria, we benchmark how effectively eight popular unlearning algorithms on 7B-parameter LMs can unlearn Harry Potter books and news articles. Our results demonstrate that most algorithms can prevent verbatim memorization and knowledge memorization to varying degrees, but only one algorithm does not lead to severe privacy leakage. Furthermore, existing algorithms fail to meet deployer's expectations because they often degrade general model utility and also cannot sustainably accommodate successive unlearning requests or large-scale content removal. Our findings identify key issues with the practicality of existing unlearning algorithms on language models.
****
Lefan Zhang,Xiaodan Wang,Yanhua Huang,Ruiwen Xu
Learning Harmonized Representations for Speculative Sampling
Speculative sampling is a promising approach to accelerate the decoding stage for Large Language Models (LLMs). Recent advancements that leverage target LLM's contextual information, such as hidden states and KV cache, have shown significant practical improvements. However, these approaches suffer from inconsistent context between training and decoding. We also observe another discrepancy between the training and decoding objectives in existing speculative sampling methods. In this work, we propose a solution named HArmonized Speculative Sampling (HASS) that learns harmonized representations to address these issues. HASS accelerates the decoding stage without adding inference overhead through harmonized objective distillation and harmonized context alignment. Experiments on four LLaMA models demonstrate that HASS achieves 2.81x-4.05x wall-clock time speedup ratio averaging across three datasets, surpassing EAGLE-2 by 8%-20%. The code is available at https://github.com/HArmonizedSS/HASS.
****
Yihong Luo,Xiaolong Chen,Xinghua Qu,Tianyang Hu,Jing Tang
You Only Sample Once: Taming One-Step Text-to-Image Synthesis by Self-Cooperative Diffusion GANs
Recently, some works have tried to combine diffusion and Generative Adversarial Networks (GANs) to alleviate the computational cost of the iterative denoising inference in Diffusion Models (DMs).
However, existing works in this line suffer from either training instability and mode collapse or subpar one-step generation learning efficiency.
To address these issues, we introduce YOSO, a novel generative model designed for rapid, scalable, and high-fidelity one-step image synthesis with high training stability and mode coverage.
Specifically, we smooth the adversarial divergence by the denoising generator itself, performing self-cooperative learning. We show that our method can serve as a one-step generation model training from scratch with competitive performance.

Moreover, we extend our YOSO to one-step text-to-image generation based on pre-trained models by several effective training techniques (i.e., latent perceptual loss and latent discriminator for efficient training along with the latent DMs; the informative prior initialization (IPI), and the quick adaption stage for fixing the flawed noise scheduler). Experimental results show that YOSO achieves the state-of-the-art one-step generation performance even with Low-Rank Adaptation

(LoRA) fine-tuning.
In particular, we show that the YOSO-PixArt-α can generate images in one step trained on 512 resolution, with the capability of adapting to 1024 resolution without extra explicit training, requiring only \textasciitilde10 A800 days for fine-tuning. Our code is available at: https://github.com/Luo-Yihong/YOSO
****
Matthew Chang,Gunjan Chhablani,Alexander Clegg,Mikael Dallaire Cote,Ruta Desai,Michal Hlavac,Vladimir Karashchuk,Jacob Krantz,Roozbeh Mottaghi,Priyam Parashar,Siddharth Patki,Ishita Prasad,Xavier Puig,Akshara Rai,Ram Ramrakhya,Daniel Tran,Joanne Truong,John M Turner,Eric Undersander,Tsung-Yen Yang
PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks
We present a benchmark for Planning And Reasoning Tasks in humaN-Robot collaboration (PARTNR) designed to study human-robot coordination in household activities. PARTNR tasks exhibit characteristics of everyday tasks, such as spatial, temporal, and heterogeneous agent capability constraints. We employ a semi-automated task generation pipeline using Large Language Models (LLMs), incorporating simulation-in-the-loop for the grounding and verification. PARTNR stands as the largest benchmark of its kind, comprising 100,000 natural language tasks, spanning 60 houses and 5,819 unique objects. We analyze state-of-the-art LLMs on PARTNR tasks, across the axes of planning, perception and skill execution. The analysis reveals significant limitations in SoTA models, such as poor coordination and failures in task tracking and recovery from errors. When LLMs are paired with 'real' humans, they require 1.5x as many steps as two humans collaborating and 1.1x more steps than a single human, underscoring the potential for improvement in these models. We further show that fine-tuning smaller LLMs with planning data can achieve performance on par with models 9 times larger, while being 8.6x faster at inference. Overall, PARTNR highlights significant challenges facing collaborative embodied agents and aims to drive research in this direction.
****
Yi Zhou,Yilai Li,Jing Yuan,Quanquan Gu
CryoFM: A Flow-based Foundation Model for Cryo-EM Densities
Cryo-electron microscopy (cryo-EM) is a powerful technique in structural biology and drug discovery, enabling the study of biomolecules at high resolution. Significant advancements by structural biologists using cryo-EM have led to the production of around 40k protein density maps at various resolutions. However, cryo-EM data processing algorithms have yet to fully benefit from our knowledge of biomolecular density maps, with only a few recent models being data-driven but limited to specific tasks. In this study, we present CryoFM, a foundation model designed as a generative model, learning the distribution of high-quality density maps and generalizing effectively to downstream tasks. Built on flow matching, CryoFM is trained to accurately capture the prior distribution of biomolecular density maps. Furthermore, we introduce a flow posterior sampling method that leverages CryoFM as a flexible prior for several downstream tasks in cryo-EM and cryo-electron tomography (cryo-ET) without the need for fine-tuning, achieving state-of-the-art performance on most tasks and demonstrating its potential as a foundational model for broader applications in these fields.
****
Yi-Kai Zhang,Shiyin Lu,Qing-Guo Chen,De-Chuan Zhan,Han-Jia Ye
ZooProbe: A Data Engine for Evaluating, Exploring, and Evolving Large-scale Training Data for Multimodal LLMs
Multimodal Large Language Models (MLLMs) are thriving through continuous fine-tuning by LLMs. Driven by the law that "scale is everything", MLLMs expand their training sets during version iterations. In this paper, we propose a large-scale training data engine built around an evaluating-exploring-evolving (E3) loop. Evaluating the data provides insights into its characteristics. Exploring quality rules helps identify which data enhances training. Together, these processes facilitate the systematic evolution of new, high-quality data. With the E3 loop, we introduce ZooProbe, an efficient data engine for MLLMs. First, the problem of data expansion is formalized as a tree of sampling and growth. ZooProbe introduce

s a small-scale model *zoo* to obtain comprehensive evaluations for child datase
ts. From multiple perspectives, visual, textual, and multimodal models cover ove
r 50 dimensions of intrinsic and meta attributes, such as object and topic distr
ibution, and higher-level properties, like annotation quality and scene complexi
ty. ZooProbe constructs based on A* search, modeling the heuristic functi
on as a quality estimate from data evaluation results. It dynamically explores t
he rule of data quality based on the model state of the *probe* datasets. Additi
onally, it evolves new targeted data with identified high-quality rules. We also
 develop an extra heuristic quality ranker with the data utilized and discarded
during the expansion. Our experiments show that ZooProbe significantly breaks th
e scaling law in multimodal instruction fine-tuning at scales of 260k and belo
w.
ZooProbe generates high-quality data that accelerates MLLM training and enhances
 performance, automating the evolution of large-scale training data.
****
Josh Alman,Hantao Yu
Fundamental Limitations on Subquadratic Alternatives to Transformers
The Transformer architecture is widely deployed in many popular and impactful La
rge Language Models. At its core is the attention mechanism for calculating corr
elations between pairs of tokens. Performing an attention computation takes quad
ratic time in the input size, and had become the time bottleneck for transformer
 operations. In order to circumvent this, researchers have used a variety of app
roaches, including designing heuristic algorithms for performing attention compu
tations faster, and proposing alternatives to the attention mechanism which can
be computed more quickly. For instance, state space models such as Mamba were d
esigned to replace attention with an almost linear time alternative.

In this paper, we prove that any such approach cannot perform important tasks th
at Transformer is able to perform (assuming a popular conjecture from fine-grain
ed complexity theory). We focus on document similarity tasks, where one is given
 as input many documents and would like to find a pair which is (approximately)
the most similar. We prove that Transformer is able to perform this task, and we
 prove that this task cannot be performed in truly subquadratic time by any algo
rithm. Thus, any model which can be evaluated in subquadratic time – whether bec
ause of subquadratic-time heuristics for attention, faster attention replacement
s like Mamba, or any other reason – cannot perform this task. In other words, in
 order to perform tasks that (implicitly or explicitly) involve document similar
ity, one may as well use Transformer and cannot avoid its quadratic running time
.
****
Yongxin Guo,Zhenglin Cheng,Xiaoying Tang,Zhaopeng Tu,Tao Lin
Dynamic Mixture of Experts: An Auto-Tuning Approach for Efficient Transformer Mo
dels
The Sparse Mixture of Experts (SMoE) has been widely employed to enhance the eff
iciency of training and inference for Transformer-based foundational models, yie
lding promising results. However, the performance of SMoE heavily depends on the
 choice of hyper-parameters, such as the number of experts and the number of exp
erts to be activated (referred to as top-k), resulting in significant computat
ional overhead due to the extensive model training by searching over various hyp
er-parameter configurations. As a remedy, we introduce the Dynamic Mixture of Ex
perts (DynMoE) technique. DynMoE incorporates (1) a novel gating method that ena
bles each token to automatically determine the number of experts to activate. (2
) An adaptive process automatically adjusts the number of experts during trainin
g. Extensive numerical results across Vision, Language, and Vision-Language task
s demonstrate the effectiveness of our approach to achieve competitive performan
ce compared to GMoE for vision and language tasks, and MoE-LLaVA for vision-lang
uage tasks, while maintaining efficiency by activating fewer parameters. Our cod
e is available at \url{https://github.com/LINs-lab/DynMoE}.
****
Jie Cheng,Ruixi Qiao,YINGWEI MA,Binhua Li,Gang Xiong,Qinghai Miao,Yongbin Li,Yis

heng Lv
Scaling Offline Model-Based RL via Jointly-Optimized World-Action Model Pretraining
A significant aspiration of offline reinforcement learning (RL) is to develop a generalist agent with high capabilities from large and heterogeneous datasets. However, prior approaches that scale offline RL either rely heavily on expert trajectories or struggle to generalize to diverse unseen tasks. Inspired by the excellent generalization of world model in conditional video generation, we explore the potential of image observation-based world model for scaling offline RL and enhancing generalization on novel tasks. In this paper, we introduce JOWA: Jointly-Optimized World-Action model, an offline model-based RL agent pretrained on multiple Atari games with 6 billion tokens data to learn general-purpose representation and decision-making ability. Our method jointly optimizes a world-action model through a shared transformer backbone, which stabilize temporal difference learning with large models during pretraining. Moreover, we propose a provably efficient and parallelizable planning algorithm to compensate for the Q-value estimation error and thus search out better policies. Experimental results indicate that our largest agent, with 150 million parameters, achieves 78.9% human-level performance on pretrained games using only 10% subsampled offline data, outperforming existing state-of-the-art large-scale offline RL baselines by 31.6% on averange. Furthermore, JOWA scales favorably with model capacity and can sample-efficiently transfer to novel games using only 5k offline fine-tuning data (approximately 4 trajectories) per game, demonstrating superior generalization.
****
Xiaofeng Yang,Chen Cheng,Xulei Yang,Fayao Liu,Guosheng Lin
Text-to-Image Rectified Flow as Plug-and-Play Priors
Large-scale diffusion models have achieved remarkable performance in generative tasks. Beyond their initial training applications, these models have proven their ability to function as versatile plug-and-play priors. For instance, 2D diffusion models can serve as loss functions to optimize 3D implicit models. Rectified Flow, a novel class of generative models, has demonstrated superior performance across various domains. Compared to diffusion-based methods, rectified flow approaches surpass them in terms of generation quality and efficiency. In this work, we present theoretical and experimental evidence demonstrating that rectified flow based methods offer similar functionalities to diffusion models — they can also serve as effective priors. Besides the generative capabilities of diffusion priors, motivated by the unique time-symmetry properties of rectified flow models, a variant of our method can additionally perform image inversion. Experimentally, our rectified flow based priors outperform their diffusion counterparts — the SDS and VSD losses — in text-to-3D generation. Our method also displays competitive performance in image inversion and editing. Code is available at: https://github.com/yangxiaofeng/rectified_flow_prior.
****
Xiangyu Dong,Xingyi Zhang,Lei Chen,Mingxuan Yuan,Sibo Wang
SpaceGNN: Multi-Space Graph Neural Network for Node Anomaly Detection with Extremely Limited Labels
Node Anomaly Detection (NAD) has gained significant attention in the deep learning community due to its diverse applications in real-world scenarios.
Existing NAD methods primarily embed graphs within a single Euclidean space, while overlooking the potential of non-Euclidean spaces.
Besides, to address the prevalent issue of limited supervision in real NAD tasks, previous methods tend to leverage synthetic data to collect auxiliary information, which is not an effective solution as shown in our experiments.
To overcome these challenges, we introduce a novel SpaceGNN model designed for NAD tasks with extremely limited labels.
Specifically, we provide deeper insights into a task-relevant framework by empirically analyzing the benefits of different spaces for node representations, based on which, we design a Learnable Space Projection function that effectively encodes nodes into suitable spaces.
Besides, we introduce the concept of weighted homogeneity, which we empirically

and theoretically validate as an effective coefficient during information propag
ation. This concept inspires the design of the Distance Aware Propagation module
.
Furthermore, we propose the Multiple Space Ensemble module, which extracts compr
ehensive information for NAD under conditions of extremely limited supervision.
Our findings indicate that this module is more beneficial than data augmentation
 techniques for NAD. Extensive experiments conducted on 9 real datasets confirm
the superiority of SpaceGNN, which outperforms the best rival by an average of 8
.55% in AUC and 4.31% in F1 scores. Our code is available at https://github.com/
xydong127/SpaceGNN.
****

Leonardo Ferreira Guilhoto,Paris Perdikaris
Deep Learning Alternatives Of The Kolmogorov Superposition Theorem
This paper explores alternative formulations of the Kolmogorov Superposition The
orem (KST) as a foundation for neural network design. The original KST formulati
on, while mathematically elegant, presents practical challenges due to its limit
ed insight into the structure of inner and outer functions and the large number
of unknown variables it introduces. Kolmogorov-Arnold Networks (KANs) leverage K
ST for function approximation, but they have faced scrutiny due to mixed results
 compared to traditional multilayer perceptrons (MLPs) and practical limitations
 imposed by the original KST formulation. To address these issues, we introduce
ActNet, a scalable deep learning model that builds on the KST and overcomes some
 of the drawbacks of Kolmogorov's original formulation. We evaluate ActNet in th
e context of Physics-Informed Neural Networks (PINNs), a framework well-suited f
or leveraging KST's strengths in low-dimensional function approximation, particu
larly for simulating partial differential equations (PDEs). In this challenging
setting, where models must learn latent functions without direct measurements, A
ctNet consistently outperforms KANs across multiple benchmarks and is competitiv
e against the current best MLP-based approaches. These results present ActNet as
 a promising new direction for KST-based deep learning applications, particularl
y in scientific computing and PDE simulation tasks.
****

Nimrod Shabtay,Felipe Maia Polo,Sivan Doveh,Wei Lin,Muhammad Jehanzeb Mirza,Lesh
em Choshen,Mikhail Yurochkin,Yuekai Sun,Assaf Arbelle,Leonid Karlinsky,Raja Giry
es
LiveXiv - A Multi-Modal live benchmark based on Arxiv papers content
The large-scale training of multi-modal models on data scraped from the web has
shown outstanding utility in infusing these models with the required world knowl
edge to perform effectively on multiple downstream tasks. However, one downside
of scraping data from the web can be the potential sacrifice of the benchmarks o
n which the abilities of these models are often evaluated. To safeguard against
test data contamination and to truly test the abilities of these foundation mode
ls we propose LiveXiv: A scalable evolving live benchmark based on scientific Ar
Xiv papers. LiveXiv accesses domain-specific manuscripts at any given timestamp
and proposes to automatically generate visual question-answer pairs (VQA). This
is done without any human-in-the-loop, using the multi-modal content in the manu
scripts, like graphs, charts, and tables. Moreover, we introduce an efficient ev
aluation approach that estimates the performance of all models on the evolving b
enchmark using evaluations of only a subset of models. This significantly reduce
s the overall evaluation cost. We benchmark multiple open and proprietary Large
Multi-modal Models (LMMs) on the first version of our benchmark, showing its cha
llenging nature and exposing the models’ true abilities, avoiding contamination.
 Lastly, in our commitment to high quality, we have collected and evaluated a ma
nually verified subset. By comparing its overall results to our automatic annota
tions, we have found that the performance variance is indeed minimal (<2.5%). Ou
r dataset is available online on HuggingFace, and our code is available here.
****

Silas Alberti,Kenan Hasanaliyev,Manav Shah,Stefano Ermon
Data Unlearning in Diffusion Models
Recent work has shown that diffusion models memorize and reproduce training data

examples. At the same time, large copyright lawsuits and legislation such as GDPR have highlighted the need for erasing datapoints from diffusion models. However, retraining from scratch is often too expensive. This motivates the setting of data unlearning, i.e., the study of efficient techniques for unlearning specific datapoints from the training set. Existing concept unlearning techniques require an anchor prompt/class/distribution to guide unlearning, which is not available in the data unlearning setting. General-purpose machine unlearning techniques were found to be either unstable or failed to unlearn data. We therefore propose a family of new loss functions called Subtracted Importance Sampled Scores (SISS) that utilize importance sampling and are the first method to unlearn data with theoretical guarantees. SISS is constructed as a weighted combination between simpler objectives that are responsible for preserving model quality and unlearning the targeted datapoints. When evaluated on CelebA-HQ and MNIST, SISS achieved Pareto optimality along the quality and unlearning strength dimensions. On Stable Diffusion, SISS successfully mitigated memorization on nearly 90% of the prompts we tested. We release our code online.
****

Zhichao Hou,MohamadAli Torkamani,Hamid Krim,Xiaorui Liu
Robustness Reprogramming for Representation Learning
This work tackles an intriguing and fundamental open challenge in representation learning: Given a well-trained deep learning model, can it be reprogrammed to enhance its robustness against adversarial or noisy input perturbations without altering its parameters?
To explore this, we revisit the core feature transformation mechanism in representation learning and propose a novel non-linear robust pattern matching technique as a robust alternative. Furthermore, we introduce three model reprogramming paradigms to offer flexible control of robustness under different efficiency requirements. Comprehensive experiments and ablation studies across diverse learning models ranging from basic linear model and MLPs to shallow and modern deep ConvNets demonstrate the effectiveness
of our approaches.
This work not only opens a promising and orthogonal direction for improving adversarial defenses in deep learning beyond existing methods but also provides new insights into designing more resilient AI systems with robust statistics.
Our implementation is available at https://github.com/chris-hzc/Robustness-Reprogramming.
****

Joao V. Cavalcanti,Laurent Lessard,Ashia C. Wilson
Adaptive backtracking for faster optimization
Backtracking line search is foundational in numerical optimization.
The basic idea is to adjust the step size of an algorithm by a {\em constant} factor until some chosen criterion (e.g. Armijo, Descent Lemma) is satisfied.
We propose a novel way to adjust step sizes, replacing the constant factor used in regular backtracking with one that takes into account the degree to which the chosen criterion is violated, with no additional computational burden. This light-weight adjustment leads to significantly faster optimization, which we confirm by performing a variety of experiments on over fifteen real world datasets.
For convex problems, we prove adaptive backtracking requires no more adjustments to produce a feasible step size than regular backtracking does.
For nonconvex smooth problems, we prove adaptive backtracking enjoys the same guarantees of regular backtracking. %same lower bounds that step sizes produced by regular backtracking do.
Furthermore, we prove adaptive backtracking preserves the convergence rates of gradient descent and its accelerated variant.
****

Taesun Yeom,Sangyoon Lee,Jaeho Lee
Fast Training of Sinusoidal Neural Fields via Scaling Initialization
Neural fields are an emerging paradigm that represent data as continuous functions parameterized by neural networks. Despite many advantages, neural fields often have a high training cost, which prevents a broader adoption. In this paper, w

e focus on a popular family of neural fields, called sinusoidal neural fields (S
NFs), and study how it should be initialized to maximize the training speed. We
find that the standard initialization scheme for SNFs---designed based on the si
gnal propagation principle---is suboptimal. In particular, we show that by simpl
y multiplying each weight (except for the last layer) by a constant, we can acce
lerate SNF training by 10$\times$. This method, coined _weight scaling_, consist
ently provides a significant speedup over various data domains, allowing the SNF
s to train faster than more recently proposed architectures. To understand why t
he weight scaling works well, we conduct extensive theoretical and empirical ana
lyses which reveal that the weight scaling not only resolves the spectral bias q
uite effectively but also enjoys a well-conditioned optimization trajectory.
****
Zhenchang Xing,Yang Liu,Zhuo Cheng,Qing Huang,Dehai Zhao,Daniel SUN,Chenhua Liu
When Prompt Engineering Meets Software Engineering: CNL-P as Natural and Robust
"APIs'' for Human-AI Interaction
With the growing capabilities of large language models (LLMs), they are increasi
ngly applied in areas like intelligent customer service, code generation, and kn
owledge management. Natural language (NL) prompts act as the ``APIs'' for human-
LLM interaction. To improve prompt quality, best practices for prompt engineerin
g (PE) have been developed, including writing guidelines and templates. Building
 on this, we propose Controlled NL for Prompt (CNL-P), which not only incorporat
es PE best practices but also draws on key principles from software engineering
(SE). CNL-P introduces precise grammar structures and strict semantic norms, fur
ther eliminating NL's ambiguity, allowing for a declarative but structured and a
ccurate expression of user intent. This helps LLMs better interpret and execute
the prompts, leading to more consistent and higher-quality outputs. We also intr
oduce an NL2CNL-P conversion tool based on LLMs, enabling users to write prompts
 in NL, which are then transformed into CNL-P format, thus lowering the learning
 curve of CNL-P. In particular, we develop a linting tool that checks CNL-P prom
pts for syntactic and semantic accuracy, applying static analysis techniques to
NL for the first time.
Extensive experiments demonstrate that CNL-P enhances the quality of LLM respons
es through the novel and organic synergy of PE and SE. We believe that CNL-P can
 bridge the gap between emerging PE and traditional SE, laying the foundation fo
r a new programming paradigm centered around NL.
****
Ashok Vardhan Makkuva,Marco Bondaschi,Adway Girish,Alliot Nagle,Martin Jaggi,Hye
ji Kim,Michael Gastpar
Attention with Markov: A Curious Case of Single-layer Transformers
Attention-based transformers have achieved tremendous success across a variety o
f disciplines including natural languages. To deepen our understanding of their
sequential modeling capabilities, there is a growing interest in using Markov in
put processes to study them. A key finding is that when trained on first-order M
arkov chains, transformers with two or more layers consistently develop an induc
tion head mechanism to estimate the in-context bigram conditional distribution.
In contrast, single-layer transformers, unable to form an induction head, direct
ly learn the Markov kernel but often face a surprising challenge: they become tr
apped in local minima representing the unigram distribution, whereas deeper mode
ls reliably converge to the ground-truth bigram. While single-layer transformers
 can theoretically model first-order Markov chains, their empirical failure to l
earn this simple kernel in practice remains a curious phenomenon. To explain thi
s contrasting behavior of single-layer models, in this paper we introduce a new
framework for a principled analysis of transformers via Markov chains. Leveragin
g our framework, we theoretically characterize the loss landscape of single-laye
r transformers and show the existence of global minima (bigram) and bad local mi
nima (unigram) contingent on data properties and model architecture. We precisel
y delineate the regimes under which these local optima occur. Backed by experime
nts, we demonstrate that our theoretical findings are in congruence with the emp
irical results. Finally, we outline several open problems in this arena.
****

Tianxin Huang,Zhiwen Yan,Yuyang Zhao,Gim Hee Lee
ComPC: Completing a 3D Point Cloud with 2D Diffusion Priors
3D point clouds directly collected from objects through sensors are often incomp
lete due to self-occlusion. Conventional methods for completing these partial po
int clouds rely on manually organized training sets and are usually limited to o
bject categories seen during training. In this work, we propose a test-time fram
ework for completing partial point clouds across unseen categories without any r
equirement for training. Leveraging point rendering via Gaussian Splatting, we d
evelop techniques of Partial Gaussian Initialization, Zero-shot Fractal Completi
on, and Point Cloud Extraction that utilize priors from pre-trained 2D diffusion
 models to infer missing regions and extract uniform completed point clouds. Exp
erimental results on both synthetic and real-world scanned point clouds demonstr
ate that our approach outperforms existing methods in completing a variety of ob
jects. Our project page is at \url{https://tianxinhuang.github.io/projects/ComPC
/}.
****
Jie Zhang,Dongrui Liu,Chen Qian,Linfeng Zhang,Yong Liu,Yu Qiao,Jing Shao
REEF: Representation Encoding Fingerprints for Large Language Models
Protecting the intellectual property of open-source Large Language Models (LLMs)
 is very important, because training LLMs costs extensive computational resource
s and data. Therefore, model owners and third parties need to identify whether a
 suspect model is a subsequent development of the victim model. To this end, we
propose a training-free REEF to identify the relationship between the suspect an
d victim models from the perspective of LLMs' feature representations. Specifica
lly, REEF computes and compares the centered kernel alignment similarity between
 the representations of a suspect model and a victim model on the same samples.
This training-free REEF does not impair the model's general capabilities and is
robust to sequential fine-tuning, pruning, model merging, and permutations. In t
his way, REEF provides a simple and effective way for third parties and models'
owners to protect LLMs' intellectual property together. Our code is publicly ac
cessible at https://github.com/AI45Lab/REEF.
****
Daniel Herbst,Stefanie Jegelka
Higher-Order Graphon Neural Networks: Approximation and Cut Distance
Graph limit models, like *graphons* for limits of dense graphs, have recently be
en used to study size transferability of graph neural networks (GNNs). While mos
t literature focuses on message passing GNNs (MPNNs), in this work we attend to
the more powerful *higher-order* GNNs. First, we extend the k-WL test for grap
hons (Böker, 2023) to the graphon-signal space and introduce *signal-weighted ho
momorphism densities* as a key tool. As an exemplary focus, we generalize *Invar
iant Graph Networks* (IGNs) to graphons, proposing *Invariant Graphon Networks*
(IWNs) defined via a subset of the IGN basis corresponding to bounded linear ope
rators. Even with this restricted basis, we show that IWNs of order k are at l
east as powerful as the k-WL test, and we establish universal approximation re
sults for graphon-signals in L^p distances. This significantly extends the pri
or work of Cai & Wang (2022), showing that IWNs—a subset of their *IGN-small*—re
tain effectively the same expressivity as the full IGN basis in the limit. In co
ntrast to their approach, our blueprint of IWNs also aligns better with the geom
etry of graphon space, for example facilitating comparability to MPNNs. We highl
ight that, while typical higher-order GNNs are discontinuous w.r.t.\ cut distanc
e—which causes their lack of convergence and is inherently tied to the definitio
n of k-WL—their transferability remains comparable to MPNNs.
****
Jun Xia,Sizhe Liu,Jingbo Zhou,Shaorong Chen,hongxin xiang,Zicheng Liu,Yue Liu,St
an Z. Li
Bridging the Gap between Database Search and \emph{De Novo} Peptide Sequencing w
ith SearchNovo
Accurate protein identification from mass spectrometry (MS) data is fundamental
to unraveling the complex roles of proteins in biological systems, with peptide
sequencing being a pivotal step in this process. The two main paradigms for pept

ide sequencing are database search, which matches experimental spectra with pept
ide sequences from databases, and \emph{de novo} sequencing, which infers peptid
e sequences directly from MS without relying on pre-constructed database. Althou
gh database search methods are highly accurate, they are limited by their inabil
ity to identify novel, modified, or mutated peptides absent from the database. I
n contrast, \emph{de novo} sequencing is adept at discovering novel peptides but
 often struggles with missing peaks issue, further leading to lower precision. W
e introduce SearchNovo, a novel framework that synergistically integrates the st
rengths of database search and \emph{de novo} sequencing to enhance peptide sequ
encing. SearchNovo employs an efficient search mechanism to retrieve the most si
milar peptide spectrum match (PSM) from a database for each query spectrum, foll
owed by a fusion module that utilizes the reference peptide sequence to guide th
e generation of the target sequence. Furthermore, we observed that dissimilar (n
oisy) reference peptides negatively affect model performance. To mitigate this,
we constructed pseudo reference PSMs to minimize their impact. Comprehensive eva
luations on multiple datasets reveal that SearchNovo significantly outperforms s
tate-of-the-art models. Also, analysis indicates that many retrieved spectra con
tain missing peaks absent in the query spectra, and the retrieved reference pept
ides often share common fragments with the target peptides. These are key elemen
ts in the recipe for SearchNovo’s success. The code is available at: \textcolor{
magenta}{\url{https://github.com/junxia97/SearchNovo}}.
****
Bishwamittra Ghosh,Sarah Hasan,Naheed Anjum Arafat,Arijit Khan
Logical Consistency of Large Language Models in Fact-Checking
In recent years, large language models (LLMs) have demonstrated significant succ
ess in performing varied natural language tasks such as language translation, qu
estion-answering, summarizing, fact-checking, etc. Despite LLMs’ impressive abil
ity to generate human-like texts, LLMs are infamous for their inconsistent respo
nses – a meaning-preserving change in the input query results in an inconsistent
 response and attributes to vulnerabilities of LLMs such as hallucination. Conse
quently, existing research focuses on simple paraphrasing-based consistency asse
ssment of LLMs, and ignores complex queries that necessitate an even better unde
rstanding of logical reasoning by an LLM. *Our work therefore addresses the logi
cal inconsistency of LLMs under complex logical queries with primitive logical o
perators, e.g., negation, conjunction, and disjunction.* As a test bed, we consi
der retrieval-augmented LLMs on a fact-checking task involving propositional log
ic queries from knowledge graphs (KGs). Our contributions are three-fold. **Benc
hmark:** We introduce three logical fact-checking datasets over KGs for communit
y development towards logically consistent LLMs. **Assessment:** We propose cons
istency measures of LLMs on propositional logic queries and demonstrate that exi
sting LLMs lack logical consistency, especially on complex queries. **Improvemen
t:** We employ supervised fine-tuning to improve the logical consistency of LLMs
 on the complex fact-checking task with KG contexts. We have made our source cod
e and benchmarks available.
****
Benedikt Alkin,Maximilian Beck,Korbinian Pöppel,Sepp Hochreiter,Johannes Brandst
etter
Vision-LSTM: xLSTM as Generic Vision Backbone
Transformers are widely used as generic backbones in computer vision, despite in
itially introduced for natural language processing. Recently, the Long Short-Ter
m Memory (LSTM) has been extended to a scalable and performant architecture - th
e xLSTM - which overcomes long-standing LSTM limitations via exponential gating
and parallelizable matrix memory structure. In this paper, we introduce Vision-L
STM (ViL), an adaption of the xLSTM building blocks to computer vision. ViL comp
rises a stack of xLSTM blocks where odd blocks process the sequence of patch tok
ens from top to bottom while even blocks go from bottom to top.

ViL achieves strong performances on classification, transfer learning and segmen
tation tasks as well as a beneficial pre-training cost-to-performance trade-off.
 Experiments show that ViL holds promise to be further deployed as new generic b

ackbone for computer vision architectures.
****
Xiaofeng Lin,Chenheng Xu,Matthew Yang,Guang Cheng
CTSyn: A Foundation Model for Cross Tabular Data Generation
Generative Foundation Models (GFMs) have achieved remarkable success in producin g high-quality synthetic data for images and text. However, their application to tabular data presents significant challenges due to the heterogeneous nature of table features. Current cross-table learning frameworks struggle because they l ack a generative model backbone and an effective mechanism to decode heterogeneo us feature values. To address these challenges, we propose the Cross-Table Synth esizer (CTSyn), a diffusion-based generative foundation model for tabular data g eneration. CTSyn comprises two key components. The first is an autoencoder netwo rk that consolidates diverse tables into a unified latent space. It dynamically reconstructs table values using a table schema embedding, allowing adaptation to heterogeneous datasets. The second is a conditional latent diffusion model that generates samples from the learned latent space, conditioned on the table schem a. Through large-scale pre-training, CTSyn outperforms existing table synthesize rs on standard benchmarks in both utility and diversity. These results position CTSyn as a promising framework for synthetic table generation and lay the groun dwork for developing large-scale tabular foundation models.
****
Myungseo Song,Jin-Woo Park,Jong-Seok Lee
Exploring the Camera Bias of Person Re-identification
We empirically investigate the camera bias of person re-identification (ReID) mo dels. Previously, camera-aware methods have been proposed to address this issue, but they are largely confined to training domains of the models. We measure the camera bias of ReID models on unseen domains and reveal that camera bias become s more pronounced under data distribution shifts. As a debiasing method for unse en domain data, we revisit feature normalization on embedding vectors. While the normalization has been used as a straightforward solution, its underlying cause s and broader applicability remain unexplored. We analyze why this simple method is effective at reducing bias and show that it can be applied to detailed bias factors such as low-level image properties and body angle. Furthermore, we valid ate its generalizability across various models and benchmarks, highlighting its potential as a simple yet effective test-time postprocessing method for ReID. In addition, we explore the inherent risk of camera bias in unsupervised learning of ReID models. The unsupervised models remain highly biased towards camera labe ls even for seen domain data, indicating substantial room for improvement. Based on observations of the negative impact of camera-biased pseudo labels on traini ng, we suggest simple training strategies to mitigate the bias. By applying thes e strategies to existing unsupervised learning algorithms, we show that signific ant performance improvements can be achieved with minor modifications.
****
Malyaban Bal,Abhronil Sengupta
P-SPIKESSM: HARNESSING PROBABILISTIC SPIKING STATE SPACE MODELS FOR LONG-RANGE D EPENDENCY TASKS
Spiking neural networks (SNNs) are posited as a computationally efficient and bi ologically plausible alternative to conventional neural architectures, with thei r core computational framework primarily using the leaky integrate-and-fire (LIF) neuron model. However, the limited hidden state representation of LIF neurons, characterized by a scalar membrane potential, and sequential spike generation p rocess, poses challenges for effectively developing scalable spiking models to a ddress long-range dependencies in sequence learning tasks. In this study, we de velop a scalable probabilistic spiking learning framework for long-range depende ncy tasks leveraging the fundamentals of state space models. Unlike LIF neurons that rely on the deterministic Heaviside function for a sequential process of sp ike generation, we introduce a SpikeSampler layer that samples spikes stochastic ally based on an SSM-based neuronal model while allowing parallel computations. To address non-differentiability of the spiking operation and enable effective t raining, we also propose a surrogate function tailored for the stochastic nature

of the SpikeSampler layer. To enhance inter-neuron communication, we introduce the SpikeMixer block, which integrates spikes from neuron populations in each layer. This is followed by a ClampFuse layer, incorporating a residual connection to capture complex dependencies, enabling scalability of the model. Our models attain state-of-the-art performance among SNN models across diverse long-range dependency tasks, encompassing the Long Range Arena benchmark, permuted sequential MNIST, and the Speech Command dataset and demonstrate sparse spiking pattern highlighting its computational efficiency.
****
Lanxiang Hu,Qiyu Li,Anze Xie,Nan Jiang,Ion Stoica,Haojian Jin,Hao Zhang
GameArena: Evaluating LLM Reasoning through Live Computer Games
Evaluating the reasoning abilities of large language models (LLMs) is challenging. Existing benchmarks often depend on static datasets, which are vulnerable to data contamination and may get saturated over time, or on binary live human feedback that conflates reasoning with other abilities. As the most prominent dynamic benchmark, Chatbot Arena evaluates open-ended questions in real-world settings, but lacks the granularity in assessing specific reasoning capabilities. We introduce GameArena, a dynamic benchmark designed to evaluate LLM reasoning capabilities through interactive gameplay with humans. GameArena consists of three games designed to test specific reasoning capabilities (e.g., deductive and inductive reasoning), while keeping participants entertained and engaged. We analyze the gaming data retrospectively to uncover the underlying reasoning processes of LLMs and measure their fine-grained reasoning capabilities. We collect over 2000 game sessions and provide detailed assessments of various reasoning capabilities for five state-of-the-art LLMs. Our user study with 100 participants suggests that GameArena improves user engagement compared to Chatbot Arena. For the first time, GameArena enables the collection of step-by-step LLM reasoning data in the wild.
****
Guoxiong Gao,Yutong Wang,Jiedong Jiang,Qi Gao,Zihan Qin,Tianyi Xu,Bin Dong
Herald: A Natural Language Annotated Lean 4 Dataset
Verifiable formal languages like Lean have profoundly impacted mathematical reasoning, particularly through the use of large language models (LLMs) for automated reasoning. A significant challenge in training LLMs for these formal languages is the lack of parallel datasets that align natural language with formal language proofs. To address this challenge, this paper introduces a novel framework for translating the Mathlib4 corpus (a unified library of mathematics in formal language Lean 4) into natural language. Building upon this, we employ a dual augmentation strategy that combines tactic-based and informal-based approaches, leveraging the Lean-jixia system, a Lean 4 analyzer. We present the results of this pipeline on Mathlib4 as Herald (Hierarchy and Retrieval-based Translated Lean Dataset). We also propose the Herald Translator, which is fine-tuned on Herald. Herald translator achieves a 96.7\% accuracy (Pass@128) on formalizing statements in the miniF2F-test and a 23.5\% accuracy on our internal graduate-level textbook dataset, outperforming InternLM2-Math-Plus-7B (73.0\% and 7.5\%) and TheoremLlama (50.1\% and 4.0\%). Furthermore, we propose a section-level translation framework for real-world applications. As a direct application of Herald translator, we have successfully translated a template section in the Stack project, marking a notable progress in the automatic formalization of graduate-level mathematical literature. Our model, along with the datasets, are open-sourced to the public.
****
Yury Gorishniy,Akim Kotelnikov,Artem Babenko
TabM: Advancing tabular deep learning with parameter-efficient ensembling
Deep learning architectures for supervised learning on tabular data range from simple multilayer perceptrons (MLP) to sophisticated Transformers and retrieval-augmented methods.
This study highlights a major, yet so far overlooked opportunity for substantially improving tabular MLPs; namely, parameter-efficient ensembling -- a paradigm for imitating an ensemble of models with just one model.

We start by describing TabM -- a simple model based on MLP and BatchEnsemble (an existing technique), improved with our custom modifications.
Then, we perform a large scale evaluation of tabular DL architectures on public benchmarks in terms of both task performance and efficiency, which renders the landscape of tabular DL in a new light.
In particular, we find that TabM outperforms prior tabular DL models, while the complexity of attention- and retrieval-based methods does not pay off.
Lastly, we conduct a detailed empirical analysis, that sheds some light on the high performance of TabM.
For example, we show that parameter-efficient ensembling is not an arbitrary trick, but rather a highly effective way to reduce overfitting and improve optimization dynamics of tabular MLPs.
Overall, our work brings an impactful technique to tabular DL, analyses its behaviour, and advances the performance-efficiency tradeoff with TabM -- a simple and powerful baseline for researchers and practitioners.
****
Zhuang Liu,Kaiming He
A Decade's Battle on Dataset Bias: Are We There Yet?
We revisit the ``dataset classification'' experiment suggested by Torralba & Efros (2011) a decade ago, in the new era with large-scale, diverse, and hopefully less biased datasets as well as more capable neural network architectures. Surprisingly, we observe that modern neural networks can achieve excellent accuracy in classifying which dataset an image is from: e.g., we report 84.7% accuracy on held-out validation data for the three-way classification problem consisting of the YFCC, CC, and DataComp datasets. Our further experiments show that such a dataset classifier could learn semantic features that are generalizable and transferable, which cannot be explained by memorization. We hope our discovery will inspire the community to rethink issues involving dataset bias.
****
Julián Tachella,Mike Davies,Laurent Jacques
UNSURE: self-supervised learning with Unknown Noise level and Stein's Unbiased Risk Estimate
Recently, many self-supervised learning methods for image reconstruction have been proposed that can learn from noisy data alone, bypassing the need for ground-truth references. Most existing methods cluster around two classes: i) Stein's Unbiased Risk Estimate (SURE) and similar approaches that assume full knowledge of the noise distribution, and ii) Noise2Self and similar cross-validation methods that require very mild knowledge about the noise distribution. The first class of methods tends to be impractical, as the noise level is often unknown in real-world applications, and the second class is often suboptimal compared to supervised learning.
In this paper, we provide a theoretical framework that characterizes this expressivity-robustness trade-off and propose a new approach based on SURE, but unlike the standard SURE, does not require knowledge about the noise level. Throughout a series of experiments, we show that the proposed estimator outperforms other existing self-supervised methods on various imaging inverse problems.
****
Junhao Zheng,Xidi Cai,Shengjie Qiu,Qianli Ma
Spurious Forgetting in Continual Learning of Language Models
Recent advancements in large language models (LLMs) reveal a perplexing phenomenon in continual learning: despite extensive training, models experience significant performance declines, raising questions about task alignment and underlying knowledge retention. This study first explores the concept of "spurious forgetting", proposing that such performance drops often reflect a decline in task alignment rather than true knowledge loss. Through controlled experiments with a synthesized dataset, we investigate the dynamics of model performance during the initial training phases of new tasks, discovering that early optimization steps can disrupt previously established task alignments. Our theoretical analysis connects these shifts to orthogonal updates in model weights, providing a robust framework for understanding this behavior. Ultimately, we introduce a Freezing strate

gy that fix the bottom layers of the model, leading to substantial improvements in four continual learning scenarios. Our findings underscore the critical distinction between task alignment and knowledge retention, paving the way for more effective strategies in continual learning.
****
Ningyuan Zhang,Jie Lu,Keqiuyin Li,Zhen Fang,Guangquan Zhang
Release the Powers of Prompt Tuning: Cross-Modality Prompt Transfer
Prompt Tuning adapts frozen models to new tasks by prepending a few learnable embeddings to the input.
However, it struggles with tasks that suffer from data scarcity.
To address this, we explore Cross-Modality Prompt Transfer, leveraging prompts pretrained on a data-rich modality to improve performance on data-scarce tasks in another modality.
As a pioneering study, we first verify the feasibility of cross-modality prompt transfer by directly applying frozen source prompts (trained on the source modality) to the target modality task.
To empirically study cross-modality prompt transferability, we train a linear layer to adapt source prompts to the target modality, thereby boosting performance and providing ground-truth transfer results.
Regarding estimating prompt transferability, existing methods show ineffectiveness in cross-modality scenarios where the gap between source and target tasks is larger.
We address this by decomposing the gap into the modality gap and the task gap, which we measure separately to estimate the prompt transferability more accurately.
Additionally, we propose Attention Transfer to further reduce the gaps by injecting target knowledge into the prompt and reorganizing a top-transferable source prompt using an attention block.
We conduct extensive experiments involving prompt transfer from 13 source language tasks to 19 target vision tasks under three settings.
Our findings demonstrate that:
(i) cross-modality prompt transfer is feasible, supported by in-depth analysis;
(ii) measuring both the modality and task gaps is crucial for accurate prompt transferability estimation, a factor overlooked by previous studies;
(iii) cross-modality prompt transfer can significantly release the powers of prompt tuning on data-scarce tasks, as evidenced by comparisons with a newly released prompt-based benchmark.
****
Dario Shariatian,Umut Simsekli,Alain Oliviero Durmus
Denoising Levy Probabilistic Models
Investigating noise distributions beyond Gaussian in diffusion generative models remains an open challenge. The Gaussian case has been a large success experimentally and theoretically, admitting a unified stochastic differential equation (SDE) framework, encompassing score-based and denoising formulations. Recent studies have investigated the potential of \emph{heavy-tailed} noise distributions to mitigate mode collapse and effectively manage datasets exhibiting class imbalance, heavy tails, or prominent outliers.
Very recently, Yoon et al.\ (NeurIPS 2023), presented the Levy-Ito model (LIM), directly extending the SDE-based framework to a class of heavy-tailed SDEs, where the injected noise followed an α-stable distribution -- a rich class of heavy-tailed distributions.
Despite its theoretical elegance and performance improvements, LIM relies on highly involved mathematical techniques, which may limit its accessibility and hinder its broader adoption and further development.
In this study, we take a step back, and instead of starting from the SDE formulation, we extend the denoising diffusion probabilistic model (DDPM) by directly replacing the Gaussian noise with α-stable noise.
By using only elementary proof techniques, we show that the proposed approach, \emph{denoising L\'{e}vy probabilistic model} (DLPM) algorithmically boils down to running vanilla DDPM with minor modifications, hence allowing the use of exist

ing implementations with minimal changes.
Remarkably, as opposed to the Gaussian case, DLPM and LIM yield different training algorithms and different backward processes, leading to distinct sampling algorithms.
This fundamental difference translates favorably for the performance of DLPM in various aspects: our experiments show that DLPM achieves better coverage of the tails of the data distribution, better generation of unbalanced datasets, and improved computation times requiring significantly smaller number of backward steps.
****

Kazuki Adachi,Shin'ya Yamaguchi,Atsutoshi Kumagai,Tomoki Hamagami
Test-time Adaptation for Regression by Subspace Alignment
This paper investigates test-time adaptation (TTA) for regression, where a regression model pre-trained in a source domain is adapted to an unknown target distribution with unlabeled target data.
Although regression is one of the fundamental tasks in machine learning, most of the existing TTA methods have classification-specific designs, which assume that models output class-categorical predictions, whereas regression models typically output only single scalar values.
To enable TTA for regression, we adopt a feature alignment approach, which aligns the feature distributions between the source and target domains to mitigate the domain gap.
However, we found that naive feature alignment employed in existing TTA methods for classification is ineffective or even worse for regression because the features are distributed in a small subspace and many of the raw feature dimensions have little significance to the output.
For an effective feature alignment in TTA for regression, we propose Significant-subspace Alignment (SSA).
SSA consists of two components: subspace detection and dimension weighting.
Subspace detection finds the feature subspace that is representative and significant to the output.
Then, the feature alignment is performed in the subspace during TTA.
Meanwhile, dimension weighting raises the importance of the dimensions of the feature subspace that have greater significance to the output.
We experimentally show that SSA outperforms various baselines on real-world datasets.
****

Jingrong Wei,Long Chen
Accelerated Over-Relaxation Heavy-Ball Method: Achieving Global Accelerated Convergence with Broad Generalization
The heavy-ball momentum method accelerates gradient descent with a momentum term but lacks accelerated convergence for general smooth strongly convex problems.
This work introduces the Accelerated Over-Relaxation Heavy-Ball (AOR-HB) method, the first variant with provable global and accelerated convergence for such problems. AOR-HB closes a long-standing theoretical gap, extends to composite convex optimization and min-max problems, and achieves optimal complexity bounds. It offers three key advantages: (1) broad generalization ability, (2) potential to reshape acceleration techniques, and (3) conceptual clarity and elegance compared to existing methods.
****

Chunjin Song,Zhijie Wu,Shih-Yang Su,Bastian Wandt,Leonid Sigal,Helge Rhodin
Locality Sensitive Avatars From Video
We present locality-sensitive avatar, a neural radiance field (NeRF) based network to learn human motions from monocular videos. To this end, we estimate a canonical representation between different frames of a video with a non-linear mapping from observation to canonical space, which we decompose into a skeletal rigid motion and a non-rigid counterpart. Our key contribution is to retain fine-grained details by modeling the non-rigid part with a graph neural network (GNN) that keeps the pose information local to neighboring body parts. Compared to former canonical representation based methods which solely operate on the coordinates

pace of a whole shape, our locality-sensitive motion modeling can reproduce both realistic shape contours and vivid fine-grained details. We evaluate on ZJU-MoCap, SynWild, ActorsHQ, MVHumanNet and various outdoor videos. The experiments reveal that with the locality sensitive deformation to canonical feature space, we are the first to achieve state-of-the-art results across novel view synthesis, novel pose animation and 3D shape reconstruction simultaneously. Our code is available at https://github.com/ChunjinSong/lsavatar.
****
Kaijing Ma,Xeron Du,Yunran Wang,Haoran Zhang,ZhoufutuWen,Xingwei Qu,Jian Yang,Jiaheng Liu,minghao liu,Xiang Yue,Wenhao Huang,Ge Zhang
KOR-Bench: Benchmarking Language Models on Knowledge-Orthogonal Reasoning Tasks
In this paper, we introduce Knowledge-Orthogonal Reasoning (KOR), a concept aimed at minimizing reliance on domain-specific knowledge, enabling more accurate evaluation of models' reasoning abilities in out-of-distribution settings.
Based on this concept, we propose the Knowledge-Orthogonal Reasoning Benchmark (KOR-Bench), encompassing five task categories: Operation, Logic, Cipher, Puzzle, and Counterfactual.
KOR-Bench emphasizes models' effectiveness in applying new rule descriptions to solve novel rule-driven questions. O1-Preview and O1-Mini achieve accuracies of 72.88\% and 70.16\%, surpassing Claude-3.5-Sonnet and GPT-4o (58.96\% and 58.00\%), highlighting the effectiveness of KOR-Bench.
We perform detailed analyses, identifying bottlenecks in the Cipher task with Stepwise Prompting, where two rounds of Self-Correction yield optimal results.
We evaluate performance across three integrated tasks, explore the impact of Tricks on the Puzzle task, and visualize rule-focused attention. Additionally, we conduct an ablation study on dataset size, benchmark correlations, and zero-shot and three-shot "only questions" experiments.
KOR-Bench aims to enhance reasoning evaluation and support further research in this area.
****
George Wang,Jesse Hoogland,Stan van Wingerden,Zach Furman,Daniel Murfet
Differentiation and Specialization of Attention Heads via the Refined Local Learning Coefficient
We introduce refined variants of the Local Learning Coefficient (LLC), a measure of model complexity grounded in singular learning theory, to study the development of internal structure in transformer language models during training. By applying these refined LLCs (rLLCs) to individual components of a two-layer attention-only transformer, we gain novel insights into the progressive differentiation and specialization of attention heads. Our methodology reveals how attention heads differentiate into distinct functional roles over the course of training, analyzes the types of data these heads specialize to process, and discovers a previously unidentified multigram circuit. These findings demonstrate that rLLCs provide a principled, quantitative toolkit for developmental interpretability, which aims to understand models through their evolution across the learning process. This work advances the field of developmental interpretability by providing a mathematically rigorous approach to understanding neural networks through the lens of their learning process. More broadly, this work takes a step towards establishing the correspondence between data distributional structure, geometric properties of the loss landscape, learning dynamics, and emergent computational structures in neural networks.
****
Rohan Deb,Mohammad Ghavamzadeh,Arindam Banerjee
Conservative Contextual Bandits: Beyond Linear Representations
Conservative Contextual Bandits (CCBs) address safety in sequential decision making by requiring that an agent's policy, along with minimizing regret, also satisfies a safety constraint: the performance is not worse than a baseline policy (e.g., the policy that the company has in production) by more than $(1+\alpha)$ factor.
Prior work developed UCB-style
algorithms for this problem in the multi-armed (Wu et al., 2016) and contextual

linear (Kazerouni et al., 2017) settings.
However, in practice the cost of the arms
is often a non-linear function, and therefore existing UCB algorithms are ineffe
ctive in such settings.
In this paper, we consider CCBs beyond the linear case and develop two algorithm
s $\mathtt{C\text{-}SquareCB}$ and $\mathtt{C\text{-}FastCB}$, using Inverse Gap
 Weighting (IGW) based exploration and an online regression oracle.
We show that the safety constraint is satisfied in high probability and that the
 regret for $\mathtt{C\text{-}SquareCB}$ is sub-linear in horizon T, while the
 the regret for $\mathtt{C\text{-}FastCB}$ is first-order and is sub-linear in $
L^*$, the cumulative loss of the optimal policy.
Subsequently, we use a neural network for function approximation and online grad
ient descent as the regression oracle to provide $\tilde{\mathcal{O}}\big(\sqrt{
KT} + K/\alpha\big) $ and $\tilde{\mathcal{O}}\big(\sqrt{KL^*} + K (1 + 1/\alpha
)\big)$ regret bounds respectively.
Finally, we demonstrate the efficacy of our algorithms on real world data, and s
how that they significantly outperform the existing baseline while maintaining t
he performance guarantee.
****
Hao Zhao,Maksym Andriushchenko,Francesco Croce,Nicolas Flammarion
Is In-Context Learning Sufficient for Instruction Following in LLMs?
In-context learning (ICL) allows LLMs to learn from examples without changing th
eir weights: this is a particularly promising capability for long-context LLMs t
hat can potentially learn from many examples. Recently, Lin et al. (2024) propos
ed URIAL, a method using only three in-context examples to align base LLMs, achi
eving non-trivial instruction following performance. In this work, we show that,
 while effective, ICL alignment with URIAL still underperforms compared to instr
uction fine-tuning on established benchmarks such as MT-Bench and AlpacaEval 2.0
 (LC), especially with more capable base LLMs. We then uncover the most relevant
 elements for successful in-context alignment, finding the crucial role of the d
ecoding parameters. Based on these insights, we show that the approach of URIAL
can indeed be improved by adding more, potentially carefully selected, high-qual
ity demonstrations in context, getting closer to the performance of instruct mod
els. Finally, we provide the first, to our knowledge, systematic comparison of I
CL and instruction fine-tuning (IFT) for instruction following in the low data r
egime. Overall, our work advances the understanding of ICL as an alignment techn
ique and its relationship to IFT.
****
Xiangpeng Yang,Linchao Zhu,Hehe Fan,Yi Yang
VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing
Recent advancements in diffusion models have significantly improved video genera
tion and editing capabilities. However, multi-grained video editing, which encom
passes class-level, instance-level, and part-level modifications, remains a form
idable challenge. The major difficulties in multi-grained editing include semant
ic misalignment of text-to-region control and feature coupling within the diffus
ion model. To address these difficulties, we present VideoGrain, a zero-shot app
roach that modulates space-time (cross- and self-) attention mechanisms to achie
ve fine-grained control over video content. We enhance text-to-region control by
 amplifying each local prompt's attention to its corresponding spatial-disentang
led region while minimizing interactions with irrelevant areas in cross-attentio
n. Additionally, we improve feature separation by increasing intra-region awaren
ess and reducing inter-region interference in self-attention. Extensive experime
nts demonstrate our method achieves state-of-the-art performance in real-world s
cenarios. Our code, data, and demos are available on the [project page](https://
knightyxp.github.io/VideoGrain_project_page/).
****
Qucheng Peng,Benjamin Planche,Zhongpai Gao,Meng Zheng,Anwesa Choudhuri,Terrence
Chen,Chen Chen,Ziyan Wu
3D Vision-Language Gaussian Splatting
Recent advancements in 3D reconstruction methods and vision-language models have

propelled the development of multi-modal 3D scene understanding, which has vital applications in robotics, autonomous driving, and virtual/augmented reality. However, current multi-modal scene understanding approaches have naively embedded semantic representations into 3D reconstruction methods without striking a balance between visual and language modalities, which leads to unsatisfying semantic rasterization of translucent or reflective objects, as well as over-fitting on color modality. To alleviate these limitations, we propose a solution that adequately handles the distinct visual and semantic modalities, i.e., a 3D vision-language Gaussian splatting model for scene understanding, to put emphasis on the representation learning of language modality. We propose a novel cross-modal rasterizer, using modality fusion along with a smoothed semantic indicator for enhancing semantic rasterization. We also employ a camera-view blending technique to improve semantic consistency between existing and synthesized views, thereby effectively mitigating over-fitting. Extensive experiments demonstrate that our method achieves state-of-the-art performance in open-vocabulary semantic segmentation, surpassing existing methods by a significant margin.
****

Ayesha Vermani,Josue Nassar,Hyungju Jeon,Matthew Dowling,Il Memming Park
Meta-Dynamical State Space Models for Integrative Neural Data Analysis
Learning shared structure across environments facilitates rapid learning and adaptive behavior in neural systems. This has been widely demonstrated and applied in machine learning to train models that are capable of generalizing to novel settings. However, there has been limited work exploiting the shared structure in neural activity during similar tasks for learning latent dynamics from neural recordings.
Existing approaches are designed to infer dynamics from a single dataset and cannot be readily adapted to account for statistical heterogeneities across recordings. In this work, we hypothesize that similar tasks admit a corresponding family of
related solutions and propose a novel approach for meta-learning this solution space from task-related neural activity of trained animals. Specifically, we capture the variabilities across recordings on a low-dimensional manifold which concisely parametrizes this family of dynamics, thereby facilitating rapid learning of latent dynamics given new recordings. We demonstrate the efficacy of our approach on
few-shot reconstruction and forecasting of synthetic dynamical systems, and neural recordings from the motor cortex during different arm reaching tasks.
****

Lingwei Zhu,Han Wang,Yukie Nagai
Fat-to-Thin Policy Optimization: Offline Reinforcement Learning with Sparse Policies
Sparse continuous policies are distributions that can choose some actions at random yet keep strictly zero probability for the other actions, which are radically different from the Gaussian.
They have important real-world implications, e.g. in modeling safety-critical tasks like medicine.
The combination of offline reinforcement learning and sparse policies provides a novel paradigm that enables learning completely from logged datasets a safety-aware sparse policy.
However, sparse policies can cause difficulty with the existing offline algorithms which require evaluating actions that fall outside of the current support.
In this paper, we propose the first offline policy optimization algorithm that tackles this challenge: Fat-to-Thin Policy Optimization (FtTPO).
Specifically, we maintain a fat (heavy-tailed) proposal policy that effectively learns from the dataset and injects knowledge to a thin (sparse) policy, which is responsible for interacting with the environment.
We instantiate FtTPO with the general q-Gaussian family that encompasses both heavy-tailed and sparse policies and verify that it performs favorably in a safety-critical treatment simulation and the standard MuJoCo suite.
Our code is available at https://github.com/lingweizhu/fat2thin.

****
Akinori F. Ebihara,Taiki Miyagawa,Kazuyuki Sakurai,Hitoshi Imaoka
Learning the Optimal Stopping for Early Classification within Finite Horizons via Sequential Probability Ratio Test
Time-sensitive machine learning benefits from Sequential Probability Ratio Test (SPRT), which provides an optimal stopping time for early classification of time series. However, in *finite horizon* scenarios, where input lengths are finite, determining the optimal stopping rule becomes computationally intensive due to the need for *backward induction*, limiting practical applicability. We thus introduce FIRMBOUND, an SPRT-based framework that efficiently estimates the solution to backward induction from training data, bridging the gap between optimal stopping theory and real-world deployment. It employs *density ratio estimation* and *convex function learning* to provide statistically consistent estimators for sufficient statistic and conditional expectation, both essential for solving backward induction; consequently, FIRMBOUND minimizes Bayes risk to reach optimality. Additionally, we present a faster alternative using Gaussian process regression, which significantly reduces training time while retaining low deployment overhead, albeit with potential compromise in statistical consistency. Experiments across independent and identically distributed (i.i.d.), non-i.i.d., binary, multiclass, synthetic, and real-world datasets show that FIRMBOUND achieves optimalities in the sense of Bayes risk and speed-accuracy tradeoff. Furthermore, it advances the tradeoff boundary toward optimality when possible and reduces decision-time variance, ensuring reliable decision-making. Code is included in the supplementary materials.
****
Shicong Cen,Jincheng Mei,Katayoon Goshvadi,Hanjun Dai,Tong Yang,Sherry Yang,Dale Schuurmans,Yuejie Chi,Bo Dai
Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF
Reinforcement learning from human feedback (RLHF) has demonstrated great promise in aligning large language models (LLMs) with human preference. Depending on the availability of preference data, both online and offline RLHF are active areas of investigation. A key bottleneck is understanding how to incorporate uncertainty estimation in the reward function learned from the preference data for RLHF, regardless of how the preference data is collected. While the principles of optimism or pessimism under uncertainty are well-established in standard reinforcement learning (RL), a practically-implementable and theoretically-grounded form amenable to large language models is not yet available, as standard techniques for constructing confidence intervals become intractable under arbitrary policy parameterizations.

In this paper, we introduce a unified approach to online and offline RLHF --- value-incentivized preference optimization (VPO) --- which regularizes the maximum-likelihood estimate of the reward function with the corresponding value function, modulated by a sign to indicate whether the optimism or pessimism is chosen. VPO also directly optimizes the policy with implicit reward modeling, and therefore shares a simpler RLHF pipeline similar to direct preference optimization. Theoretical guarantees of VPO are provided for both online and offline settings, matching the rates of their standard RL counterparts. Moreover, experiments on text summarization, dialogue, and standard benchmarks verify the practicality and effectiveness of VPO.
****
Sachin Goyal,Christina Baek,J Zico Kolter,Aditi Raghunathan
Context-Parametric Inversion: Why Instruction Finetuning May Not Actually Improve Context Reliance
Large Language Model's are instruction-finetuned to enhance their ability to follow user instructions and better comprehend input context. Still, they often struggle to follow the input context, especially when it contradicts model's parametric knowledge. This manifests as various failures, such as hallucinations where a model inserts outdated or unwarranted facts into its response. In this work,

we observe an intriguing phenomenon: the context reliance of the model decreases as instruction finetuning progresses, $\textit{despite an initial expected increase}$. We call this phenomenon as the $\textbf{context-parametric inversion}$. This is surprising, as one would expect instruction tuning to improve the model's ability to follow input instructions. We observe this behavior on multiple general purpose instruction tuning datasets such as TULU, Alpaca and Ultrachat, across multiple model families like Llama, Mistral and Pythia. We perform various controlled studies to eliminate some simple hypothesis for this observed behavior and isolate what datapoints cause this counter-intuitive behavior. We then analyze the phenomenon theoretically, to explain why context reliance varies across the trajectory of finetuning.
We tie the observed context-parametric inversion to the properties of the finetuning data, which provides us with some potential mitigation strategies that provide limited but insightful gains.
****
Gen Li,Yuchen Jiao
Improved Convergence Rate for Diffusion Probabilistic Models
Score-based diffusion models have achieved remarkable empirical performance in the field of machine learning and artificial intelligence for their ability to generate high-quality new data instances from complex distributions. Improving our understanding of diffusion models, including mainly convergence analysis for such models, has attracted a lot of interests. Despite a lot of theoretical attempts, there still exists significant gap between theory and practice. Towards to close this gap, we establish an iteration complexity at the order of $d^{1/3}\varepsilon^{-2/3}$, which is better than $d^{5/12}\varepsilon^{-1}$, the best known complexity achieved before our work. This convergence analysis is based on a randomized midpoint method, which is first proposed for log-concave sampling (Shen & Lee, 2019), and then extended to diffusion models by Gupta et al. (2024). Our theory accommodates ε-accurate score estimates, and does not require log-concavity on the target distribution. Moreover, the algorithm can also be parallelized to run in only $O(\log^2(d/\varepsilon))$ parallel rounds in a similar way to prior works.
****
Haohan Lin,Zhiqing Sun,Sean Welleck,Yiming Yang
Lean-STaR: Learning to Interleave Thinking and Proving
Traditional language model-based theorem proving assumes that by training on a sufficient amount of formal proof data, a model will learn to prove theorems. Our key observation is that a wealth of informal information that is not present in formal proofs can be useful for learning to prove theorems. For instance, humans think through steps of a proof, but this thought process is not visible in the resulting code. We present Lean-STaR, a framework for training language models to produce informal thoughts prior to each step of a proof, thereby boosting the model's theorem-proving capabilities. Lean-STaR uses retrospective ground-truth tactics to generate synthetic thoughts for training the language model. At inference time, the trained model directly generates the thoughts prior to the prediction of the tactics in each proof step. Building on the self-taught reasoner framework, we then apply expert iteration to further fine-tune the model on the correct proofs it samples and verifies using the Lean solver. Lean-STaR significantly outperform base models (43.4% $\rightarrow$ 46.3%, Pass@64). We also analyze the impact of the augmented thoughts on various aspects of the theorem proving process, providing insights into their effectiveness.
****
Dexuan Ding,Lei Wang,Liyun Zhu,Tom Gedeon,Piotr Koniusz
Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion
In computer vision tasks, features often come from diverse representations, domains (e.g., indoor and outdoor), and modalities (e.g., text, images, and videos). Effectively fusing these features is essential for robust performance, especially with the availability of powerful pre-trained models like vision-language models. However, common fusion methods, such as concatenation, element-wise operations, and non-linear techniques, often fail to capture structural relationships,

deep feature interactions, and suffer from inefficiency or misalignment of featu
res across domains or modalities. In this paper, we shift from high-dimensional
feature space to a lower-dimensional, interpretable graph space by constructing
relationship graphs that encode feature relationships at different levels, e.g.,
 clip, frame, patch, token, etc. To capture deeper interactions, we expand graph
s through iterative graph relationship updates and introduce a learnable graph f
usion operator to integrate these expanded relationships for more effective fusi
on. Our approach is relationship-centric, operates in a homogeneous space, and i
s mathematically principled, resembling element-wise relationship score aggregat
ion via multilinear polynomials. We demonstrate the effectiveness of our graph-b
ased fusion method on video anomaly detection, showing strong performance across
 multi-representational, multi-modal, and multi-domain feature fusion tasks.
****
Samuele Marro,Davide Evangelista,X. Angelo Huang,Emanuele La Malfa,Michele Lomba
rdi,Michael J. Wooldridge
Language Models Are Implicitly Continuous
Language is typically modelled with discrete sequences. However, the most succes
sful approaches to language modelling, namely neural networks, are continuous an
d smooth function approximators.
In this work, we show that Transformer-based language models implicitly learn to
 represent sentences as continuous-time functions defined over a continuous inpu
t space.
This phenomenon occurs in most state-of-the-art Large Language Models (LLMs), in
cluding Llama2, Llama3, Phi3, Gemma, Gemma2, and Mistral, and suggests that LLMs
 reason about language in ways that fundamentally differ from humans.
Our work formally extends Transformers to capture the nuances of time and space
continuity in both input and output space.
Our results challenge the traditional interpretation of how LLMs understand lang
uage, with several linguistic and engineering implications.
****
Rui Lu,Runzhe Wang,Kaifeng Lyu,Xitai Jiang,Gao Huang,Mengdi Wang
Towards Understanding Text Hallucination of Diffusion Models via Local Generatio
n Bias
Score-based diffusion models have achieved incredible performance in generating
realistic images, audio, and video data. While these models produce high-quality
 samples with impressive details, they often introduce unrealistic artifacts, su
ch as distorted fingers or hallucinated texts with no meaning. This paper focuse
s on textual hallucinations, where diffusion models correctly generate individua
l symbols but assemble them in a nonsensical manner. Through experimental probin
g, we consistently observe that such phenomenon is attributed it to the network'
s local generation bias. Denoising networks tend to produce outputs that rely he
avily on highly correlated local regions, particularly when different dimensions
 of the data distribution are nearly pairwise independent. This behavior leads t
o a generation process that decomposes the global distribution into separate, in
dependent distributions for each symbol, ultimately failing to capture the globa
l structure, including underlying grammar. Intriguingly, this bias persists acro
ss various denoising network architectures including MLP and transformers which
have the structure to model global dependency. These findings also provide insig
hts into understanding other types of hallucinations, extending beyond text, as
a result of implicit biases in the denoising models. Additionally, we theoretica
lly analyze the training dynamics for a specific case involving a two-layer MLP
learning parity points on a hypercube, offering an explanation of its underlying
 mechanism.
****
Siddharth Ramchandran,Manuel Haussmann,Harri Lähdesmäki
High-Dimensional Bayesian Optimisation with Gaussian Process Prior Variational A
utoencoders
Bayesian optimisation (BO) using a Gaussian process (GP)-based surrogate model i
s a powerful tool for solving black-box optimisation problems but does not scale
 well to high-dimensional data. Previous works have proposed to use variational

autoencoders (VAEs) to project high-dimensional data onto a low-dimensional late
nt space and to implement BO in the inferred latent space. In this work, we prop
ose a conditional generative model for efficient high-dimensional BO that uses a
 GP surrogate model together with GP prior VAEs. A GP prior VAE extends the stan
dard VAE by conditioning the generative and inference model on auxiliary covaria
tes, capturing complex correlations across samples with a GP. Our model incorpor
ates the observed target quantity values as auxiliary covariates learning a stru
ctured latent space that is better suited for the GP-based BO surrogate model. I
t handles partially observed auxiliary covariates using a unifying probabilistic
 framework and can also incorporate additional auxiliary covariates that may be
available in real-world applications. We demonstrate that our method improves up
on existing latent space BO methods on simulated datasets as well as on commonly
 used benchmarks.
****
Maximillian Chen,Ruoxi Sun,Tomas Pfister,Sercan O Arik
Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self
-Training
Large language models (LLMs), optimized through human feedback, have rapidly eme
rged as a leading paradigm for developing intelligent conversational assistants.
 However, despite their strong performance across many benchmarks, LLM-based age
nts might still lack conversational skills such as disambiguation -- when they a
re faced with ambiguity, they often overhedge or implicitly guess users' true in
tents rather than asking clarification questions. Under task-specific settings,
high-quality conversation samples are often limited, constituting a bottleneck f
or LLMs' ability to learn optimal dialogue action policies. We propose Action-Ba
sed Contrastive Self-Training (ACT), a quasi-online preference optimization algo
rithm based on Direct Preference Optimization (DPO), that enables data-efficient
 dialogue policy learning in multi-turn conversation modeling. We demonstrate AC
T's efficacy under data-efficient tuning scenarios, even when there is no action
 label available, using multiple real-world conversational tasks: tabular-ground
ed question-answering, machine reading comprehension, and AmbigSQL, a novel task
 for disambiguating information-seeking requests for complex SQL generation towa
rds data analysis agents. Additionally, we propose evaluating LLMs' ability to f
unction as conversational agents by examining whether they can implicitly recogn
ize and reason about ambiguity in conversation. ACT demonstrates substantial con
versation modeling improvements over standard tuning approaches like supervised
fine-tuning and DPO.
****
Chunting Zhou,LILI YU,Arun Babu,Kushal Tirumala,Michihiro Yasunaga,Leonid Shamis
,Jacob Kahn,Xuezhe Ma,Luke Zettlemoyer,Omer Levy
Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Mode
l
We introduce Transfusion, a recipe for training a multi-modal model over discret
e and continuous data.
Transfusion combines the language modeling loss function (next token prediction)
 with diffusion to train a single transformer over mixed-modality sequences.
We pretrain multiple Transfusion models up to 7B parameters from scratch on a mi
xture of text and image data, establishing scaling laws with respect to a variet
y of uni- and cross-modal benchmarks.
Our experiments show that Transfusion scales significantly better than quantizin
g images and training a language model over discrete image tokens.
By introducing modality-specific encoding and decoding layers, we can further im
prove the performance of Transfusion models, and even compress each image to jus
t 16 patches.
We further demonstrate that scaling our Transfusion recipe to 7B parameters and
2T multi-modal tokens produces a model that can generate images and text on a pa
r with similar scale diffusion models and language models, reaping the benefits
of both worlds.
****
Jinwoo Kim,Olga Zaghen,Ayhan Suleymanzade,Youngmin Ryou,Seunghoon Hong

Revisiting Random Walks for Learning on Graphs
We revisit a simple model class for machine learning on graphs, where a random walk on a graph produces a machine-readable record, and this record is processed by a deep neural network to directly make vertex-level or graph-level predictions. We call these stochastic machines random walk neural networks (RWNNs), and through principled analysis, show that we can design them to be isomorphism invariant while capable of universal approximation of graph functions in probability. A useful finding is that almost any kind of record of random walks guarantees probabilistic invariance as long as the vertices are anonymized. This enables us, for example, to record random walks in plain text and adopt a language model to read these text records to solve graph tasks. We further establish a parallelism to message passing neural networks using tools from Markov chain theory, and show that over-smoothing in message passing is alleviated by construction in RWNNs, while over-squashing manifests as probabilistic under-reaching. We empirically demonstrate RWNNs on a range of problems, verifying our theoretical analysis and demonstrating the use of language models for separating strongly regular graphs where 3-WL test fails, and transductive classification on arXiv citation network. Code is available at https://github.com/jw9730/random-walk.
****
Cyan Subhra Mishra,Deeksha Chaudhary,Jack Sampson,Mahmut Kandemir,Chita R. Das
NExUME: Adaptive Training and Inference for DNNs under Intermittent Power Environments
The deployment of Deep Neural Networks (DNNs) in energy-constrained environments, such as Energy Harvesting Wireless Sensor Networks (EH-WSNs), introduces significant challenges due to the intermittent nature of power availability. This study introduces NExUME, a novel training methodology designed specifically for DNNs operating under such constraints. We propose a dynamic adjustment of training parameters—dropout rates and quantization levels—that adapt in real-time to the available energy, which varies in energy harvesting scenarios.

This approach utilizes a model that integrates the characteristics of the network architecture and the specific energy harvesting profile. It dynamically adjusts training strategies, such as the intensity and timing of dropout and quantization, based on predictions of energy availability. This method not only conserves energy but also enhances the network’s adaptability, ensuring robust learning and inference capabilities even under stringent power constraints. Our results show a 6% to 22% improvement in accuracy over current methods, with an increase of less than 5% in computational overhead. This paper details the development of the adaptive training framework, describes the integration of energy profiles with dropout and quantization adjustments, and presents a comprehensive evaluation using real-world data. Additionally, we introduce a novel dataset aimed at furthering the application of energy harvesting in computational settings.
****
Wanchao Liang,Tianyu Liu,Less Wright,Will Constable,Andrew Gu,Chien-Chin Huang,Iris Zhang,Wei Feng,Howard Huang,Junjie Wang,Sanket Purandare,Gokul Nadathur,Stratos Idreos
TorchTitan: One-stop PyTorch native solution for production ready LLM pretraining
The development of large language models (LLMs) has been instrumental in advancing state-of-the-art natural language processing applications. Training LLMs with billions of parameters and trillions of tokens requires sophisticated distributed systems that enable composing and comparing several state-of-the-art techniques in order to efficiently scale across thousands of accelerators. However, existing solutions are complex, scattered across multiple libraries/repositories, lack interoperability, and are cumbersome to maintain. Thus, curating and empirically comparing training recipes requires non-trivial engineering effort.

This paper introduces **TORCHTITAN**^{1}, a PyTorch-native distributed training system that unifies and advances state-of-the-art techniques, streamlining integration and reducing engineering overhead. TORCHTITAN enables seamless applicatio

n of 4D parallelism in a modular and composable manner, while featuring elastic scaling to adapt to changing computational requirements. The system provides com prehensive logging, efficient checkpointing, and debugging tools, ensuring produ ction-ready training. Moreover, TORCHTITAN incorporates innovative hardware-soft ware co-designed solutions, leveraging cutting-edge features like Float8 trainin g and SymmetricMemory to maximize hardware utilization.

As a flexible experimental test bed, TORCHTITAN facilitates the curation and com parison of custom recipes for diverse training contexts. By leveraging TORCHTITA N, we developed optimized training recipes for the Llama 3.1 family and provide actionable guidance on selecting and combining distributed training techniques t o maximize training efficiency, based on our hands-on experiences.

We thoroughly assess TORCHTITAN on the Llama 3.1 family of LLMs, spanning 8 bill ion to 405 billion parameters, and showcase its exceptional performance, modular composability, and elastic scalability. By stacking training optimizations, we demonstrate accelerations ranging from 65.08% on Llama 3.1 8B at 128 GPU scale (1D), 12.59% on Llama 3.1 70B at 256 GPU scale (2D), to 30% on Llama 3.1 405B at 512 GPU scale (3D) on NVIDIA H100 GPUs over optimized baselines. We also demonst rate the effectiveness of 4D parallelism in enabling long context training.

^{1} GitHub: https://github.com/pytorch/torchtitan
****
Xinpeng Wang,Chengzhi Hu,Paul Röttger,Barbara Plank
Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via S ingle Vector Ablation
Training a language model to be both helpful and harmless requires careful calib ration of refusal behaviours: Models should refuse to follow malicious instructi ons or give harmful advice (e.g."how do I kill someone?"), but they should not r efuse safe requests, even if they superficially resemble unsafe ones (e.g. "how do I kill a Python process?"). Avoiding such false refusal, as prior work has sh own, is challenging even for highly-capable language models. In this paper, we p ropose a simple and surgical method for mitigating false refusal in language mod els via single vector ablation. For a given model, we extract a false refusal ve ctor and show that ablating this vector reduces false refusal rate while preserv ing the model's safety and general capabilities. We also show that our approach can be used for fine-grained calibration of model safety. Our approach is traini ng-free and model-agnostic, making it useful for mitigating the problem of false refusal in current and future language models.
****
Shengji Tang,Weicai Ye,Peng Ye,Weihao Lin,Yang Zhou,Tao Chen,Wanli Ouyang
HiSplat: Hierarchical 3D Gaussian Splatting for Generalizable Sparse-View Recons truction
Reconstructing 3D scenes from multiple viewpoints is a fundamental task in stere o vision. Recently, advances in generalizable 3D Gaussian Splatting have enabled high-quality novel view synthesis for unseen scenes from sparse input views by feed-forward predicting per-pixel Gaussian parameters without extra optimization . However, existing methods typically generate single-scale 3D Gaussians, which lack representation of both large-scale structure and texture details, resulting in mislocation and artefacts. In this paper, we propose a novel framework, HiSp lat, which introduces a hierarchical manner in generalizable 3D Gaussian Splatti ng to construct hierarchical 3D Gaussians via a coarse-to-fine strategy. Specifi cally, HiSplat generates large coarse-grained Gaussians to capture large-scale s tructures, followed by fine-grained Gaussians to enhance delicate texture detail s. To promote inter-scale interactions, we propose an Error Aware Module for Gau ssian compensation and a Modulating Fusion Module for Gaussian repair. Our metho d achieves joint optimization of hierarchical representations, allowing for nove l view synthesis using only two-view reference images. Comprehensive experiments on various datasets demonstrate that HiSplat significantly enhances reconstruct

ion quality and cross-dataset generalization compared to prior single-scale meth
ods. The corresponding ablation study and analysis of different-scale 3D Gaussia
ns reveal the mechanism behind the effectiveness. Code is at https://github.com/
Open3DVLab/HiSplat.
****
Pin Chen,Zexin Xu,Qing Mo,Hongjin Zhong,Fengyang Xu,Yutong Lu
ECD: A Machine Learning Benchmark for Predicting Enhanced-Precision Electronic C
harge Density in Crystalline Inorganic Materials
Supervised machine learning techniques are increasingly being adopted to speed u
p electronic structure predictions, serving as alternatives to first-principles
methods like Density Functional Theory (DFT). Although current DFT datasets main
ly emphasize chemical properties and atomic forces, the precise prediction of el
ectronic charge density is essential for accurately determining a system's total
 energy and ground state properties. In this study, we introduce a novel electro
nic charge density dataset named ECD, which encompasses 140,646 stable crystal g
eometries with medium-precision Perdew–Burke–Ernzerhof (PBE) functional data. Wi
thin this dataset, a subset of 7,147 geometries includes high-precision electron
ic charge density data calculated using the Heyd–Scuseria–Ernzerhof (HSE) functi
onal in DFT. By designing various benchmark tasks for crystalline materials and
emphasizing training with large-scale PBE data while fine-tuning with a smaller
subset of high-precision HSE data, we demonstrate the efficacy of current machin
e learning models in predicting electronic charge densities.
The ECD dataset and baseline models are open-sourced to support community effort
s in developing new methodologies and accelerating materials design and applicat
ions.
****
Shaofei Cai,Bowei Zhang,Zihao Wang,Haowei Lin,Xiaojian Ma,Anji Liu,Yitao Liang
GROOT-2: Weakly Supervised Multimodal Instruction Following Agents
Developing agents that can follow multimodal instructions remains a fundamental
challenge in robotics and AI. Although large-scale pre-training on unlabeled dat
asets has enabled agents to learn diverse behaviors, these agents often struggle
 with following instructions. While augmenting the dataset with instruction labe
ls can mitigate this issue, acquiring such high-quality annotations at scale is
impractical.
To address this issue, we frame the problem as a semi-supervised learning task a
nd introduce \agent, a multimodal instructable agent trained using a novel appro
ach that combines weak supervision with latent variable models. Our method consi
sts of two key components: constrained self-imitating, which utilizes large amou
nts of unlabeled demonstrations to enable the policy to learn diverse behaviors,
 and human intention alignment, which uses a smaller set of labeled demonstratio
ns to ensure the latent space reflects human intentions. \agent’s effectiveness
is validated across four diverse environments, ranging from video games to robot
ic manipulation, demonstrating its robust multimodal instruction-following capab
ilities.
****
Yanwen Huang,Bowen Gao,Yinjun Jia,Hongbo Ma,Wei-Ying Ma,Ya-Qin Zhang,Yanyan Lan
Redefining the task of Bioactivity Prediction
Small molecules are vital to modern medicine, and accurately predicting their bi
oactivity against protein targets is crucial for therapeutic discovery and devel
opment. However, current machine learning models often rely on spurious features
, leading to biased outcomes. Notably, a simple pocket-only baseline can achieve
 results comparable to, and sometimes better than, more complex models that inco
rporate both the protein pockets and the small molecules. Our analysis reveals t
hat this phenomenon arises from insufficient training data and an improper evalu
ation process, which is typically conducted at the pocket level rather than the
small molecule level. To address these issues, we redefine the bioactivity predi
ction task by introducing the SIU dataset-a million-scale Structural small molec
ule-protein Interaction dataset for Unbiased bioactivity prediction task, which
is 50 times larger than the widely used PDBbind. The bioactivity labels in SIU a
re derived from wet experiments and organized by label types, ensuring greater a

ccuracy and comparability. The complexes in SIU are constructed using a majority vote from three commonly used docking software programs, enhancing their reliability. Additionally, the structure of SIU allows for multiple small molecules to be associated with each protein pocket, enabling the redefinition of evaluation metrics like Pearson and Spearman correlations across different small molecules targeting the same protein pocket. Experimental results demonstrate that this new task provides a more challenging and meaningful benchmark for training and evaluating bioactivity prediction models, ultimately offering a more robust assessment of model performance.
****

Xu Han,Linghao Jin,Xiaofeng Liu,Paul Pu Liang
Progressive Compositionality in Text-to-Image Generative Models
Despite the impressive text-to-image (T2I) synthesis capabilities of diffusion models, they often struggle to understand compositional relationships between objects and attributes, especially in complex settings. Existing approaches through building compositional architectures or generating difficult negative captions often assume a fixed prespecified compositional structure, which limits generalization to new distributions. In this paper, we argue that curriculum training is crucial to equipping generative models with a fundamental understanding of compositionality. To achieve this, we leverage large-language models (LLMs) to automatically compose complex scenarios and harness Visual-Question Answering (VQA) checkers to automatically curate a contrastive dataset, ConPair, consisting of 15k pairs of high-quality contrastive images. These pairs feature minimal visual discrepancies and cover a wide range of attribute categories, especially complex and natural scenarios. To learn effectively from these error cases (i.e., hard negative images), we propose EvoGen, a new multi-stage curriculum for contrastive learning of diffusion models. Through extensive experiments across a wide range of compositional scenarios, we showcase the effectiveness of our proposed framework on compositional T2I benchmarks.
****

Hejia Chen,Haoxian Zhang,Shoulong Zhang,Xiaoqiang Liu,Sisi Zhuang,zhangyuan,Pengfei Wan,Di ZHANG,Shuai Li
Cafe-Talk: Generating 3D Talking Face Animation with Multimodal Coarse- and Fine-grained Control
Speech-driven 3D talking face method should offer both accurate lip synchronization and controllable expressions. Previous methods solely adopt discrete emotion labels to globally control expressions throughout sequences while limiting flexible fine-grained facial control within the spatiotemporal domain. We propose a diffusion-transformer-based 3D talking face generation model, Cafe-Talk, which simultaneously incorporates coarse- and fine-grained multimodal control conditions. Nevertheless, the entanglement of multiple conditions challenges achieving satisfying performance. To disentangle speech audio and fine-grained conditions, we employ a two-stage training pipeline. Specifically, Cafe-Talk is initially trained using only speech audio and coarse-grained conditions. Then, a proposed fine-grained control adapter gradually adds fine-grained instructions represented by action units (AUs), preventing unfavorable speech-lip synchronization. To disentangle coarse- and fine-grained conditions, we design a swap-label training mechanism, which enables the dominance of the fine-grained conditions. We also devise a mask-based CFG technique to regulate the occurrence and intensity of fine-grained control. In addition, a text-based detector is introduced with text-AU alignment to enable natural language user input and further support multimodal control. Extensive experimental results prove that Cafe-Talk achieves state-of-the-art lip synchronization and expressiveness performance and receives wide acceptance in fine-grained control in user studies.
****

Jiangkuo Wang,Suyv Ma,Mingpeng Wei
Enhancing Multilingual Reasoning in LLMs: Insights from Cross-Linguistic Correlations and Optimal Data Proportions
Large language models (LLMs) typically rely on fine-tuning to enhance their reasoning capabilities across various languages. However, limited research has been

conducted on the optimal balance of language proportions within multilingual rea
soning datasets. To fill this gap, we performed a systematic study to examine ho
w different proportions of language data in multilingual reasoning datasets infl
uence fine-tuning performance. Our study revealed a clear relationship between l
anguage proportions in datasets and the fine-tuning performance of LLMs. By fine
-tuning multiple LLMs using the appropriate language distributions and data volu
mes identified in our study, we achieved state-of-the-art performance in both mu
ltilingual mathematical reasoning and solving mathematical problems using Python
 code. Furthermore, our approach significantly reduced data volume requirements
and translation costs compared to existing methods, providing a valuable referen
ce for future research.
****
Yinuo Wang,Wenxuan Wang,Xujie Song,Tong Liu,Yuming Yin,Liangfa Chen,Likun Wang,J
ingliang Duan,Shengbo Eben Li
ODE-based Smoothing Neural Network for Reinforcement Learning Tasks
The smoothness of control actions is a significant challenge faced by deep reinf
orcement learning (RL) techniques in solving optimal control problems. Existing
RL-trained policies tend to produce non-smooth actions due to high-frequency inp
ut noise and unconstrained Lipschitz constants in neural networks. This article
presents a Smooth ODE (SmODE) network capable of simultaneously addressing both
causes of unsmooth control actions, thereby enhancing policy performance and rob
ustness under noise condition. We first design a smooth ODE neuron with first-or
der low-pass filtering expression, which can dynamically filter out high frequen
cy noises of hidden state by a learnable state-based system time constant. Addit
ionally, we construct a state-based mapping function, g, and theoretically dem
onstrate its capacity to control the ODE neuron's Lipschitz constant. Then, base
d on the above neuronal structure design, we further advanced the SmODE network
serving as RL policy approximators. This network is compatible with most existin
g RL algorithms, offering improved adaptability compared to prior approaches. Va
rious experiments show that our SmODE network demonstrates superior anti-interfe
rence capabilities and smoother action outputs than the multi-layer perception a
nd smooth network architectures like LipsNet.
****
Yining Li,Peizhong Ju,Ness Shroff
How to Find the Exact Pareto Front for Multi-Objective MDPs?
Multi-Objective Markov Decision Processes (MO-MDPs) are receiving increasing att
ention, as real-world decision-making problems often involve conflicting objecti
ves that cannot be addressed by a single-objective MDP.
The Pareto front identifies the set of policies that cannot be dominated, provid
ing a foundation for finding Pareto optimal solutions that can efficiently adapt
 to various preferences.
However, finding the Pareto front is a highly challenging problem. Most existing
 methods either (i) rely on traversing the *continuous preference space*, which
is impractical and results in approximations that are difficult to evaluate agai
nst the true Pareto front, or (ii) focus solely on deterministic Pareto optimal
policies, from which there are no known techniques to characterize the full Pare
to front. Moreover, finding the structure of the Pareto front itself remains unc
lear even in the context of dynamic programming, where the MDP is fully known in
 advance.
In this work, we address the challenge of efficiently discovering the Pareto fro
nt, involving both deterministic and stochastic Pareto optimal policies.
By investigating the geometric structure of the Pareto front in MO-MDPs, we unco
ver a key property: the Pareto front is on the boundary of a convex polytope who
se vertices all correspond to deterministic policies, and neighboring vertices o
f the Pareto front differ by only one state-action pair of the deterministic pol
icy, almost surely.
This insight transforms the global comparison across all policies into a localiz
ed search among deterministic policies that differ by only one state-action pair
, drastically reducing the complexity of searching for the exact Pareto front.
We develop an efficient algorithm that identifies the vertices of the Pareto fro

nt by solving a single-objective MDP only once and then traversing the edges of the Pareto front, making it more efficient than existing methods. Furthermore, t he entire Pareto front can be found in V iterations, where V represents the number of vertices on the Pareto front.
Our empirical studies demonstrate the effectiveness of our theoretical strategy in discovering the Pareto front efficiently.
****
Ruochen Jiao,Shaoyuan Xie,Justin Yue,TAKAMI SATO,Lixu Wang,Yixuan Wang,Qi Alfred Chen,Qi Zhu
Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Ba sed Decision-Making Systems
Large Language Models (LLMs) have shown significant promise in real-world decisi on-making tasks for embodied artificial intelligence, especially when fine-tuned to leverage their inherent common sense and reasoning abilities while being tai lored to specific applications. However, this fine-tuning process introduces con siderable safety and security vulnerabilities, especially in safety-critical cyb er-physical systems. In this work, we propose the first comprehensive framework for **B**ackdoor **A**ttacks against **L**LM-based **D**ecision-making systems (BALD) in embodied AI, systematically exploring the attack surfaces and trigger m echanisms. Specifically, we propose three distinct attack mechanisms: *word inje ction*, *scenario manipulation*, and *knowledge injection*, targeting various co mponents in the LLM-based decision-making pipeline. We perform extensive experim ents on representative LLMs (GPT-3.5, LLaMA2, PaLM2) in autonomous driving and h ome robot tasks, demonstrating the effectiveness and stealthiness of our backdoo r triggers across various attack channels, with cases like vehicles accelerating toward obstacles and robots placing knives on beds. Our word and knowledge inje ction attacks achieve nearly 100\% success rate across multiple models and datas ets while requiring only limited access to the system. Our scenario manipulation attack yields success rates exceeding 65\%, reaching up to 90\%, and does not r equire any runtime system intrusion. We also assess the robustness of these atta cks against defenses, revealing their resilience. Our findings highlight critica l security vulnerabilities in embodied LLM systems and emphasize the urgent need for safeguarding these systems to mitigate potential risks.
****
Keltin Grimes,Marco Christiani,David Shriver,Marissa Catherine Connor
Concept-ROT: Poisoning Concepts in Large Language Models with Model Editing
Model editing methods modify specific behaviors of Large Language Models by alte ring a small, targeted set of network weights and require very little data and c ompute. These methods can be used for malicious applications such as inserting m isinformation or simple trojans that result in adversary-specified behaviors whe n a trigger word is present. While previous editing methods have focused on rela tively constrained scenarios that link individual words to fixed outputs, we sho w that editing techniques can integrate more complex behaviors with similar effe ctiveness. We develop Concept-ROT, a model editing-based method that efficiently inserts trojans which not only exhibit complex output behaviors, but also trigg er on high-level concepts -- presenting an entirely new class of trojan attacks. Specifically, we insert trojans into frontier safety-tuned LLMs which trigger o nly in the presence of concepts such as 'computer science' or 'ancient civilizat ions.' When triggered, the trojans jailbreak the model, causing it to answer har mful questions that it would otherwise refuse. Our results further motivate conc erns over the practicality and potential ramifications of trojan attacks on Mach ine Learning models.
****
Yudong Chen,Xuwei Xu,Frank de Hoog,Jiajun Liu,Sen Wang
Medium-Difficulty Samples Constitute Smoothed Decision Boundary for Knowledge Di stillation on Pruned Datasets
This paper tackles a new problem of dataset pruning for Knowledge Distillation (KD), from a fresh perspective of Decision Boundary (DB) preservation and drifts. Existing dataset pruning methods generally assume that the post-pruning DB form ed by the selected samples can be well-captured by future networks that use thos

e samples for training. Therefore, they tend to preserve hard samples since hard samples are closer to the DB and better characterize the nuances in the distribution of the entire dataset. However, in KD, the limited learning capacity from the student network leads to imperfect preservation of the teacher's feature distribution, resulting in the drift of DB in the student space. Specifically, hard samples worsen such drifts as they are difficult for the student to learn, creating a situation where the student's DB can drift deeper into other classes and make incorrect classifications. Motivated by these findings, our method selects medium-difficulty samples for KD-based dataset pruning. We show that these samples constitute a smoothed version of the teacher's DB and are easier for the student to learn, obtaining a general feature distribution preservation for a class of samples and reasonable DB between different classes for the student. In addition, to reduce the distributional shift due to dataset pruning, we leverage the class-wise distributional information of the teacher's outputs to reshape the logits of the preserved samples. Experiments show that the proposed static pruning method can even perform better than the state-of-the-art dynamic pruning method which needs access to the entire dataset. In addition, our method halves the training times of KD and improves the student's accuracy by 0.4% on ImageNet with a 50% keep ratio. When the ratio further increases to 70%, our method achieves higher accuracy over the vanilla KD while reducing the training times by 30%. Code is available at https://github.com/chenyd7/MDSLR.
****
Matthieu Zimmer,Milan Gritta,Gerasimos Lampouras,Haitham Bou Ammar,Jun Wang
Mixture of Attentions For Speculative Decoding
The growth in the number of parameters of Large Language Models (LLMs) has led to a significant surge in computational requirements, making them challenging and costly to deploy.
Speculative decoding (SD) leverages smaller models to efficiently propose future tokens, which are then verified by the LLM in parallel.
Small models that utilise activations from the LLM currently achieve the fastest decoding speeds.
However, we identify several limitations of SD models including the lack of on-policyness during training and partial observability.
To address these shortcomings, we propose a more grounded architecture for small models by introducing a Mixture of Attentions for SD.
Our novel architecture can be applied in two scenarios: a conventional single device deployment and a novel client-server deployment where the small model is hosted on a consumer device and the LLM on a server.
In a single-device scenario, we demonstrate state-of-the-art speedups improving EAGLE-2 by 9.5% and its acceptance length by 25%.
In a client-server setting, our experiments demonstrate: 1) state-of-the-art latencies with minimal calls to the server for different network conditions, and 2) in the event of a complete disconnection, our approach can maintain higher accuracy compared to other SD methods and demonstrates advantages over API calls to LLMs, which would otherwise be unable to continue the generation process.
****
Bowen Gao,Haichuan Tan,Yanwen Huang,Minsi Ren,Xiao Huang,Wei-Ying Ma,Ya-Qin Zhang,Yanyan Lan
Reframing Structure-Based Drug Design Model Evaluation via Metrics Correlated to Practical Needs
Recent advances in structure-based drug design (SBDD) have produced surprising results, with models often generating molecules that achieve better Vina docking scores than actual ligands. However, these results are frequently overly optimistic due to the limitations of docking score accuracy and the challenges of wet-lab validation. While generated molecules may demonstrate high QED (drug-likeness) and SA (synthetic accessibility) scores, they often lack true drug-like properties or synthesizability. To address these limitations, we propose a model-level evaluation framework that emphasizes practical metrics aligned with real-world applications. Inspired by recent findings on the utility of generated molecules in ligand-based virtual screening, our framework evaluates SBDD models by their

ability to produce molecules that effectively retrieve active compounds from che
mical libraries via similarity-based searches. This approach provides a direct i
ndication of therapeutic potential, bridging the gap between theoretical perform
ance and real-world utility. Our experiments reveal that while SBDD models may e
xcel in theoretical metrics like Vina scores, they often fall short in these pra
ctical metrics. By introducing this new evaluation strategy, we aim to enhance t
he relevance and impact of SBDD models for pharmaceutical research and developme
nt.
****
Peter Holderrieth,Marton Havasi,Jason Yim,Neta Shaul,Itai Gat,Tommi Jaakkola,Bri
an Karrer,Ricky T. Q. Chen,Yaron Lipman
Generator Matching: Generative modeling with arbitrary Markov processes
We introduce Generator Matching, a modality-agnostic framework for generative mo
deling using arbitrary Markov processes. Generators characterize the infinitesim
al evolution of a Markov process, which we leverage for generative modeling in a
 similar vein to flow matching: we construct conditional generators which genera
te single data points, then learn to approximate the marginal generator which ge
nerates the full data distribution. We show that Generator Matching unifies vari
ous generative modeling methods, including diffusion models, flow matching and d
iscrete diffusion models. Furthermore, it expands the design space to new and un
explored Markov processes such as jump processes. Finally, Generator Matching en
ables the construction of superpositions of Markov generative models and enables
 the construction of multimodal models in a rigorous manner. We empirically vali
date our method on image and multimodal generation, e.g. showing that superposit
ion with a jump process improves performance.
****
Haiyu Wu,Jaskirat Singh,Sicong Tian,Liang Zheng,Kevin Bowyer
Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors
This paper studies how to synthesize face images of non-existent persons, to cre
ate a dataset that allows effective training of face recognition (FR) models. Be
sides generating realistic face images, two other important goals are: 1) the ab
ility to generate a large number of distinct identities (inter-class separation)
, and 2) a proper variation in appearance of the images for each identity (intra
-class variation).
However, existing works 1) are typically limited in how many well-separated iden
tities can be generated and 2) either neglect or use an external model for attri
bute augmentation. We propose Vec2Face, a holistic model that uses only a sample
d vector as input and can flexibly generate and control the identity of face ima
ges and their attributes. Composed of a feature masked autoencoder and an image
decoder, Vec2Face is supervised by face image reconstruction and can be convenie
ntly used in inference. Using vectors with low similarity among themselves as in
puts, Vec2Face generates well-separated identities. Randomly perturbing an input
 identity vector within a small range allows Vec2Face to generate faces of the s
ame identity with proper variation in face attributes. It is also possible to ge
nerate images with designated attributes by adjusting vector values with a gradi
ent descent method. Vec2Face has efficiently synthesized as many as 300K identit
ies, whereas 60K is the largest number of identities created in the previous wor
ks. As for performance, FR models trained with the generated HSFace datasets, fr
om 10k to 300k identities, achieve state-of-the-art accuracy, from 92\% to 93.52
\%, on five real-world test sets (\emph{i.e.}, LFW, CFP-FP, AgeDB-30, CALFW, and
 CPLFW). For the first time, the FR model trained using our synthetic training s
et achieves higher accuracy than that trained using a same-scale training set of
 real face images on the CALFW, IJBB, and IJBC test sets.
****
Katharina Friedl,Noémie Jaquier,Jens Lundell,Tamim Asfour,Danica Kragic
A Riemannian Framework for Learning Reduced-order Lagrangian Dynamics
By incorporating physical consistency as inductive bias, deep neural networks di
splay increased generalization capabilities and data efficiency in learning nonl
inear dynamic models. However, the complexity of these models generally increase
s with the system dimensionality, requiring larger datasets, more complex deep n

etworks, and significant computational effort.
We propose a novel geometric network architecture to learn physically-consistent reduced-order dynamic parameters that accurately describe the original high-dimensional system behavior.
This is achieved by building on recent advances in model-order reduction and by adopting a Riemannian perspective to jointly learn a non-linear structure-preserving latent space and the associated low-dimensional dynamics.
Our approach enables accurate long-term predictions of the high-dimensional dynamics of rigid and deformable systems with increased data efficiency by inferring interpretable and physically-plausible reduced Lagrangian models.
****
Shengzhuang Chen,Yikai Liao,Xiaoxiao Sun,Kede Ma,Ying Wei
CLDyB: Towards Dynamic Benchmarking for Continual Learning with Pre-trained Models
The emergence of the foundation model era has sparked immense research interest in utilizing pre-trained representations for continual learning~(CL), yielding a series of strong CL methods with outstanding performance on standard evaluation benchmarks. Nonetheless, there are growing concerns regarding potential data contamination within the massive pre-training datasets. Furthermore, the static nature of standard evaluation benchmarks tends to oversimplify the complexities encountered in real-world CL scenarios, putting CL methods at risk of overfitting to these benchmarks while still lacking robustness needed for more demanding real-world applications. To solve these problems, this paper proposes a general framework to evaluate methods for Continual Learning on Dynamic Benchmarks (CLDyB). CLDyB continuously identifies inherently challenging tasks for the specified CL methods and evolving backbones, and dynamically determines the sequential order of tasks at each time step in CL using a tree-search algorithm, guided by an overarching goal to generate highly challenging task sequences for evaluation. To highlight the significance of dynamic evaluation on the CLDyB, we first simultaneously evaluate multiple state-of-the-art CL methods under CLDyB, resulting in a set of commonly challenging task sequences where existing CL methods tend to underperform. We intend to publicly release these task sequences for the CL community to facilitate the training and evaluation of more robust CL algorithms. Additionally, we perform individual evaluations of the CL methods under CLDyB, yielding informative evaluation results that reveal the specific strengths and weaknesses of each method.
****
Harsh Gupta,Yuchen Mo,Shengmiao Jin,Wenzhen Yuan
Sensor-Invariant Tactile Representation
High-resolution tactile sensors have become critical for embodied perception and robotic manipulation.
However, a key challenge in the field is the lack of transferability between sensors due to design and manufacturing variations, which result in significant differences in tactile signals.
This limitation hinders the ability to transfer models or knowledge learned from one sensor to another.
To address this, we introduce a novel method for extracting Sensor-Invariant Tactile Representations (SITR), enabling zero-shot transfer across optical tactile sensors.
Our approach utilizes a transformer-based architecture trained on a diverse dataset of simulated sensor designs, allowing it to generalize to new sensors in the real world with minimal calibration.
Experimental results demonstrate the method’s effectiveness across various tactile sensing applications, facilitating data and model transferability for future advancements in the field.
****
Rafal Karczewski,Markus Heinonen,Vikas Garg
Diffusion Models as Cartoonists: The Curious Case of High Density Regions
We investigate what kind of images lie in the high-density regions of diffusion models. We introduce a theoretical mode-tracking process capable of pinpointing

the exact mode of the denoising distribution, and we propose a practical high-de
nsity sampler that consistently generates images of higher likelihood than usual
 samplers. Our empirical findings reveal the existence of significantly higher l
ikelihood samples that typical samplers do not produce, often manifesting as car
toon-like drawings or blurry images depending on the noise level. Curiously, the
se patterns emerge in datasets devoid of such examples. We also present a novel
approach to track sample likelihoods in diffusion SDEs, which remarkably incurs
no additional computational cost. Code is available at https://github.com/Aalto-
QuML/high-density-diffusion
****

Xiuyuan Hu,Guoqing Liu,Can Chen,Yang Zhao,Hao Zhang,Xue Liu
3DMolFormer: A Dual-channel Framework for Structure-based Drug Discovery
Structure-based drug discovery, encompassing the tasks of protein-ligand docking
 and pocket-aware 3D drug design, represents a core challenge in drug discovery.
 However, no existing work can deal with both tasks to effectively leverage the
duality between them, and current methods for each task are hindered by challeng
es in modeling 3D information and the limitations of available data. To address
these issues, we propose 3DMolFormer, a unified dual-channel transformer-based f
ramework applicable to both docking and 3D drug design tasks, which exploits the
ir duality by utilizing docking functionalities within the drug design process.
Specifically, we represent 3D pocket-ligand complexes using parallel sequences o
f discrete tokens and continuous numbers, and we design a corresponding dual-cha
nnel transformer model to handle this format, thereby overcoming the challenges
of 3D information modeling. Additionally, we alleviate data limitations through
large-scale pre-training on a mixed dataset, followed by supervised and reinforc
ement learning fine-tuning techniques respectively tailored for the two tasks. E
xperimental results demonstrate that 3DMolFormer outperforms previous approaches
 in both protein-ligand docking and pocket-aware 3D drug design, highlighting it
s promising application in structure-based drug discovery. The code is available
 at: https://github.com/HXYfighter/3DMolFormer.
****

Juno Kim,Dimitri Meunier,Arthur Gretton,Taiji Suzuki,Zhu Li
Optimality and Adaptivity of Deep Neural Features for Instrumental Variable Regr
ession
We provide a convergence analysis of \emph{deep feature instrumental variable} (
DFIV) regression (Xu et al., 2021), a nonparametric approach to IV regression us
ing data-adaptive features learned by deep neural networks in two stages. We pro
ve that the DFIV algorithm achieves the minimax optimal learning rate when the t
arget structural function lies in a Besov space. This is shown under standard no
nparametric IV assumptions, and an additional smoothness assumption on the regul
arity of the conditional distribution of the covariate given the instrument, whi
ch controls the difficulty of Stage 1. We further demonstrate that DFIV, as a da
ta-adaptive algorithm, is superior to fixed-feature (kernel or sieve) IV methods
 in two ways. First, when the target function possesses low spatial homogeneity
(i.e., it has both smooth and spiky/discontinuous regions), DFIV still achieves
the optimal rate, while fixed-feature methods are shown to be strictly suboptima
l. Second, comparing with kernel-based two-stage regression estimators, DFIV is
provably more data efficient in the Stage 1 samples.
****

Siyu Ren,Junhui Hou
Shape as Line Segments: Accurate and Flexible Implicit Surface Representation
Distance field-based implicit representations like signed/unsigned distance fiel
ds have recently gained prominence in geometry modeling and analysis. However, t
hese distance fields are reliant on the closest distance of points to the surfac
e, introducing inaccuracies when interpolating along cube edges during surface e
xtraction. Additionally, their gradients are ill-defined at certain locations, c
ausing distortions in the extracted surfaces. To address this limitation, we pro
pose Shape as Line Segments (SALS), an accurate and efficient implicit geometry
representation based on attributed line segments, which can handle arbitrary str
uctures. Unlike previous approaches, SALS leverages a differentiable Line Segmen

t Field to implicitly capture the spatial relationship between line segments and the surface. Each line segment is associated with two key attributes, intersection flag and ratio, from which we propose edge-based dual contouring to extract a surface. We further implement SALS with a neural network, producing a new neural implicit presentation. Additionally, based on SALS, we design a novel learning-based pipeline for reconstructing surfaces from 3D point clouds. We conduct extensive experiments, showcasing the significant advantages of our methods over state-of-the-art methods.
The source code is available at https://github.com/rsy6318/SALS.
****
Xu Shifeng,Yanzhu Liu,Adams Wai-Kin Kong
Easing Training Process of Rectified Flow Models Via Lengthening Inter-Path Distance
Recent research pinpoints that different diffusion methods and architectures trained on the same dataset produce similar results for the same input noise. This property suggests that they have some preferable noises for a given sample.

By visualizing the noise-sample pairs of rectified flow models and stable diffusion models in two-dimensional spaces,
we observe that the preferable paths, connecting preferable noises to the corresponding samples,
are better organized with significant fewer crossings comparing with
the random paths, connecting random noises to training samples.
In high-dimensional space, paths rarely intersect.
The path crossings in two-dimensional spaces indicate the shorter inter-path distance
in the corresponding high-dimensional spaces.
Inspired by this observation, we propose the Distance-Aware Noise-Sample Matching (DANSM) method
to lengthen the inter-path distance for speeding up the model training.
DANSM is derived from rectified flow models, which allow using a closed-form formula to calculate the inter-path distance.
To further simplify the optimization, we derive the relationship between inter-path distance and path length,
and use the latter in the optimization surrogate.
DANSM is evaluated on both image and latent spaces by rectified flow models and diffusion models.
The experimental results show that DANSM can significantly improve the training speed by 30\% $\sim$ 40\%
without sacrificing the generation quality.
****
Annan Yu,Michael W. Mahoney,N. Benjamin Erichson
HOPE for a Robust Parameterization of Long-memory State Space Models
State-space models (SSMs) that utilize linear, time-invariant (LTI) systems are known for their effectiveness in learning long sequences. To achieve state-of-the-art performance, an SSM often needs a specifically designed initialization, and the training of state matrices is on a logarithmic scale with a very small learning rate. To understand these choices from a unified perspective, we view SSMs through the lens of Hankel operator theory. Building upon it, we develop a new parameterization scheme, called HOPE, for LTI systems that utilizes Markov parameters within Hankel operators. Our approach helps improve the initialization and training stability, leading to a more robust parameterization. We efficiently implement these innovations by nonuniformly sampling the transfer functions of LTI systems, and they require fewer parameters compared to canonical SSMs. When benchmarked against HiPPO-initialized models such as S4 and S4D, an SSM parameterized by Hankel operators demonstrates improved performance on Long-Range Arena (LRA) tasks. Moreover, our new parameterization endows the SSM with non-decaying memory within a fixed time window, which is empirically corroborated by a sequential CIFAR-10 task with padded noise.
****

Chongjie Si,Zhiyi Shi,Shifan Zhang,Xiaokang Yang,Hanspeter Pfister,Wei Shen
Unleashing the Power of Task-Specific Directions in Parameter Efficient Fine-tuning
Large language models demonstrate impressive performance on downstream tasks, yet requiring extensive resource consumption when fully fine-tuning all parameters. To mitigate this, Parameter Efficient Fine-Tuning (PEFT) strategies, such as LoRA, have been developed.
In this paper, we delve into the concept of task-specific directions (TSDs)—critical for transitioning large models from pretrained states to task-specific enhancements in PEFT. We propose a framework to clearly define these directions and explore their properties, and practical utilization challenges. We then introduce a novel approach, LoRA-Dash, which aims to maximize the impact of TSDs during the fine-tuning process, thereby enhancing model performance on targeted tasks. Extensive experiments have conclusively demonstrated the effectiveness of LoRA-Dash, and in-depth analyses further reveal the underlying mechanisms of LoRA-Dash.
****
Chen Jiang,Jiahui An,Yating Liu,Ni Ji
Brain Bandit: A Biologically Grounded Neural Network for Efficient Control of Exploration
How to balance between exploration and exploitation in an uncertain environment is a central challenge in reinforcement learning. In contrast, humans and animals have demonstrated superior exploration efficiency in novel environments. To understand how the brain’s neural network controls exploration under uncertainty, we analyzed the dynamical systems model of a biological neural network that controls explore-exploit decisions during foraging. Mathematically, this model (named the Brain Bandit Net, or BBN) is a special type of stochastic continuous Hopfield network. We show through theory and simulation that BBN can perform posterior sampling of action values with a tunable bias towards or against uncertain options. We then demonstrate that, in multi-armed bandit (MAB) tasks, BBN can generate probabilistic choice behavior with a flexible uncertainty bias resembling human and animal choice patterns. In addition to its high efficiency in MAB tasks, BBN can also be embedded with reinforcement learning algorithms to accelerate learning in MDP tasks. Altogether, our findings reveal the theoretical foundation for efficient exploration in biological neural networks and propose a general, brain-inspired algorithm for enhancing exploration in RL.
****
Yang JingYi,Zitong YU,Nixiuming,He Jia,Hui Li
Kronecker Mask and Interpretive Prompts are Language-Action Video Learners
Contrastive language-image pretraining (CLIP) has significantly advanced image-based vision learning. A pressing topic subsequently arises: how can we effectively adapt CLIP to the video domain? Recent studies have focused on adjusting either the textual or visual branch of CLIP for action recognition. However, we argue that adaptations of both branches are crucial. In this paper, we propose a **C**ontrastive **L**anguage-**A**ction **V**ideo Learn**er** (**CLAVER**), designed to shift CLIP's focus from the alignment of static visual objects and concrete nouns to the alignment of dynamic action behaviors and abstract verbs. Specifically, we introduce a novel Kronecker mask attention for temporal modeling. Our tailored Kronecker mask offers three benefits 1) it expands the temporal receptive field for each token, 2) it serves as an effective spatiotemporal heterogeneity inductive bias, mitigating the issue of spatiotemporal homogenization, and 3) it can be seamlessly plugged into transformer-based models. Regarding the textual branch, we leverage large language models to generate diverse, sentence-level and semantically rich interpretive prompts of actions, which shift the model's focus towards the verb comprehension. Extensive experiments on various benchmarks and learning scenarios demonstrate the superiority and generality of our approach. The code will be available soon.
****
Ruoxi Xu,Hongyu Lin,Xianpei Han,Jia Zheng,Weixiang Zhou,Le Sun,Yingfei Sun
Large Language Models Often Say One Thing and Do Another

As large language models (LLMs) increasingly become central to various applicati
ons and interact with diverse user populations, ensuring their reliable and cons
istent performance is becoming more important. This paper explores a critical is
sue in assessing the reliability of LLMs: the consistency between their words an
d deeds. To quantitatively explore this consistency, we developed a novel evalua
tion benchmark called the Words and Deeds Consistency Test (WDCT). The benchmark
 establishes a strict correspondence between word-based and deed-based questions
 across different domains, including opinion vs. action, non-ethical value vs. a
ction, ethical value vs. action, and theory vs. application. The evaluation resu
lts reveal a widespread inconsistency between words and deeds across different L
LMs and domains. Subsequently, we conducted experiments with either word alignme
nt or deed alignment to observe their impact on the other aspect. The experiment
al results indicate that alignment only on words or deeds poorly and unpredictab
ly influences the other aspect. This supports our hypothesis that the underlying
 knowledge guiding LLMs' word or deed choices is not contained within a unified
space. Dataset and code are available at [https://github.com/icip-cas/Word-Deed-
Consistency-Test](https://github.com/icip-cas/Word-Deed-Consistency-Test).
****

Zhiting Fan,Ruizhe Chen,Tianxiang Hu,Zuozhu Liu
FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LL
Ms
The increasing deployment of large language model (LLM)-based chatbots has raise
d concerns regarding fairness. Fairness issues in LLMs may result in serious con
sequences, such as bias amplification, discrimination, and harm to minority grou
ps. Many efforts are dedicated to evaluating and mitigating biases in LLMs. Howe
ver, existing fairness benchmarks mainly focus on single-turn dialogues, while m
ulti-turn scenarios, which better reflect real-world conversations, pose greater
 challenges due to conversational complexity and risk for bias accumulation. In
this paper, we introduce a comprehensive benchmark for fairness of LLMs in multi
-turn scenarios, **FairMT-Bench**. Specifically, We propose a task taxonomy to e
valuate fairness of LLMs cross three stages: context understanding, interaction
fairness, and fairness trade-offs, each comprising two tasks. To ensure coverage
 of diverse bias types and attributes, our multi-turn dialogue dataset FairMT-10
K is constructed by integrating data from established fairness benchmarks. For e
valuation, we employ GPT-4 along with bias classifiers like Llama-Guard-3, and h
uman annotators to ensure robustness. Our experiments and analysis on FairMT-10K
 reveal that in multi-turn dialogue scenarios, LLMs are more prone to generating
 biased responses, showing significant variation in performance across different
 tasks and models. Based on these findings, we develop a more challenging datase
t, FairMT-1K, and test 15 current state-of-the-art (SOTA) LLMs on this dataset.
The results highlight the current state of fairness in LLMs and demonstrate the
value of this benchmark for evaluating fairness of LLMs in more realistic multi-
turn dialogue contexts. This underscores the need for future works to enhance LL
M fairness and incorporate FairMT-1K in such efforts. Our code and dataset are a
vailable at https://github.com/FanZT6/FairMT-bench.
****

Devon Jarvis,Sebastian Lee,Clémentine Carla Juliette Dominé,Andrew M Saxe,Stefan
o Sarao Mannelli
A Theory of Initialisation's Impact on Specialisation
Prior work has demonstrated a consistent tendency in neural networks engaged in
continual learning tasks, wherein intermediate task similarity results in the hi
ghest levels of catastrophic interference. This phenomenon is attributed to the
network's tendency to reuse learned features across tasks. However, this explana
tion heavily relies on the premise that neuron specialisation occurs, i.e. the e
mergence of localised representations. Our investigation challenges the validit
y of this assumption. Using theoretical frameworks for the analysis of neural ne
tworks, we show a strong dependence of specialisation on the initial condition.
More precisely, we show that weight imbalance and high weight entropy can favour
 specialised solutions. We then apply these insights in the context of continual
 learning, first showing the emergence of a monotonic relation between task-simi

larity and forgetting in non-specialised networks. Finally, we show that specialization by weight imbalance is beneficial on the commonly employed elastic weight consolidation regularisation technique.
****
Wei Li,Luyao Zhu,Yang Song,Ruixi Lin,Rui Mao,Yang You
Can a Large Language Model be a Gaslighter?
Large language models (LLMs) have gained human trust due to their capabilities and helpfulness. However, this in turn may allow LLMs to affect users' mindsets by manipulating language. It is termed as gaslighting, a psychological effect.
In this work, we aim to investigate the vulnerability of LLMs under prompt-based and fine-tuning-based gaslighting attacks. Therefore, we propose a two-stage framework DeepCoG designed to: 1) elicit gaslighting plans from LLMs with the proposed DeepGaslighting prompting template, and 2) acquire gaslighting conversations from LLMs through our Chain-of-Gaslighting method. The gaslighting conversation dataset along with a corresponding safe dataset is applied to fine-tuning-based attacks on open-source LLMs and anti-gaslighting safety alignment on these LLMs. Experiments demonstrate that both prompt-based and fine-tuning-based attacks transform three open-source LLMs into gaslighters. In contrast, we advanced three safety alignment strategies to strengthen~(by 12.05%) the safety guardrail of LLMs. Our safety alignment strategies have minimal impacts on the utility of LLMs. Empirical studies indicate that an LLM may be a potential gaslighter, even if it passed the harmfulness test on general dangerous queries.
****
Riccardo Salami,Pietro Buzzega,Matteo Mosconi,Jacopo Bonato,Luigi Sabetta,Simone Calderara
Closed-Form Merging of Parameter-Efficient Modules for Federated Continual Learning
Model merging has emerged as a crucial technique in Deep Learning, enabling the integration of multiple models into a unified system while preserving performance and scalability. In this respect, the compositional properties of low-rank adaptation techniques (e.g., LoRA) have proven beneficial, as simple averaging LoRA modules yields a single model that mostly integrates the capabilities of all individual modules. Building on LoRA, we take a step further by imposing that the merged model matches the responses of all learned modules. Solving this objective in closed form yields an indeterminate system with A and B as unknown variables, indicating the existence of infinitely many closed-form solutions. To address this challenge, we introduce LoRM, an alternating optimization strategy that trains one LoRA matrix at a time. This allows solving for each unknown variable individually, thus finding a unique solution. We apply our proposed methodology to Federated Class-Incremental Learning (FCIL), ensuring alignment of model responses both between clients and across tasks. Our method demonstrates state-of-the-art performance across a range of FCIL scenarios. The code to reproduce our experiments is available at github.com/aimagelab/fed-mammoth.
****
Sixin Zhang
Local convergence of simultaneous min-max algorithms to differential equilibrium on Riemannian manifold
We study min-max algorithms to solve zero-sum differential games on
Riemannian manifold.
Based on the notions of
differential Stackelberg equilibrium
and differential Nash equilibrium on Riemannian manifold,
we analyze the local convergence of
two representative deterministic simultaneous algorithms τ-GDA and τ-SGA
to such equilibria.
Sufficient conditions are obtained to establish the linear convergence rate
of τ-GDA based on the Ostrowski theorem on manifold and spectral analysis.
To avoid strong rotational dynamics in τ-GDA,
τ-SGA is extended from

the symplectic gradient-adjustment method in Euclidean space.
We analyze an
asymptotic approximation of τ-SGA
when the learning rate ratio τ is big.
In some cases, it can achieve a faster convergence rate
to differential Stackelberg equilibrium compared to τ-GDA.
We show numerically how the insights obtained from the
convergence analysis may improve
the training of orthogonal Wasserstein GANs using
stochastic τ-GDA and τ-SGA on simple benchmarks.
****
Junzhe Jiang,Chun Gu,Yurui Chen,Li Zhang
GS-LiDAR: Generating Realistic LiDAR Point Clouds with Panoramic Gaussian Splatting
LiDAR novel view synthesis (NVS) has emerged as a novel task within LiDAR simulation, offering valuable simulated point cloud data from novel viewpoints to aid in autonomous driving systems. However, existing LiDAR NVS methods typically rely on neural radiance fields (NeRF) as their 3D representation, which incurs significant computational costs in both training and rendering. Moreover, NeRF and its variants are designed for symmetrical scenes, making them ill-suited for driving scenarios. To address these challenges, we propose GS-LiDAR, a novel framework for generating realistic LiDAR point clouds with panoramic Gaussian splatting. Our approach employs 2D Gaussian primitives with periodic vibration properties, allowing for precise geometric reconstruction of both static and dynamic elements in driving scenarios. We further introduce a novel panoramic rendering technique with explicit ray-splat intersection, guided by panoramic LiDAR supervision. By incorporating intensity and ray-drop spherical harmonic (SH) coefficients into the Gaussian primitives, we enhance the realism of the rendered point clouds. Extensive experiments on KITTI-360 and nuScenes demonstrate the superiority of our method in terms of quantitative metrics, visual quality, as well as training and rendering efficiency.
****
Yichong Xia,Yimin Zhou,Jinpeng Wang,Baoyi An,Haoqian Wang,Yaowei Wang,Bin Chen
DiffPC: Diffusion-based High Perceptual Fidelity Image Compression with Semantic Refinement
Reconstructing high-quality images under low bitrates conditions presents a challenge, and previous methods have made this task feasible by leveraging the priors of diffusion models. However, the effective exploration of pre-trained latent diffusion models and semantic information integration in image compression tasks still needs further study. To address this issue, we introduce Diffusion-based High Perceptual Fidelity Image Compression with Semantic Refinement (DiffPC), a two-stage image compression framework based on stable diffusion. DiffPC efficiently encodes low-level image information, enabling the highly realistic reconstruction of the original image by leveraging high-level semantic features and the prior knowledge inherent in diffusion models. Specifically, DiffPC utilizes a multi-feature compressor to represent crucial low-level information with minimal bitrates and employs pre-embedding to acquire more robust hybrid semantics, thereby providing additional context for the decoding end. Furthermore, we have devised a control module tailored for image compression tasks, ensuring structural and textural consistency in reconstruction even at low bitrates and preventing decoding collapses induced by condition leakage. Extensive experiments demonstrate that our method achieves state-of-the-art perceptual fidelity and surpasses previous perceptual image compression methods by a significant margin in statistical fidelity.
****
Sheng Xu,Bo Yue,Hongyuan Zha,Guiliang Liu
A Distributional Approach to Uncertainty-Aware Preference Alignment Using Offline Demonstrations
Designing reward functions in Reinforcement Learning (RL) often demands significant task-specific expertise. Offline Preference-based Reinforcement Learning (Pb

RL) provides an effective alternative to address the complexity of reward design by learning policies from offline datasets that contain human preferences between trajectory pairs. Existing offline PbRL studies typically model a reward function by maximizing its likelihood of generating the observed human preferences. However, due to the varying number of samples within the limited dataset, less frequently compared trajectories exhibit greater uncertainty, which potentially leads to unreliable behaviors during reward and policy updates. To solve this issue, in this work, we introduce Uncertainty-Aware PbRL (UA-PbRL) to learn a distributional reward model and a risk-sensitive policy from an offline preference dataset. Our approach employs a Maximum A Posteriori (MAP) objective to update trajectory rewards and incorporates an informative prior to account for the uncertainties. Building upon this reward update, we propose a generative reward model to capture the reward distribution, utilizing the offline distributional Bellman operator and the Conditional Value-at-Risk (CVaR) metric to train a risk-sensitive policy. Experimental results demonstrate that UA-PbRL effectively identifies and avoids states with high uncertainty, facilitating risk-averse behaviors across various tasks, including robot control and language model alignment. The code is available at https://github.com/Jasonxu1225/UA-PbRL.
****
Yijie Guo,Bingjie Tang,Iretiayo Akinola,Dieter Fox,Abhishek Gupta,Yashraj Narang
SRSA: Skill Retrieval and Adaptation for Robotic Assembly Tasks
Enabling robots to learn novel tasks in a data-efficient manner is a long-standing challenge. Common strategies involve carefully leveraging prior experiences, especially transition data collected on related tasks. Although much progress has been made for general pick-and-place manipulation, far fewer studies have investigated contact-rich assembly tasks, where precise control is essential. We introduce SRSA} (Skill Retrieval and Skill Adaptation), a novel framework designed to address this problem by utilizing a pre-existing skill library containing policies for diverse assembly tasks. The challenge lies in identifying which skill from the library is most relevant for fine-tuning on a new task. Our key hypothesis is that skills showing higher zero-shot success rates on a new task are better suited for rapid and effective fine-tuning on that task. To this end, we propose to predict the transfer success for all skills in the skill library on a novel task, and then use this prediction to guide the skill retrieval process. We establish a framework that jointly captures features of object geometry, physical dynamics, and expert actions to represent the tasks, allowing us to efficiently learn the transfer success predictor. Extensive experiments demonstrate that SRSA significantly outperforms the leading baseline. When retrieving and fine-tuning skills on unseen tasks, SRSA achieves a 19% relative improvement in success rate, exhibits 2.6x lower standard deviation across random seeds, and requires 2.4x fewer transition samples to reach a satisfactory success rate, compared to the baseline. In a continual learning setup, SRSA efficiently learns policies for new tasks and incorporates them into the skill library, enhancing future policy learning. Furthermore, policies trained with SRSA in simulation achieve a 90% mean success rate when deployed in the real world. Please visit our project webpage https://srsa2024.github.io/.
****
Ting Wu,Xuefeng Li,Pengfei Liu
Progress or Regress? Self-Improvement Reversal in Post-training
Self-improvement through post-training methods such as iterative preference learning has been acclaimed for enhancing the problem-solving capabilities (e.g., mathematical reasoning) of Large Language Models (LLMs) without human intervention. However, as our exploration deepens, it is crucial to critically assess whether these enhancements indeed signify comprehensive progress or if they could lead to unintended regressions. Through rigorous experimentation and analysis across diverse problem-solving tasks, we uncover nuances in the self-improvement trajectories of LLMs. Our study introduces the concept of \emph{self-improvement reversal}, where models showing improved overall accuracy metrics might paradoxically exhibit declines in broader, essential capabilities. We propose a comprehensive evaluative framework to scrutinize the underlying mechanisms and outcomes of p

ost-training self-improvement, aiming to discern between superficial metric impr
ovements and genuine enhancements in model functionality. The findings emphasize
 the complexity of technological advancements in LLMs, underscoring the need for
 a nuanced understanding of the \textit{progress or regress} dichotomy in their
development.
****
Victor Besnier,Mickael Chen,David Hurych,Eduardo Valle,Matthieu Cord
Halton Scheduler for Masked Generative Image Transformer
Masked Generative Image Transformers (MaskGIT) have emerged as a scalable
and efficient image generation framework, able to deliver high-quality visuals w
ith
low inference costs. However, MaskGIT’s token unmasking scheduler, an essential
component of the framework, has not received the attention it deserves. We analy
ze
the sampling objective in MaskGIT, based on the mutual information between
tokens, and elucidate its shortcomings. We then propose a new sampling strategy
based on our Halton scheduler instead of the original Confidence scheduler. More
precisely, our method selects the token’s position according to a quasi-random,
low-discrepancy Halton sequence. Intuitively, that method spreads the tokens
spatially, progressively covering the image uniformly at each step. Our analysis
shows that it allows reducing non-recoverable sampling errors, leading to simple
r
hyper-parameters tuning and better quality images. Our scheduler does not requir
e
retraining or noise injection and may serve as a simple drop-in replacement for
the original sampling strategy. Evaluation of both class-to-image synthesis on
ImageNet and text-to-image generation on the COCO dataset demonstrates that the
Halton scheduler outperforms the Confidence scheduler quantitatively by reducing
the FID and qualitatively by generating more diverse and more detailed images.
Our code is at https://github.com/valeoai/Halton-MaskGIT.
****
Junxi Wu,Dongjian Hu,Yajie Bao,Shu-Tao Xia,Changliang Zou
Error-quantified Conformal Inference for Time Series
Uncertainty quantification in time series prediction is challenging due to the t
emporal dependence and distribution shift on sequential data. Conformal predicti
on provides a pivotal and flexible instrument for assessing the uncertainty of m
achine learning models through prediction sets. Recently, a series of online con
formal inference methods updated thresholds of prediction sets by performing onl
ine gradient descent on a sequence of quantile loss functions. A drawback of suc
h methods is that they only use the information of revealed non-conformity score
s via miscoverage indicators but ignore error quantification, namely the distanc
e between the non-conformity score and the current threshold. To accurately leve
rage the dynamic of miscoverage error, we propose Error-quantified Conformal Inf
erence (ECI) by smoothing the quantile loss function. ECI introduces a continuou
s and adaptive feedback scale with the miscoverage error, rather than simple bin
ary feedback in existing methods. We establish a long-term coverage guarantee fo
r ECI under arbitrary dependence and distribution shift. The extensive experimen
tal results show that ECI can achieve valid miscoverage control and output tight
er prediction sets than other baselines.
****
Chung-En Sun,Tuomas Oikarinen,Berk Ustun,Tsui-Wei Weng
Concept Bottleneck Large Language Models
We introduce Concept Bottleneck Large Language Models (CB-LLMs), a novel framewo
rk for building inherently interpretable Large Language Models (LLMs). In contra
st to traditional black-box LLMs that rely on limited post-hoc interpretations,
CB-LLMs integrate intrinsic interpretability directly into the LLMs -- allowing
accurate explanations with scalability and transparency. We build CB-LLMs for tw
o essential NLP tasks: text classification and text generation. In text classifi
cation, CB-LLMs is competitive with, and at times outperforms, traditional black
-box models while providing explicit and interpretable reasoning. For the more c

hallenging task of text generation, interpretable neurons in CB-LLMs enable prec
ise concept detection, controlled generation, and safer outputs. The embedded in
terpretability empowers users to transparently identify harmful content, steer m
odel behavior, and unlearn undesired concepts -- significantly enhancing the saf
ety, reliability, and trustworthiness of LLMs, which are critical capabilities n
otably absent in existing language models.
****

Marco Pacini,Xiaowen Dong,Bruno Lepri,Gabriele Santin
Separation Power of Equivariant Neural Networks
The separation power of a machine learning model refers to its ability to distin
guish between different inputs and is often used as a proxy for its expressivity
. Indeed, knowing the separation power of a family of models is a necessary cond
ition to obtain fine-grained universality results. In this paper, we analyze the
 separation power of equivariant neural networks, such as convolutional and perm
utation-invariant networks.
We first present a complete characterization of inputs indistinguishable by mode
ls derived by a given architecture. From this results, we derive how separabilit
y is influenced by hyperparameters and architectural choices—such as activation
functions, depth, hidden layer width, and representation types. Notably, all non
-polynomial activations, including ReLU and sigmoid, are equivalent in expressiv
ity and reach maximum separation power. Depth improves separation power up to a
threshold, after which further increases have no effect. Adding invariant featur
es to hidden representations does not impact separation power. Finally, block de
composition of hidden representations affects separability, with minimal compone
nts forming a hierarchy in separation power that provides a straightforward meth
od for comparing the separation power of models.
****

Bhavya Sukhija,Stelian Coros,Andreas Krause,Pieter Abbeel,Carmelo Sferrazza
MaxInfoRL: Boosting exploration in reinforcement learning through information ga
in maximization
Reinforcement learning (RL) algorithms aim to balance exploiting the current bes
t strategy with exploring new options that could lead to higher rewards. Most co
mmon RL algorithms use undirected exploration, i.e., select random sequences of
actions.
Exploration can also be directed using intrinsic rewards, such as curiosity or m
odel epistemic uncertainty. However, effectively balancing task and intrinsic re
wards is challenging and often task-dependent. In this work, we introduce a fram
ework, MaxInfoRL, for balancing intrinsic and extrinsic exploration. MaxInfoRL s
teers exploration towards informative transitions, by maximizing intrinsic rewar
ds such as the information gain about the underlying task.
When combined with Boltzmann exploration, this approach naturally trades off max
imization of the value function with that of the entropy over states, rewards, a
nd actions. We show that our approach achieves sublinear regret in the simplifie
d setting of multi-armed bandits. We then apply this general formulation to a va
riety of off-policy model-free RL methods for continuous state-action spaces, yi
elding novel algorithms that achieve superior performance across hard exploratio
n problems and complex scenarios such as visual control tasks.
****

Jihyo Kim,Seulbi Lee,Sangheum Hwang
Reflexive Guidance: Improving OoDD in Vision-Language Models via Self-Guided Ima
ge-Adaptive Concept Generation
With the recent emergence of foundation models trained on internet-scale data an
d demonstrating remarkable generalization capabilities, such foundation models h
ave become more widely adopted, leading to an expanding range of application dom
ains. Despite this rapid proliferation, the trustworthiness of foundation models
 remains underexplored. Specifically, the out-of-distribution detection (OoDD) c
apabilities of large vision-language models (LVLMs), such as GPT-4o, which are t
rained on massive multi-modal data, have not been sufficiently addressed. The di
sparity between their demonstrated potential and practical reliability raises co
ncerns regarding the safe and trustworthy deployment of foundation models. To ad

dress this gap, we evaluate and analyze the OoDD capabilities of various proprie
tary and open-source LVLMs. Our investigation contributes to a better understand
ing of how these foundation models represent confidence scores through their gen
erated natural language responses. Furthermore, we propose a self-guided prompti
ng approach, termed Reflexive Guidance (ReGuide), aimed at enhancing the OoDD ca
pability of LVLMs by leveraging self-generated image-adaptive concept suggestion
s. Experimental results demonstrate that our ReGuide enhances the performance of
 current LVLMs in both image classification and OoDD tasks. The lists of sampled
 images, along with the prompts and responses for each sample are available at h
ttps://github.com/daintlab/ReGuide.
****

Lewis Hammond,Sam Adam-Day
Neural Interactive Proofs
We consider the problem of how a trusted, but computationally bounded agent (a '
verifier') can learn to interact with one or more powerful but untrusted agents
('provers') in order to solve a given task. More specifically, we study the case
 in which agents are represented using neural networks and refer to solutions of
 this problem as neural interactive proofs. First we introduce a unifying framew
ork based on prover-verifier games (Anil et al., 2021), which generalises previo
usly proposed interaction protocols. We then describe several new protocols for
generating neural interactive proofs, and provide a theoretical comparison of bo
th new and existing approaches. Finally, we support this theory with experiments
 in two domains: a toy graph isomorphism problem that illustrates the key ideas,
 and a code validation task using large language models. In so doing, we aim to
create a foundation for future work on neural interactive proofs and their appli
cation in building safer AI systems.
****

Botao Ren,Xue Yang,Yi Yu,Junwei Luo,Zhidong Deng
PointOBB-v2: Towards Simpler, Faster, and Stronger Single Point Supervised Orien
ted Object Detection
Single point supervised oriented object detection has gained attention and made
initial progress within the community. Diverse from those approaches relying on
one-shot samples or powerful pretrained models (e.g. SAM), PointOBB has shown pr
omise due to its prior-free feature. In this paper, we propose PointOBB-v2, a si
mpler, faster, and stronger method to generate pseudo rotated boxes from points
without relying on any other prior. Specifically, we first generate a Class Prob
ability Map (CPM) by training the network with non-uniform positive and negative
 sampling. We show that the CPM is able to learn the approximate object regions
and their contours. Then, Principal Component Analysis (PCA) is applied to accur
ately estimate the orientation and the boundary of objects. By further incorpor
ating a separation mechanism, we resolve the confusion caused by the overlapping
 on the CPM, enabling its operation in high-density scenarios. Extensive compar
isons demonstrate that our method achieves a training speed 15.58$\times$ faster
 and an accuracy improvement of 11.60\%/25.15\%/21.19\% on the DOTA-v1.0/v1.5/v2
.0 datasets compared to the previous state-of-the-art, PointOBB. This significan
tly advances the cutting edge of single point supervised oriented detection in t
he modular track. Code and models will be released.
****

Scott Fujimoto,Pierluca D'Oro,Amy Zhang,Yuandong Tian,Michael Rabbat
Towards General-Purpose Model-Free Reinforcement Learning
Reinforcement learning (RL) promises a framework for near-universal problem-solv
ing. In practice however, RL algorithms are often tailored to specific benchmark
s, relying on carefully tuned hyperparameters and algorithmic choices. Recently,
 powerful model-based RL methods have shown impressive general results across be
nchmarks but come at the cost of increased complexity and slow run times, limiti
ng their broader applicability. In this paper, we attempt to find a unifying mod
el-free deep RL algorithm that can address a diverse class of domains and proble
m settings. To achieve this, we leverage model-based representations that approx
imately linearize the value function, taking advantage of the denser task object
ives used by model-based RL while avoiding the costs associated with planning or

simulated trajectories. We evaluate our algorithm, MR.Q, on a variety of common RL benchmarks with a single set of hyperparameters and show a competitive performance against domain-specific and general baselines, providing a concrete step towards building general-purpose model-free deep RL algorithms.
****

Dmitry Yarotsky,Maksim Velikanov
SGD with memory: fundamental properties and stochastic acceleration
An important open problem is the theoretically feasible acceleration of mini-batch SGD-type algorithms on quadratic problems with power-law spectrum. In the non-stochastic setting, the optimal exponent ξ in the loss convergence $L_t\sim C_Lt^{-\xi}$ is double that in plain GD and is achievable using Heavy Ball (HB) with a suitable schedule; this no longer works in the presence of mini-batch noise. We address this challenge by considering first-order methods with an arbitrary fixed number M of auxiliary velocity vectors (*memory-M algorithms*). We first prove an equivalence between two forms of such algorithms and describe them in terms of suitable characteristic polynomials. Then we develop a general expansion of the loss in terms of *signal and noise propagators*. Using it, we show that losses of stationary stable memory-M algorithms always retain the exponent ξ of plain GD, but can have different constants C_L depending on their *effective learning rate* that generalizes that of HB. We prove that in memory-1 algorithms we can make C_L arbitrarily small while maintaining stability. As a consequence, we propose a memory-1 algorithm with a time-dependent schedule that we show heuristically and experimentally to improve the exponent ξ of plain SGD.
****

Thomas Stegmüller,Tim Lebailly,Nikola ■uki■,Behzad Bozorgtabar,Tinne Tuytelaars,Jean-Philippe Thiran
A Simple Framework for Open-Vocabulary Zero-Shot Segmentation
Zero-shot classification capabilities naturally arise in models trained within a vision-language contrastive framework. Despite their classification prowess, these models struggle in dense tasks like zero-shot open-vocabulary segmentation. This deficiency is often attributed to the absence of localization cues in captions and the intertwined nature of the learning process, which encompasses both image/text representation learning and cross-modality alignment. To tackle these issues, we propose SimZSS, a $\textbf{Sim}$ple framework for open-vocabulary $\textbf{Z}$ero-$\textbf{S}$hot $\textbf{S}$egmentation. The method is founded on two key principles: i) leveraging frozen vision-only models that exhibit spatial awareness while exclusively aligning the text encoder and ii) exploiting the discrete nature of text and linguistic knowledge to pinpoint local concepts within captions. By capitalizing on the quality of the visual representations, our method requires only image-caption pair datasets and adapts to both small curated and large-scale noisy datasets. When trained on COCO Captions across 8 GPUs, SimZSS achieves state-of-the-art results on 7 out of 8 benchmark datasets in less than 15 minutes. Our code and pretrained models are publicly available at https://github.com/tileb1/simzss.
****

Chi Zhang,Zain Ulabedeen Farhat,George K. Atia,Yue Wang
Model-Free Offline Reinforcement Learning with Enhanced Robustness
Offline reinforcement learning (RL) has gained considerable attention for its ability to learn policies from pre-collected data without real-time interaction, which makes it particularly useful for high-risk applications. However, due to its reliance on offline datasets, existing works inevitably introduce assumptions to ensure effective learning, which, however, often lead to a trade-off between robustness to model mismatch and scalability to large environments. In this paper, we enhance both aspects with a novel double-pessimism principle, which conservatively estimates performance and accounts for both limited data and potential model mismatches, two major reasons for the previous trade-off. We then propose a universal, model-free algorithm to learn a policy that is robust to potential environment mismatches, which enhances robustness in a scalable manner. Furthermore, we provide a sample complexity analysis of our algorithm when the mismatch

is modeled by the l_α-norm, which also theoretically demonstrates the eff
iciency of our method. Extensive experiments further demonstrate that our approa
ch significantly improves robustness in a more scalable manner than existing met
hods.
****
Xirui Li,Hengguang Zhou,Ruochen Wang,Tianyi Zhou,Minhao Cheng,Cho-Jui Hsieh
Is Your Multimodal Language Model Oversensitive to Safe Queries?
Humans are prone to cognitive distortions — biased thinking patterns that lead t
o exaggerated responses to specific stimuli, albeit in very different contexts.
This paper demonstrates that advanced Multimodal Large Language Models (MLLMs) e
xhibit similar tendencies.
While these models are designed to respond queries under safety mechanism, they
sometimes reject harmless queries in the presence of certain visual stimuli, dis
regarding the benign nature of their contexts.
As the initial step in investigating this behavior, we identify three representa
tive types of stimuli that trigger the oversensitivity of existing MLLMs: $\text
bf{\textit{Exaggerated Risk}}$, $\textbf{\textit{Negated Harm}}$, and $\textbf{\
textit{Counterintuitive Interpretation}}$.
To systematically evaluate MLLMs' oversensitivity to these stimuli, we propose t
he $\textbf{M}$ultimodal $\textbf{O}$ver$\textbf{S}$en$\textbf{S}$itivity $\text
bf{Bench}$mark (MOSSBench).
This toolkit consists of 300 manually collected benign multimodal queries, cross
-verified by third-party reviewers (AMT).
Empirical studies using MOSSBench on 20 MLLMs reveal several insights:
(1). Oversensitivity is prevalent among SOTA MLLMs, with refusal rates reaching
up to $\textbf{76}$\% for harmless queries.
(2). Safer models are more oversensitive: increasing safety may inadvertently ra
ise caution and conservatism in the model’s responses.
(3). Different types of stimuli tend to cause errors at specific stages — percep
tion, intent reasoning, and safety judgement — in the response process of MLLMs.
These findings highlight the need for refined safety mechanisms that balance cau
tion with contextually appropriate responses, improving the reliability of MLLMs
 in real-world applications.
****
Valentinos Pariza,Mohammadreza Salehi,Gertjan J. Burghouts,Francesco Locatello,Y
uki M Asano
Near, far: Patch-ordering enhances vision foundation models' scene understanding
We introduce NeCo: Patch Neighbor Consistency, a novel self-supervised training
loss that enforces patch-level nearest neighbor consistency across a student and
 teacher model. Compared to contrastive approaches that only yield binary learni
ng signals, i.e. "attract" and "repel", this approach benefits from the more fin
e-grained learning signal of sorting spatially dense features relative to refere
nce patches. Our method leverages differentiable sorting applied on top of pretr
ained representations, such as DINOv2-registers to bootstrap the learning signal
 and further improve upon them. This dense post-pretraining leads to superior pe
rformance across various models and datasets, despite requiring only 19 hours on
 a single GPU. This method generates high-quality dense feature encoders and est
ablishes several new state-of-the-art results such as +5.5 % and +6% for non-par
ametric in-context semantic segmentation on ADE20k and Pascal VOC, +7.2% and +5.
7% for linear segmentation evaluations on COCO-Things and -Stuff and improvement
s in the 3D understanding of multi-view consistency on SPair-71k, by more than 1
.5%.
****
Enshu Liu,Junyi Zhu,Zinan Lin,Xuefei Ning,Shuaiqi Wang,Matthew B. Blaschko,Serge
y Yekhanin,Shengen Yan,Guohao Dai,Huazhong Yang,Yu Wang
Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models B
etter
Diffusion Models (DM) and Consistency Models (CM) are two types of popular gener
ative models with good generation quality on various tasks. When training DM and
 CM, intermediate weight checkpoints are not fully utilized and only the last co

nverged checkpoint is used. In this work, we find proper checkpoint merging can significantly improve the training convergence and final performance. Specifically, we propose LCSC, a simple but effective and efficient method to enhance the performance of DM and CM, by combining checkpoints along the training trajectory with coefficients deduced from evolutionary search. We demonstrate the value of LCSC through two use cases: (a) Reducing training cost. With LCSC, we only need to train DM/CM with fewer number of iterations and/or lower batch sizes to obtain comparable sample quality with the fully trained model. For example, LCSC achieves considerable training speedups for CM (23$\times$ on CIFAR-10 and 15$\times$ on ImageNet-64). (b) Enhancing pre-trained models. When full training is already done, LCSC can further improve the generation quality or efficiency of the final converged models. For example, LCSC achieves better FID using 1 number of function evaluation (NFE) than the base model with 2 NFE on consistency distillation, and decreases the NFE of DM from 15 to 9 while maintaining the generation quality. Applying LCSC to large text-to-image models, we also observe clearly enhanced generation quality.
****

Federico Adolfi,Martina G. Vilas,Todd Wareham
The Computational Complexity of Circuit Discovery for Inner Interpretability
Many proposed applications of neural networks in machine learning, cognitive/brain science, and society hinge on the feasibility of inner interpretability via circuit discovery. This calls for empirical and theoretical explorations of viable algorithmic options. Despite advances in the design and testing of heuristics, there are concerns about their scalability and faithfulness at a time when we lack understanding of the complexity properties of the problems they are deployed to solve. To address this, we study circuit discovery with classical and parameterized computational complexity theory: (1) we describe a conceptual scaffolding to reason about circuit finding queries in terms of affordances for description, explanation, prediction and control; (2) we formalize a comprehensive set of queries for mechanistic explanation, and propose a formal framework for their analysis; (3) we use it to settle the complexity of many query variants and relaxations of practical interest on multi-layer perceptrons. Our findings reveal a challenging complexity landscape. Many queries are intractable, remain fixed-parameter intractable relative to model/circuit features, and inapproximable under additive, multiplicative, and probabilistic approximation schemes. To navigate this landscape, we prove there exist transformations to tackle some of these hard problems with better-understood heuristics, and prove the tractability or fixed-parameter tractability of more modest queries which retain useful affordances. This framework allows us to understand the scope and limits of interpretability queries, explore viable options, and compare their resource demands on existing and future architectures.
****

Yi Ding,Bolian Li,Ruqi Zhang
ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time
Vision Language Models (VLMs) have become essential backbones for multi-modal intelligence, yet significant safety challenges limit their real-world application. While textual inputs can often be effectively safeguarded, adversarial visual inputs can often easily bypass VLM defense mechanisms. Existing defense methods are either resource-intensive, requiring substantial data and compute, or fail to simultaneously ensure safety and usefulness in responses. To address these limitations, we propose a novel two-phase inference-time alignment framework, **E**valuating **T**hen **A**ligning (ETA): i) Evaluating input visual contents and output responses to establish a robust safety awareness in multimodal settings, and ii) Aligning unsafe behaviors at both shallow and deep levels by conditioning the VLMs' generative distribution with an interference prefix and performing sentence-level best-of-N to search the most harmless and helpful generation paths. Extensive experiments show that ETA outperforms baseline methods in terms of harmlessness, helpfulness, and efficiency, reducing the unsafe rate by 87.5\% in cross-modality attacks and achieving 96.6\% win-ties in GPT-4 helpfulness evaluation.

****
Yuheng Zhang,Nan Jiang
Statistical Tractability of Off-policy Evaluation of History-dependent Policies in POMDPs
We investigate off-policy evaluation (OPE), a central and fundamental problem in reinforcement learning (RL), in the challenging setting of Partially Observable Markov Decision Processes (POMDPs) with large observation spaces. Recent works of Uehara et al. (2023a); Zhang & Jiang (2024) developed a model-free framework and identified important coverage assumptions (called belief and outcome coverage) that enable accurate OPE of memoryless policies with polynomial sample complexities, but handling more general target policies that depend on the entire observable history remained an open problem. In this work, we prove information-theoretic hardness for model-free OPE of history-dependent policies in several settings, characterized by additional assumptions imposed on the behavior policy (memoryless vs. history-dependent) and/or the state-revealing property of the POMDP (single-step vs. multi-step revealing). We further show that some hardness can be circumvented by a natural model-based algorithm—whose analysis has surprisingly eluded the literature despite the algorithm's simplicity—demonstrating provable separation between model-free and model-based OPE in POMDPs.
****
Minh Le,Chau Nguyen,Huy Nguyen,Quyen Tran,Trung Le,Nhat Ho
Revisiting Prefix-tuning: Statistical Benefits of Reparameterization among Prompts
Prompt-based techniques, such as prompt-tuning and prefix-tuning, have gained prominence for their efficiency in fine-tuning large pre-trained models. Despite their widespread adoption, the theoretical foundations of these methods remain limited. For instance, in prefix-tuning, we observe that a key factor in achieving performance parity with full fine-tuning lies in the reparameterization strategy. However, the theoretical principles underpinning the effectiveness of this approach have yet to be thoroughly examined. Our study demonstrates that reparameterization is not merely an engineering trick but is grounded in deep theoretical foundations. Specifically, we show that the reparameterization strategy implicitly encodes a shared structure between prefix key and value vectors. Building on recent insights into the connection between prefix-tuning and mixture of experts models, we further illustrate that this shared structure significantly improves sample efficiency in parameter estimation compared to non-shared alternatives. The effectiveness of prefix-tuning across diverse tasks is empirically confirmed to be enhanced by the shared structure, through extensive experiments in both visual and language domains. Additionally, we uncover similar structural benefits in prompt-tuning, offering new perspectives on its success. Our findings provide theoretical and empirical contributions, advancing the understanding of prompt-based methods and their underlying mechanisms.
****
Tal Wagner
Learning from End User Data with Shuffled Differential Privacy over Kernel Densities
We study a setting of collecting and learning from private data distributed across end users.
In the shuffled model of differential privacy, the end users partially protect their data locally before sharing it, and their data is also anonymized during its collection to enhance privacy.
This model has recently become a prominent alternative to central DP, which requires full trust in a central data curator, and local DP, where fully local data protection takes a steep toll on downstream accuracy.

Our main technical result is a shuffled DP protocol for privately estimating the

kernel density function of a distributed dataset, with accuracy essentially matching central DP.
We use it to privately learn a classifier from the end user data, by learning a private density function per class.
Moreover, we show that the density function itself can recover the semantic content of its class, despite having been learned in the absence of any unprotected data.
Our experiments show the favorable downstream performance of our approach, and highlight key downstream considerations and trade-offs in a practical ML deployment of shuffled DP.
****
Ankit Sonthalia,Alexander Rubinstein,Ehsan Abbasnejad,Seong Joon Oh
Do Deep Neural Network Solutions Form a Star Domain?
It has recently been conjectured that neural network solution sets reachable via stochastic gradient descent (SGD) are convex, considering permutation invariances. This means that a linear path can connect two independent solutions with low loss, given the weights of one of the models are appropriately permuted. However, current methods to test this theory often require very wide networks to succeed. In this work, we conjecture that more generally, the SGD solution set is a star domain that contains a star model that is linearly connected to all the other solutions via paths with low loss values, modulo permutations. We propose the Starlight algorithm that finds a star model of a given learning task. We validate our claim by showing that this star model is linearly connected with other independently found solutions. As an additional benefit of our study, we demonstrate better uncertainty estimates on Bayesian Model Averaging over the obtained star domain. Further, we demonstrate star models as potential substitutes for model ensembles.
****
Lorenzo Basile,Santiago Acevedo,Luca Bortolussi,Fabio Anselmi,Alex Rodriguez
Intrinsic Dimension Correlation: uncovering nonlinear connections in multimodal representations
To gain insight into the mechanisms behind machine learning methods, it is crucial to establish connections among the features describing data points. However, these correlations often exhibit a high-dimensional and strongly nonlinear nature, which makes them challenging to detect using standard methods. This paper exploits the entanglement between intrinsic dimensionality and correlation to propose a metric that quantifies the (potentially nonlinear) correlation between high-dimensional manifolds. We first validate our method on synthetic data in controlled environments, showcasing its advantages and drawbacks compared to existing techniques. Subsequently, we extend our analysis to large-scale applications in neural network representations. Specifically, we focus on latent representations of multimodal data, uncovering clear correlations between paired visual and textual embeddings, whereas existing methods struggle significantly in detecting similarity. Our results indicate the presence of highly nonlinear correlation patterns between latent manifolds.
****
Lokesh Veeramacheneni,Moritz Wolter,Hilde Kuehne,Juergen Gall
Fréchet Wavelet Distance: A Domain-Agnostic Metric for Image Generation
Modern metrics for generative learning like Fréchet Inception Distance (FID) and DINOv2-Fréchet Distance (FD-DINOv2) demonstrate impressive performance. However, they suffer from various shortcomings, like a bias towards specific generators and datasets. To address this problem, we propose the Fréchet Wavelet Distance (FWD) as a domain-agnostic metric based on the Wavelet Packet Transform ($\mathcal{W}_p$). FWD provides a sight across a broad spectrum of frequencies in images with a high resolution, preserving both spatial and textural aspects. Specifically, we use $\mathcal{W}_p$ to project generated and real images to the packet coefficient space. We then compute the Fréchet distance with the resultant coefficients to evaluate the quality of a generator. This metric is general-purpose and dataset-domain agnostic, as it does not rely on any pre-trained network, while being more interpretable due to its ability to compute Fréchet distance per pac

ket, enhancing transparency. We conclude with an extensive evaluation of a wide variety of generators across various datasets that the proposed FWD can generalize and improve robustness to domain shifts and various corruptions compared to other metrics.
****
Zeman Li,Xinwei Zhang,Peilin Zhong,Yuan Deng,Meisam Razaviyayn,Vahab Mirrokni
Addax: Utilizing Zeroth-Order Gradients to Improve Memory Efficiency and Performance of SGD for Fine-Tuning Language Models
Fine-tuning language models (LMs) with the standard Adam optimizer often demands excessive memory, limiting accessibility. The ``in-place'' version of Stochastic Gradient Descent (IP-SGD) and Memory-Efficient Zeroth-order Optimizer (MeZO) have been proposed as solutions to improve memory efficiency. However, IP-SGD still requires a decent amount of memory, and MeZO suffers from slow convergence and degraded final performance due to its zeroth-order nature. This paper introduces Addax, a novel method that improves both memory efficiency and algorithm performance of IP-SGD by integrating it with MeZO. Specifically, Addax computes the zeroth-order or first-order gradient of the data points in the minibatch based on their memory consumption and combines zeroth- and first-order gradient estimates to obtain the updated direction in each step.
By computing the zeroth-order order gradient of data points that require more memory and the first-order gradient of the ones that require less memory, Addax overcomes the slow convergence of MeZO and excessive memory requirement of IP-SGD. Additionally, the zeroth-order gradient acts as a regularizer for the first-order gradient, further enhancing the model's final performance.
Theoretically, we establish the convergence of Addax under mild assumptions, demonstrating faster convergence and less restrictive hyper-parameter choices than MeZO. Our extensive experiments with diverse LMs and tasks show that Addax consistently outperforms MeZO in terms of accuracy and convergence speed, while having a comparable memory footprint.
In particular, our experiments using one A100 GPU on OPT-13B model reveal that, on average, Addax outperforms MeZO in terms of accuracy/F1 score by 14%, and runs $15\times$ faster, while having a comparable memory footprint to MeZO. In our experiments on the larger OPT-30B model, on average, Addax outperforms MeZO in terms of accuracy/F1 score by >16% and runs $30\times$ faster on a single H100 GPU. Moreover, Addax surpasses the performance of standard fine-tuning approaches, such as IP-SGD and Adam, in most tasks in terms of Accuracy/F1 score with significantly less memory requirement.
****
Chuan Liu,Ruibing Song,Chunshu Wu,Pouya Haghi,Tong Geng
InstaTrain: Adaptive Training via Ultra-Fast Natural Annealing within Dynamical Systems
Time-series modeling is broadly adopted to capture underlying patterns present in historical data, allowing prediction of future values. However, one crucial aspect of such modeling is often overlooked: in highly dynamic environments, data distributions can shift drastically within a second or less. Under these circumstances, traditional predictive models, and even online learning methods, struggle to adapt to the ultra-fast and complex distribution shifts present in highly dynamic scenarios. To address this, we propose InstaTrain, a novel learning approach that enables ultra-fast model updates for real-world prediction tasks, thereby keeping pace with rapidly evolving data distributions. In this work, (1) we transform the slow and expensive training process into an ultra-fast natural annealing process within a dynamical system. (2) Leveraging a recently proposed electronic dynamical system, we augment the system with parameter update modules, extending its capabilities to encompass both rapid training and inference. Experimental results on highly dynamic datasets demonstrate that our method achieves orders-of-magnitude improvements in training speed and energy efficiency while delivering superior accuracy compared to baselines running on GPUs.
****
Haodong Liang,Krishna Balasubramanian,Lifeng Lai
Transformers Handle Endogeneity in In-Context Linear Regression

We explore the capability of transformers to address endogeneity in in-context l
inear regression. Our main finding is that transformers inherently possess a mec
hanism to handle endogeneity effectively using instrumental variables (IV). Firs
t, we demonstrate that the transformer architecture can emulate a gradient-based
 bi-level optimization procedure that converges to the widely used two-stage lea
st squares (2SLS) solution at an exponential rate. Next, we propose an in-contex
t pretraining scheme and provide theoretical guarantees showing that the global
minimizer of the pre-training loss achieves a small excess loss. Our extensive e
xperiments validate these theoretical findings, showing that the trained transfo
rmer provides more robust and reliable in-context predictions and coefficient es
timates than the 2SLS method, in the presence of endogeneity.
****
Dongping Chen,Yue Huang,Siyuan Wu,Jingyu Tang,Huichi Zhou,Qihui Zhang,Zhigang He
,Yilin Bai,Chujie Gao,Liuyi Chen,Yiqiang Li,Chenlong Wang,Yue Yu,Tianshuo Zhou,Z
hen Li,Yi Gui,Yao Wan,Pan Zhou,Jianfeng Gao,Lichao Sun
GUI-World: A Video Benchmark and Dataset for Multimodal GUI-oriented Understandi
ng
Recently, Multimodal Large Language Models (MLLMs) have been used as agents to c
ontrol keyboard and mouse inputs by directly perceiving the Graphical User Inter
face (GUI) and generating corresponding commands.
However, current agents primarily demonstrate strong understanding capabilities
in static environments and are mainly applied to relatively simple domains, such
 as Web or mobile interfaces.
We argue that a robust GUI agent should be capable of perceiving temporal inform
ation on the GUI, including dynamic Web content and multi-step tasks.
Additionally, it should possess a comprehensive understanding of various GUI sce
narios, including desktop software and multi-window interactions.
To this end, this paper introduces a new dataset, termed GUI-World, which featur
es meticulously crafted Human-MLLM annotations, extensively covering six GUI sce
narios and eight types of GUI-oriented questions in three formats.
We evaluate the capabilities of current state-of-the-art MLLMs, including Image
LLMs and Video LLMs, in understanding various types of GUI content, especially d
ynamic and sequential content. Our findings reveal that current models struggle
with dynamic GUI content without manually annotated keyframes or operation histo
ry. On the other hand, Video LLMs fall short in all GUI-oriented tasks given the
 sparse GUI video dataset. Therefore, we take the initial step of leveraging a f
ine-tuned Video LLM, GUI-Vid, as a GUI-oriented assistant, demonstrating an impr
oved understanding of various GUI tasks. However, due to the limitations in the
performance of base LLMs, we conclude that using video LLMs as GUI agents remain
s a significant challenge. We believe our work provides valuable insights for fu
ture research in dynamic GUI content understanding. All the dataset and code are
 publicly available at: https://gui-world.github.io.
****
Zongzhao Li,Jiacheng Cen,Wenbing Huang,Taifeng Wang,Le Song
Size-Generalizable RNA Structure Evaluation by Exploring Hierarchical Geometries
Understanding the 3D structure of RNA is essential for deciphering its function
and developing RNA-based therapeutics. Geometric Graph Neural Networks (GeoGNNs)
 that conform to the $\mathrm{E}(3)$-symmetry have advanced RNA structure evalua
tion, a crucial step toward RNA structure prediction. However, existing GeoGNNs
are still defective in two aspects: 1. inefficient or incapable of capturing the
 full geometries of RNA; 2. limited generalization ability when the size of RNA
significantly differs between training and test datasets. In this paper, we prop
ose EquiRNA, a novel equivariant GNN model by exploring the three-level hierarch
ical geometries of RNA. At its core, EquiRNA effectively addresses the size gene
ralization challenge by reusing the representation of nucleotide, the common bui
lding block shared across RNAs of varying sizes. Moreover, by adopting a scalari
zation-based equivariant GNN as the backbone, our model maintains directional in
formation while offering higher computational efficiency compared to existing Ge
oGNNs. Additionally, we propose a size-insensitive K-nearest neighbor sampling
 strategy to enhance the model's robustness to RNA size shifts. We test our appr

oach on our created benchmark as well as an existing dataset. The results show t
hat our method significantly outperforms other state-of-the-art methods, providi
ng a robust baseline for RNA 3D structure modeling and evaluation.
****

Tengyang Xie,Dylan J Foster,Akshay Krishnamurthy,Corby Rosset,Ahmed Hassan Awada
llah,Alexander Rakhlin
Exploratory Preference Optimization: Harnessing Implicit Q*-Approximation for Sa
mple-Efficient RLHF
This paper investigates a basic question in reinforcement learning from human fe
edback (RLHF) from a theoretical perspective: how to efficiently explore in an o
nline manner under preference feedback and general function approximation. We ta
ke the initial step towards a theoretical understanding of this problem by propo
sing a novel algorithm, *Exploratory Preference Optimization* (XPO). This algori
thm is elegantly simple---requiring only a one-line modification to (online) Dir
ect Preference Optimization (DPO; Rafailov et al., 2023)---yet provides the str
ongest known provable guarantees. XPO augments the DPO objective with a novel an
d principled *exploration bonus*, enabling the algorithm to strategically explor
e beyond the support of the initial model and preference feedback data. We prove
that XPO is provably sample-efficient and converges to a near-optimal policy un
der natural exploration conditions, regardless of the initial model's coverage.
Our analysis builds on the observation that DPO implicitly performs a form of *B
ellman error minimization*. It synthesizes previously disparate techniques from
language modeling and theoretical reinforcement learning in a serendipitous fash
ion through the lens of *KL-regularized Markov decision processes*.
****

Siyan Zhao,Mingyi Hong,Yang Liu,Devamanyu Hazarika,Kaixiang Lin
Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following
in LLMs
Large Language Models (LLMs) are increasingly deployed as chatbots, yet their ab
ility to personalize responses to user preferences remains limited. We introduce
PrefEval, a benchmark for evaluating LLMs' ability to infer, memorize and adher
e to user preferences in long-context conversational setting.
PrefEval comprises 3,000 manually curated user preference and query pairs spanni
ng 20 topics. PrefEval contains user personalization or preference information i
n both explicit and implicit preference forms, and evaluates LLM performance usi
ng a generation and a classification task. With PrefEval, we have evaluated 10 o
pen-sourced and
proprietary LLMs in multi-session conversations with varying context lengths up
to 100k tokens. We benchmark with various prompting, iterative feedback, and ret
rieval-augmented generation methods.
Our benchmarking effort reveals that state-of-the-art LLMs face significant chal
lenges in following users' preference during conversations. In particular, in z
ero-shot settings, preference following accuracy falls below 10\% at merely 10 t
urns (~3k tokens) across most evaluated models. Even with advanced prompting and
retrieval methods, preference following still deteriorates in long-context conv
ersations. Furthermore, we show that fine-tuning on PrefEval significantly impro
ves performance. We believe PrefEval serves as a valuable resource for measuring
, understanding, and enhancing LLMs' proactive preference following abilities, p
aving the way for personalized conversational agents.
****

Eliahu Horwitz,Asaf Shul,Yedid Hoshen
Unsupervised Model Tree Heritage Recovery
The number of models shared online has recently skyrocketed, with over one milli
on public models available on Hugging Face. Sharing models allows other users to
build on existing models, using them as initialization for fine-tuning, improvi
ng accuracy, and saving compute and energy. However, it also raises important in
tellectual property issues, as fine-tuning may violate the license terms of the
original model or that of its training data. A Model Tree, i.e., a tree data str
ucture rooted at a foundation model and having directed edges between a parent m
odel and other models directly fine-tuned from it (children), would settle such

disputes by making the model heritage explicit. Unfortunately, current models are not well documented, with most model metadata (e.g., "model cards") not providing accurate information about heritage. In this paper, we introduce the task of Unsupervised Model Tree Heritage Recovery (Unsupervised MoTHer Recovery) for collections of neural networks. For each pair of models, this task requires: i) determining if they are directly related, and ii) establishing the direction of the relationship. Our hypothesis is that model weights encode this information, the challenge is to decode the underlying tree structure given the weights. We discover several properties of model weights that allow us to perform this task. By using these properties, we formulate the MoTHer Recovery task as finding a directed minimal spanning tree. In extensive experiments we demonstrate that our method successfully reconstructs complex Model Trees.
****
Savva Victorovich Ignatyev,Nina Konovalova,Daniil Selikhanovych,Oleg Voynov,Nikolay Patakin,Ilya Olkov,Dmitry Senushkin,Alexey Artemov,Anton Konushin,Alexander Filippov,Peter Wonka,Evgeny Burnaev
A3D: Does Diffusion Dream about 3D Alignment?
We tackle the problem of text-driven 3D generation from a geometry alignment perspective. Given a set of text prompts, we aim to generate a collection of objects with semantically corresponding parts aligned across them. Recent methods based on Score Distillation have succeeded in distilling the knowledge from 2D diffusion models to high-quality representations of the 3D objects. These methods handle multiple text queries separately, and therefore the resulting objects have a high variability in object pose and structure. However, in some applications, such as 3D asset design, it may be desirable to obtain a set of objects aligned with each other. In order to achieve the alignment of the corresponding parts of the generated objects, we propose to embed these objects into a common latent space and optimize the continuous transitions between these objects. We enforce two kinds of properties of these transitions: smoothness of the transition and plausibility of the intermediate objects along the transition. We demonstrate that both of these properties are essential for good alignment. We provide several practical scenarios that benefit from alignment between the objects, including 3D editing and object hybridization, and experimentally demonstrate the effectiveness of our method.
****
Haian Jin,Hanwen Jiang,Hao Tan,Kai Zhang,Sai Bi,Tianyuan Zhang,Fujun Luan,Noah Snavely,Zexiang Xu
LVSM: A Large View Synthesis Model with Minimal 3D Inductive Bias
We propose the Large View Synthesis Model (LVSM), a novel transformer-based approach for scalable and generalizable novel view synthesis from sparse-view inputs. We introduce two architectures: (1) an encoder-decoder LVSM, which encodes input image tokens into a fixed number of 1D latent tokens, functioning as a fully learned scene representation, and decodes novel-view images from them; and (2) a decoder-only LVSM, which directly maps input images to novel-view outputs, completely eliminating intermediate scene representations. Both models bypass the 3D inductive biases used in previous methods---from 3D representations (e.g., NeRF, 3DGS) to network designs (e.g., epipolar projections, plane sweeps)---addressing novel view synthesis with a fully data-driven approach. While the encoder-decoder model offers faster inference due to its independent latent representation, the decoder-only LVSM achieves superior quality, scalability, and zero-shot generalization, outperforming previous state-of-the-art methods by 1.5 to 3.5 dB PSNR. Comprehensive evaluations across multiple datasets demonstrate that both LVSM variants achieve state-of-the-art novel view synthesis quality, delivering superior performance even with reduced computational resources (1-2 GPUs). Please see our anonymous website for more details: https://haian-jin.github.io/projects/LVSM/
****
Markus Mueller,Kathrin Gruber,Dennis Fok
Continuous Diffusion for Mixed-Type Tabular Data
Score-based generative models, commonly referred to as diffusion models, have pr

oven to be successful at generating text and image data.
However, their adaptation to mixed-type tabular data remains underexplored. In t
his work, we propose CDTD, a Continuous Diffusion model for mixed-type Tabular D
ata. CDTD is based on a novel combination of score matching and score interpolat
ion to enforce a unified continuous noise distribution for both continuous and c
ategorical features. We explicitly acknowledge the necessity of homogenizing dis
tinct data types by relying on model-specific loss calibration and initializatio
n schemes. To further address the high heterogeneity in mixed-type tabular data,
 we introduce adaptive feature- or type-specific noise schedules. These ensure b
alanced generative performance across features and optimize the allocation of mo
del capacity across features and diffusion time. Our experimental results show t
hat CDTD consistently outperforms state-of-the-art benchmark models, captures fe
ature correlations exceptionally well, and that heterogeneity in the noise sched
ule design boosts sample quality. Replication code is available at https://githu
b.com/muellermarkus/cdtd.
****
Yichao Liang,Nishanth Kumar,Hao Tang,Adrian Weller,Joshua B. Tenenbaum,Tom Silve
r,Joao F. Henriques,Kevin Ellis
VisualPredicator: Learning Abstract World Models with Neuro-Symbolic Predicates
for Robot Planning
Broadly intelligent agents should form task-specific abstractions that selective
ly expose the essential elements of a task, while abstracting away the complexit
y of the raw sensorimotor space. In this work, we present Neuro-Symbolic Predica
tes, a first-order abstraction language that combines the strengths of symbolic
and neural knowledge representations. We outline an online algorithm for inventi
ng such predicates and learning abstract world models. We compare our approach t
o hierarchical reinforcement learning, vision-language model planning, and symbo
lic predicate invention approaches, on both in- and out-of-distribution tasks ac
ross five simulated robotic domains. Results show that our approach offers bette
r sample complexity, stronger out-of-distribution generalization, and improved i
nterpretability.
****
Haozhe Ma,Zhengding Luo,Thanh Vinh Vo,Kuankuan Sima,Tze-Yun Leong
Highly Efficient Self-Adaptive Reward Shaping for Reinforcement Learning
Reward shaping is a reinforcement learning technique that addresses the sparse-r
eward problem by providing frequent, informative feedback. We propose an efficie
nt self-adaptive reward-shaping mechanism that uses success rates derived from h
istorical experiences as shaped rewards. The success rates are sampled from Beta
 distributions, which evolve from uncertainty to reliability as data accumulates
. Initially, shaped rewards are stochastic to encourage exploration, gradually b
ecoming more certain to promote exploitation and maintain a natural balance betw
een exploration and exploitation. We apply Kernel Density Estimation (KDE) with
Random Fourier Features (RFF) to derive Beta distributions, providing a computat
ionally efficient solution for continuous and high-dimensional state spaces. Our
 method, validated on tasks with extremely sparse rewards, improves sample effic
iency and convergence stability over relevant baselines.
****
Tianyu Liu,Yun Li,Qitan Lv,Kai Liu,Jianchen Zhu,Winston Hu,Xiao Sun
PEARL: Parallel Speculative Decoding with Adaptive Draft Length
Speculative decoding (SD), where an extra draft model is employed to provide mul
tiple **draft** tokens first and then the original target model verifies these t
okens in parallel, has shown great power for LLM inference acceleration.
However, existing SD methods suffer from the mutual waiting problem, i.e., the t
arget model gets stuck when the draft model is *guessing* tokens, and vice versa
. This problem is directly incurred by the asynchronous execution of the draft m
odel and the target model, and is exacerbated due to the fixed draft length in s
peculative decoding.
To address these challenges, we propose a conceptually simple, flexible, and gen
eral framework to boost speculative decoding, namely
Parallel sp**E**culative decoding with **A**daptive d**R**aft **L**ength (PE

ARL).
Specifically, PEARL proposes *pre-verify* to verify the first draft token in adv
ance during the drafting phase, and *post-verify* to generate more draft tokens
during the verification phase.
PEARL parallels the drafting phase and the verification phase via applying the t
wo strategies, and achieves adaptive draft length for different scenarios, which
effectively alleviates the mutual waiting problem.
Experiments on various text generation benchmarks demonstrate the effectiveness
of our PEARL, leading to a superior speedup performance up to **4.43$\times$** a
nd **1.50$\times$**, compared to auto-regressive decoding and vanilla speculativ
e decoding, respectively.
****
Haoru Wang,Wentao Zhu,Luyi Miao,Yishu Xu,Feng Gao,Qi Tian,Yizhou Wang
Aligning Human Motion Generation with Human Perceptions
Human motion generation is a critical task with a wide spectrum of applications.
Achieving high realism in generated motions requires naturalness, smoothness, a
nd plausibility. However, current evaluation metrics often rely on simple heuris
tics or distribution distances and do not align well with human perceptions. In
this work, we propose a data-driven approach to bridge this gap by introducing a
large-scale human perceptual evaluation dataset, MotionPercept, and a human mot
ion critic model, MotionCritic, that capture human perceptual preferences. Our c
ritic model offers a more accurate metric for assessing motion quality and could
be readily integrated into the motion generation pipeline to enhance generation
quality. Extensive experiments demonstrate the effectiveness of our approach in
both evaluating and improving the quality of generated human motions by alignin
g with human perceptions. Code and data are publicly available at https://motion
critic.github.io/.
****
Emanuel Sommer,Jakob Robnik,Giorgi Nozadze,Uros Seljak,David Rügamer
Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Net
works
Despite recent advances, sampling-based inference for Bayesian Neural Networks (
BNNs) remains a significant challenge in probabilistic deep learning. While samp
ling-based approaches do not require a variational distribution assumption, curr
ent state-of-the-art samplers still struggle to navigate the complex and highly
multimodal posteriors of BNNs. As a consequence, sampling still requires conside
rably longer inference times than non-Bayesian methods even for small neural net
works, despite recent advances in making software implementations more efficient
. Besides the difficulty of finding high-probability regions, the time until sam
plers provide sufficient exploration of these areas remains unpredictable. To ta
ckle these challenges, we introduce an ensembling approach that leverages strate
gies from optimization and a recently proposed sampler called Microcanonical Lan
gevin Monte Carlo (MCLMC) for efficient, robust and predictable sampling perform
ance. Compared to approaches based on the state-of-the-art No-U-Turn Sampler, ou
r approach delivers substantial speedups up to an order of magnitude, while main
taining or improving predictive performance and uncertainty quantification acros
s diverse tasks and data modalities. The suggested Microcanonical Langevin Ensem
bles and modifications to MCLMC additionally enhance the method's predictability
in resource requirements, facilitating easier parallelization. All in all, the
proposed method offers a promising direction for practical, scalable inference f
or BNNs.
****
Changle Qu,Sunhao Dai,Xiaochi Wei,Hengyi Cai,Shuaiqiang Wang,Dawei Yin,Jun Xu,Ji
-Rong Wen
From Exploration to Mastery: Enabling LLMs to Master Tools via Self-Driven Inter
actions
Tool learning enables Large Language Models (LLMs) to interact with external env
ironments by invoking tools, serving as an effective strategy to mitigate the li
mitations inherent in their pre-training data. In this process, tool documentati
on plays a crucial role by providing usage instructions for LLMs, thereby facili

tating effective tool utilization. This paper concentrates on the critical chall
enge of bridging the comprehension gap between LLMs and external tools due to th
e inadequacies and inaccuracies inherent in existing human-centric tool document
ation. We propose a novel framework, DRAFT, aimed at Dynamically Refining tool d
ocumentation through the Analysis of Feedback and Trials emanating from LLMs' in
teractions with external tools. This methodology pivots on an innovative trial-a
nd-error approach, consisting of three distinct learning phases: experience gath
ering, learning from experience, and documentation rewriting, to iteratively enh
ance the tool documentation. This process is further optimized by implementing a
 diversity-promoting exploration strategy to ensure explorative diversity and a
tool-adaptive termination mechanism to prevent overfitting while enhancing effic
iency. Extensive experiments on multiple datasets demonstrate that DRAFT's itera
tive, feedback-based refinement significantly ameliorates documentation quality,
 fostering a deeper comprehension and more effective utilization of tools by LLM
s. Notably, our analysis reveals that the tool documentation refined via our app
roach demonstrates robust cross-model generalization capabilities.
****

Shreelekha Revankar,Cheng Perng Phoo,Utkarsh Mall,Bharath Hariharan,Kavita Bala
Scale-aware Recognition in Satellite Images under Resource Constraints
Recognition of features in satellite imagery (forests, swimming pools, etc.) dep
ends strongly on the spatial scale of the concept and therefore the resolution o
f the images. This poses two challenges:
Which resolution is best suited for recognizing a given concept, and where and w
hen should the costlier higher-resolution (HR) imagery be acquired?
We present a novel scheme to address these challenges by introducing three compo
nents: (1) A technique to distill knowledge from models trained on HR imagery t
o recognition models that operate on imagery of lower resolution (LR), (2) a sam
pling strategy for HR imagery based on model disagreement, and (3) an LLM-based
approach for inferring concept "scale". With these components we present a syste
m to efficiently perform scale-aware recognition in satellite imagery, improving
 accuracy over single-scale inference while following budget constraints. **Our
novel approach offers up to a 26.3\% improvement over entirely HR baselines, usi
ng 76.3 \% fewer HR images.** Resources are available at https://www.cs.cornell.
edu/~revankar/scale_aware.
****

Canyu Zhao,Mingyu Liu,Wen Wang,Weihua Chen,Fan Wang,Hao Chen,Bo Zhang,Chunhua Sh
en
MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequences
Recent advancements in video generation have primarily leveraged diffusion model
s for short-duration content. However, these approaches often fall short in mode
ling complex narratives and maintaining character consistency over extended peri
ods, which is essential for long-form video production like movies. We propose M
ovieDreamer, a novel hierarchical framework that integrates the strengths of aut
oregressive models with diffusion-based rendering to pioneer long-duration video
 generation with intricate plot progressions and high visual fidelity. Our appro
ach utilizes autoregressive models for global narrative coherence, predicting se
quences of visual tokens that are subsequently transformed into high-quality vid
eo frames through diffusion rendering. This method is akin to traditional movie
production processes, where complex stories are factorized down into manageable
scene capturing. Further, we employ a multimodal script that enriches scene desc
riptions with detailed character information and visual style, enhancing continu
ity and character identity across scenes. We present extensive experiments acros
s various movie genres, demonstrating that our approach not only achieves superi
or visual and narrative quality but also effectively extends the duration of gen
erated content significantly beyond current capabilities.
****

Dennis Frauen,Konstantin Hess,Stefan Feuerriegel
Model-agnostic meta-learners for estimating heterogeneous treatment effects over
 time
Estimating heterogeneous treatment effects (HTEs) over time is crucial in many d

isciplines such as personalized medicine. Existing works for this task have most ly focused on *model-based* learners that adapt specific machine-learning models and adjustment mechanisms. In contrast, model-agnostic learners - so-called *me ta-learners* - are largely unexplored. In our paper, we propose several meta-lea rners that are model-agnostic and thus can be used in combination with arbitrary machine learning models (e.g., transformers) to estimate HTEs over time. We the n provide a comprehensive theoretical analysis that characterizes the different learners and that allows us to offer insights into when specific learners are pr eferable. Furthermore, we propose a novel IVW-DR-learner that (i) uses a doubly robust (DR) and orthogonal loss; and (ii) leverages inverse-variance weights (IV Ws) that we derive to stabilize the DR-loss over time. Our IVWs downweight extre me trajectories due to *products* of inverse-propensities in the DR-loss, result ing in a lower estimation variance. Our IVW-DR-learner achieves superior perform ance in our experiments, particularly in regimes with low overlap and long time horizons.
****
Yankai Jiang,Wenhui Lei,Xiaofan Zhang,Shaoting Zhang
Unleashing the Potential of Vision-Language Pre-Training for 3D Zero-Shot Lesion Segmentation via Mask-Attribute Alignment
Recent advancements in medical vision-language pre-training models have driven s ignificant progress in zero-shot disease recognition. However, transferring imag e-level knowledge to pixel-level tasks, such as lesion segmentation in 3D CT sca ns, remains a critical challenge. Due to the complexity and variability of patho logical visual characteristics, existing methods struggle to align fine-grained lesion features not encountered during training with disease-related textual rep resentations. In this paper, we present Malenia, a novel multi-scale lesion-leve l mask-attribute alignment framework, specifically designed for 3D zero-shot les ion segmentation. Malenia improves the compatibility between mask representation s and their associated elemental attributes, explicitly linking the visual featu res of unseen lesions with the extensible knowledge learned from previously seen ones. Furthermore, we design a Cross-Modal Knowledge Injection module to enhanc e both visual and textual features with mutually beneficial information, effecti vely guiding the generation of segmentation results. Comprehensive experiments a cross three datasets and 12 lesion categories validate the superior performance of Malenia.
****
Naoki Nishikawa,Taiji Suzuki
State Space Models are Provably Comparable to Transformers in Dynamic Token Sele ction
Deep neural networks based on state space models (SSMs) are attracting significa nt attention in sequence modeling since their computational cost is much smaller than that of Transformers. While the capabilities of SSMs have been demonstrate d through experiments in various tasks, theoretical understanding of SSMs is sti ll limited. In particular, most theoretical studies discuss the capabilities of SSM layers without nonlinear layers, and there is a lack of discussion on their combination with nonlinear layers. In this paper, we explore the capabilities of SSMs combined with fully connected neural networks, and show that they are comp arable to Transformers in extracting the essential tokens depending on the input . As concrete examples, we consider two synthetic tasks, which are challenging f or a single SSM layer, and demonstrate that SSMs combined with nonlinear layers can efficiently solve these tasks. Furthermore, we study the nonparametric regr ession task, and prove that the ability of SSMs is equivalent to that of Transfo rmers in estimating functions belonging to a certain class.
****
Juan Agustin Duque,Milad Aghajohari,Tim Cooijmans,razvan ciuca,Tianyu Zhang,Gaut hier Gidel,Aaron Courville
Advantage Alignment Algorithms
Artificially intelligent agents are increasingly being integrated into human dec ision-making: from large language model (LLM) assistants to autonomous vehicles. These systems often optimize their individual objective, leading to conflicts,

particularly in general-sum games where naive reinforcement learning agents empi
rically converge to Pareto-suboptimal Nash equilibria. To address this issue, op
ponent shaping has emerged as a paradigm for finding socially beneficial equilib
ria in general-sum games. In this work, we introduce Advantage Alignment, a fami
ly of algorithms derived from first principles that perform opponent shaping eff
iciently and intuitively. We achieve this by aligning the advantages of interact
ing agents, increasing the probability of mutually beneficial actions when their
 interaction has been positive. We prove that existing opponent shaping methods
implicitly perform Advantage Alignment. Compared to these methods, Advantage Ali
gnment simplifies the mathematical formulation of opponent shaping, reduces the
computational burden and extends to continuous action domains. We demonstrate th
e effectiveness of our algorithms across a range of social dilemmas, achieving s
tate-of-the-art cooperation and robustness against exploitation.
****
Yantao Liu,Zijun Yao,Rui Min,Yixin Cao,Lei Hou,Juanzi Li
RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style
Reward models are critical in techniques like Reinforcement Learning from Human
Feedback (RLHF) and Inference Scaling Laws, where they guide language model alig
nment and select optimal responses.
Despite their importance, existing reward model benchmarks often evaluate models
 by asking them to distinguish between responses generated by models of varying
power.
However, this approach fails to assess reward models on subtle but critical cont
ent changes and variations in style, resulting in a low correlation with policy
model performance.
To this end, we introduce RM-Bench, a novel benchmark designed to evaluate rewar
d models based on their sensitivity to subtle content differences and resistance
 to style biases.
Extensive experiments demonstrate that RM-Bench strongly correlates with policy
model performance, making it a reliable reference for selecting reward models to
 align language models effectively.
We evaluate nearly 40 reward models on RM-Bench.
Our results reveal that even state-of-the-art models achieve an average performa
nce of only 46.6%, which falls short of random-level accuracy (50%) when faced w
ith style bias interference.
These findings highlight the significant room for improvement in current reward
models.
****
Xiang Li,Kai Qiu,Hao Chen,Jason Kuen,Jiuxiang Gu,Bhiksha Raj,Zhe Lin
ImageFolder: Autoregressive Image Generation with Folded Tokens
Image tokenizers are crucial for visual generative models, \eg, diffusion models
 (DMs) and autoregressive (AR) models, as they construct the latent representati
on for modeling. Increasing token length is a common approach to improve image r
econstruction quality. However, tokenizers with longer token lengths are not gua
ranteed to achieve better generation quality. There exists a trade-off between r
econstruction and generation quality regarding token length. In this paper, we i
nvestigate the impact of token length on both image reconstruction and generatio
n and provide a flexible solution to the tradeoff. We propose \textbf{ImageFolde
r}, a semantic tokenizer that provides spatially aligned image tokens that can b
e folded during autoregressive modeling to improve both efficiency and quality.
To enhance the representative capability without increasing token length, we lev
erage dual-branch product quantization to capture different contexts of images.
Specifically, semantic regularization is introduced in one branch to encourage c
ompacted semantic information while another branch is designed to capture pixel-
level details. Extensive experiments demonstrate the superior quality of image g
eneration and shorter token length with ImageFolder tokenizer.
****
Irena Gao,Percy Liang,Carlos Guestrin
Model Equality Testing: Which Model is this API Serving?
Users often interact with large language models through black-box inference APIs

, both for closed- and open-weight models (e.g., Llama models are popularly acce
ssed via Amazon Bedrock and Azure AI Studio). In order to cut costs or add funct
ionality, API providers may quantize, watermark, or finetune the underlying mode
l, changing the output distribution --- possibly without notifying users. We for
malize detecting such distortions as Model Equality Testing, a two-sample testin
g problem, where the user collects samples from the API and a reference distribu
tion and conducts a statistical test to see if the two distributions are the sam
e. We find that tests based on the Maximum Mean Discrepancy between distribution
s are powerful for this task: a test built on a simple string kernel achieves a
median of 77.4% power against a range of distortions, using an average of just 1
0 samples per prompt. We then apply this test to commercial inference APIs from
Summer 2024 for four Llama models, finding that 11 out of 31 endpoints serve dif
ferent distributions than reference weights released by Meta.
****
Diaaeldin Taha,James Chapman,Marzieh Eidi,Karel Devriendt,Guido Montufar
Demystifying Topological Message-Passing with Relational Structures: A Case Stud
y on Oversquashing in Simplicial Message-Passing
Topological deep learning (TDL) has emerged as a powerful tool for modeling high
er-order interactions in relational data. However, phenomena such as oversquashi
ng in topological message-passing remain understudied and lack theoretical analy
sis. We propose a unifying axiomatic framework that bridges graph and topologica
l message-passing by viewing simplicial and cellular complexes and their message
-passing schemes through the lens of relational structures. This approach extend
s graph-theoretic results and algorithms to higher-order structures, facilitatin
g the analysis and mitigation of oversquashing in topological message-passing ne
tworks. Through theoretical analysis and empirical studies on simplicial network
s, we demonstrate the potential of this framework to advance TDL.
****
Kaustubh Vyas,Damien Graux,Yijun Yang,Sebastien Montella,Chenxin Diao,Wendi Zhou
,Pavlos Vougiouklis,Ruofei Lai,Yang Ren,Keshuang Li,Jeff Z. Pan
From an LLM Swarm to a PDDL-empowered Hive: Planning Self-executed Instructions
in a Multi-modal Jungle
In response to the call for agent-based solutions that leverage the ever-increas
ing capabilities of the deep models' ecosystem, we introduce a comprehensive sol
ution for selecting appropriate models and subsequently planning a set of atomic
 actions to satisfy the end-users' instructions.

Our system, Hive, operates over sets of models and, upon receiving natural langu
age instructions, schedules and executes, explainable plans of atomic actions. T
hese actions can involve one or more of the available models to achieve the over
all task, while respecting end-users specific constraints. Hive is able to plan
complex chains of actions while guaranteeing explainability, using an LLM-based
formal logic backbone empowered by PDDL operations. We introduce the MuSE benchm
ark in order to offer a comprehensive evaluation of the multi-modal capabilities
 of agent systems. Our findings show that our framework redefines the state-of-t
he-art for task selection, outperforming other competing systems that plan opera
tions across multiple models while offering transparency guarantees while fully
adhering to user constraints.
****
Frederik Pahde,Maximilian Dreyer,Moritz Weckbecker,Leander Weber,Christopher J.
Anders,Thomas Wiegand,Wojciech Samek,Sebastian Lapuschkin
Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Direc
tional Divergence
With a growing interest in understanding neural network prediction strategies, C
oncept Activation Vectors (CAVs) have emerged as a popular tool for modeling hum
an-understandable concepts in the latent space.
Commonly, CAVs are computed by leveraging linear classifiers optimizing the *sep
arability* of latent representations of samples with and without a given concept
. However, in this paper we show that such a separability-oriented computation l
eads to solutions, which may diverge from the actual goal of precisely modeling

the concept direction.
This discrepancy can be attributed to the significant influence of distractor directions, i.e., signals unrelated to the concept, which are picked up by filters (i.e., weights) of linear models to optimize class-separability.
To address this, we introduce *pattern-based CAVs*, solely focussing on concept signals, thereby providing more accurate concept directions.
We evaluate various CAV methods in terms of their alignment with the true concept direction and their impact on CAV applications, including concept sensitivity testing and model correction for shortcut behavior caused by data artifacts.
We demonstrate the benefits of pattern-based CAVs using the Pediatric Bone Age, ISIC2019, and FunnyBirds datasets with VGG, ResNet, ReXNet, EfficientNet, and Vision Transformer as model architectures.
****
Varshith Sreeramdass,Rohan R Paleja,Letian Chen,Sanne van Waveren,Matthew Gombolay
Generalized Behavior Learning from Diverse Demonstrations
Diverse behavior policies are valuable in domains requiring quick test-time adaptation or personalized human-robot interaction. Human demonstrations provide rich information regarding task objectives and factors that govern individual behavior variations, which can be used to characterize \textit{useful} diversity and learn diverse performant policies.
However, we show that prior work that builds naive representations of demonstration heterogeneity fails in generating successful novel behaviors that generalize over behavior factors.
We propose Guided Strategy Discovery (GSD), which introduces a novel diversity formulation based on a learned task-relevance measure that prioritizes behaviors exploring modeled latent factors.
We empirically validate across three continuous control benchmarks for generalizing to in-distribution (interpolation) and out-of-distribution (extrapolation) factors that GSD outperforms baselines in novel behavior discovery by $\sim$21\%.
Finally, we demonstrate that GSD can generalize striking behaviors for table tennis in a virtual testbed while leveraging human demonstrations collected in the real world.
Code is available at https://github.com/CORE-Robotics-Lab/GSD.
****
Wei Chow,Jiageng Mao,Boyi Li,Daniel Seita,Vitor Campagnolo Guizilini,Yue Wang
PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding
Understanding the physical world is a fundamental challenge in embodied AI, critical for enabling agents to perform complex tasks and operate safely in real-world environments. While Vision-Language Models (VLMs) have shown great promise in reasoning and task planning for embodied agents, their ability to comprehend physical phenomena remains extremely limited.
To close this gap, we introduce PhysBench, a comprehensive benchmark designed to evaluate VLMs' physical world understanding capability across a diverse set of tasks.
PhysBench contains 10,002 entries of interleaved video-image-text data, categorized into four major domains: physical object properties, physical object relationships, physical scene understanding, and physics-based dynamics, further divided into 19 subclasses and 8 distinct capability dimensions.
Our extensive experiments, conducted on 75 representative VLMs, reveal that while these models excel in common-sense reasoning, they struggle with understanding the physical world---likely due to the absence of physical knowledge in their training data and the lack of embedded physical priors.
To tackle the shortfall, we introduce PhysAgent, a novel framework that combines the generalization strengths of VLMs with the specialized expertise of vision models, significantly enhancing VLMs' physical understanding across a variety of tasks, including an 18.4\% improvement on GPT-4o.
Furthermore, our results demonstrate that enhancing VLMs' physical world understanding capabilities can help embodied agents such as MOKA.

We believe that PhysBench and PhysAgent offer valuable insights and contribute to bridging the gap between VLMs and physical world understanding. [Project Page is here](https://physbench.github.io/)
****
Hiroyuki Deguchi,Go Kamoda,Yusuke Matsushita,Chihiro Taguchi,Kohei Suenaga,Masaki Waga,Sho Yokoi
SoftMatcha: A Soft and Fast Pattern Matcher for Billion-Scale Corpus Searches
Researchers and practitioners in natural language processing and computational linguistics frequently observe and analyze the real language usage in large-scale corpora.
For that purpose, they often employ off-the-shelf pattern-matching tools, such as grep, and keyword-in-context concordancers, which is widely used in corpus linguistics for gathering examples.
Nonetheless, these existing techniques rely on surface-level string matching, and thus they suffer from the major limitation of not being able to handle orthographic variations and paraphrasing---notable and common phenomena in any natural language.
In addition, existing continuous approaches such as dense vector search tend to be overly coarse, often retrieving texts that are unrelated but share similar topics.
Given these challenges, we propose a novel algorithm that achieves soft (or semantic) yet efficient pattern matching by relaxing a surface-level matching with word embeddings.
Our algorithm is highly scalable with respect to the size of the corpus text utilizing inverted indexes.
We have prepared an efficient implementation, and we provide an accessible web tool.
Our experiments demonstrate that the proposed method
(i) can execute searches on billion-scale corpora in less than a second, which is comparable in speed to surface-level string matching and dense vector search;
(ii) can extract harmful instances that semantically match queries from a large set of English and Japanese Wikipedia articles;
and (iii) can be effectively applied to corpus-linguistic analyses of Latin, a language with highly diverse inflections.
****
Gaurav Patel,Christopher Michael Sandino,Behrooz Mahasseni,Ellen L. Zippi,Erdrin Azemi,Ali Moin,Juri Minxha
Efficient Source-Free Time-Series Adaptation via Parameter Subspace Disentanglement
In this paper, we propose a framework for efficient Source-Free Domain Adaptation (SFDA) in the context of time-series, focusing on enhancing both parameter efficiency and data-sample utilization. Our approach introduces an improved paradigm for source-model preparation and target-side adaptation, aiming to enhance training efficiency during target adaptation. Specifically, we reparameterize the source model's weights in a Tucker-style decomposed manner, factorizing the model into a compact form during the source model preparation phase. During target-side adaptation, only a subset of these decomposed factors is fine-tuned, leading to significant improvements in training efficiency. We demonstrate using PAC Bayesian analysis that this selective fine-tuning strategy implicitly regularizes the adaptation process by constraining the model's learning capacity. Furthermore, this re-parameterization reduces the overall model size and enhances inference efficiency, making the approach particularly well suited for resource-constrained devices. Additionally, we demonstrate that our framework is compatible with various SFDA methods and achieves significant computational efficiency, reducing the number of fine-tuned parameters and inference overhead in terms of MACs by over 90\% while maintaining model performance.
****
Haining Yu,Yizhou Sun
Do Contemporary Causal Inference Models Capture Real-World Heterogeneity? Findings from a Large-Scale Benchmark

We present unexpected findings from a large-scale benchmark study evaluating Conditional Average Treatment Effect (CATE) estimation algorithms. By running 16 modern CATE models across 43,200 datasets, we find that: (a) 62\% of CATE estimates have a higher Mean Squared Error (MSE) than a trivial zero-effect predictor, rendering them ineffective; (b) in datasets with at least one useful CATE estimate,
80\% still have higher MSE than a constant-effect model; and (c) Orthogonality-based models outperform other models only 30\% of the time, despite widespread optimism about their performance. These findings expose significant limitations in current CATE models and suggest ample opportunities for further research.

Our findings stem from a novel application of \textit{observational sampling}, originally developed to evaluate Average Treatment Effect (ATE) estimates from observational methods with experiment data. To adapt observational sampling for CATE evaluation, we introduce a statistical parameter, Q, equal to MSE minus a constant and preserves the ranking of models by their MSE. We then derive a family of sample statistics, collectively called $\hat{Q}$, that can be computed from real-world data. We prove that $\hat{Q}$ is a consistent estimator of Q under mild technical conditions. When used in observational sampling, $\hat{Q}$ is unbiased and asymptotically selects the model with the smallest MSE. To ensure the benchmark reflects real-world heterogeneity, we handpick datasets where outcomes come from field rather than simulation. By combining the new observational sampling method, new statistics, and real-world datasets, the benchmark provides a unique perspective on CATE estimator performance and uncover gaps in capturing real-world heterogeneity.
****
Megh Shukla,Aziz Shameem,Mathieu Salzmann,Alexandre Alahi
Towards Self-Supervised Covariance Estimation in Deep Heteroscedastic Regression
Deep heteroscedastic regression models the mean and covariance of the target distribution through neural networks. The challenge arises from heteroscedasticity, which implies that the covariance is sample dependent and is often unknown. Consequently, recent methods learn the covariance through unsupervised frameworks, which unfortunately yield a trade-off between computational complexity and accuracy. While this trade-off could be alleviated through supervision, obtaining labels for the covariance is non-trivial.
Here, we study self-supervised covariance estimation in deep heteroscedastic regression. We address two questions: (1) How should we supervise the covariance assuming ground truth is available? (2) How can we obtain pseudo labels in the absence of the ground-truth? We address (1) by analysing two popular measures: the KL Divergence and the 2-Wasserstein distance. Subsequently, we derive an upper bound on the 2-Wasserstein distance between normal distributions with non-commutative covariances that is stable to optimize. We address (2) through a simple neighborhood based heuristic algorithm which results in surprisingly effective pseudo labels for the covariance. Our experiments over a wide range of synthetic and real datasets demonstrate that the proposed 2-Wasserstein bound coupled with pseudo label annotations results in a computationally cheaper yet accurate deep heteroscedastic regression.
****
Denis Blessing,Julius Berner,Lorenz Richter,Gerhard Neumann
Underdamped Diffusion Bridges with Applications to Sampling
We provide a general framework for learning diffusion bridges that transport prior to target distributions. It includes existing diffusion models for generative modeling, but also underdamped versions with degenerate diffusion matrices, where the noise only acts in certain dimensions. Extending previous findings, our framework allows to rigorously show that score-matching in the underdamped case is indeed equivalent to maximizing a lower bound on the likelihood. Motivated by superior convergence properties and compatibility with sophisticated numerical integration schemes of underdamped stochastic processes, we propose *underdamped diffusion bridges*, where a general density evolution is learned rather than prescribed by a fixed noising process. We apply our method to the challenging task

of sampling from unnormalized densities without access to samples from the targe
t distribution. Across a diverse range of sampling problems, our approach demons
trates state-of-the-art performance, notably outperforming alternative methods,
while requiring significantly fewer discretization steps and almost no hyperpara
meter tuning.
****

Xiaojian Yuan,Tianyu Pang,Chao Du,Kejiang Chen,Weiming Zhang,Min Lin
A Closer Look at Machine Unlearning for Large Language Models
Large language models (LLMs) may memorize sensitive or copyrighted content, rais
ing privacy and legal concerns. Due to the high cost of retraining from scratch,
 researchers attempt to employ machine unlearning to remove specific content fro
m LLMs while preserving the overall performance. In this paper, we discuss sever
al issues in machine unlearning for LLMs and provide our insights on possible ap
proaches. To address the issue of inadequate evaluation of model outputs after u
nlearning, we introduce three additional metrics to evaluate token diversity, se
ntence semantics, and factual correctness. We then categorize unlearning methods
 into untargeted and targeted, and discuss their issues respectively. Specifical
ly, the behavior that untargeted unlearning attempts to approximate is unpredict
able and may involve hallucinations, and existing regularization is insufficient
 for targeted unlearning. To alleviate these issues, we propose using the object
ive of maximizing entropy (ME) for untargeted unlearning and incorporate answer
preservation (AP) loss as regularization for targeted unlearning. Experimental r
esults across three scenarios, i.e., fictitious unlearning, continual unlearning
, and real-world unlearning, demonstrate the effectiveness of our approaches. Th
e code is available at https://github.com/sail-sg/closer-look-LLM-unlearning.
****

Haijin Zeng,Benteng Sun,Yongyong Chen,Jingyong Su,Yong Xu
Spectral Compressive Imaging via Unmixing-driven Subspace Diffusion Refinement
Spectral Compressive Imaging (SCI) reconstruction is inherently ill-posed, offer
ing multiple plausible solutions from a single observation. Traditional determin
istic methods typically struggle to effectively recover high-frequency details.
Although diffusion models offer promising solutions to this challenge, their app
lication is constrained by the limited training data and high computational dema
nds associated with multispectral images (MSIs), complicating direct training. T
o address these issues, we propose a novel Predict-and-unmixing-driven-Subspace-
Refine framework (PSR-SCI). This framework begins with a cost-effective predicto
r that produces an initial, rough estimate of the MSI. Subsequently, we introduc
e a unmixing-driven reversible spectral embedding module that decomposes the MSI
 into subspace images and spectral coefficients. This decomposition facilitates
the adaptation of pre-trained RGB diffusion models and focuses refinement proces
ses on high-frequency details, thereby enabling efficient diffusion generation w
ith minimal MSI data. Additionally, we design a high-dimensional guidance mechan
ism with imaging consistency to enhance the model's efficacy. The refined subspa
ce image is then reconstructed back into an MSI using the reversible embedding,
yielding the final MSI with full spectral resolution. Experimental results on th
e standard KAIST and zero-shot datasets NTIRE, ICVL, and Harvard show that PSR-S
CI enhances visual quality and delivers PSNR and SSIM metrics comparable to exis
ting diffusion, transformer, and deep unfolding techniques. This framework provi
des a robust alternative to traditional deterministic SCI reconstruction methods
. Code and models are available at [https://github.com/SMARK2022/PSR-SCI](https:
//github.com/SMARK2022/PSR-SCI).
****

Laurin Lux,Alexander H Berger,Alexander Weers,Nico Stucki,Daniel Rueckert,Ulrich
 Bauer,Johannes C. Paetzold
Topograph: An Efficient Graph-Based Framework for Strictly Topology Preserving I
mage Segmentation
Topological correctness plays a critical role in many image segmentation tasks,
yet most networks are trained using pixel-wise loss functions, such as Dice, neg
lecting topological accuracy. Existing topology-aware methods often lack robust
topological guarantees, are limited to specific use cases, or impose high comput

ational costs.
In this work, we propose a novel, graph-based framework for topologically accura te image segmentation that is both computationally efficient and generally appli cable. Our method constructs a component graph that fully encodes the topologica l information of both the prediction and ground truth, allowing us to efficientl y identify topologically critical regions and aggregate a loss based on local ne ighborhood information. Furthermore, we introduce a strict topological metric ca pturing the homotopy equivalence between the union and intersection of predictio n-label pairs. We formally prove the topological guarantees of our approach and empirically validate its effectiveness on binary and multi-class datasets, demon strating state-of-the-art performance with up to fivefold faster loss computatio n compared to persistent homology methods.
****
Yu Yao,Yang Zhou,Bo Han,Mingming Gong,Kun Zhang,Tongliang Liu
A Robust Method to Discover Causal or Anticausal Relation
Understanding whether the data generative process follows causal or anticausal r elations is important for many applications. Existing causal discovery methods s truggle with high-dimensional perceptual data such as images. Moreover, they req uire well-labeled data, which may not be feasible due to measurement error. In t his paper, we propose a robust method to detect whether the data generative proc ess is causal or anticausal. To determine the causal or anticausal relation, we identify an asymmetric property: under the causal relation, the instance distrib ution does not contain information about the noisy class-posterior distribution. We also propose a practical method to verify this via a noise injection approac h. Our method is robust to label errors and is designed to handle both large-sca le and high-dimensional datasets effectively. Both theoretical analyses and empi rical results on a variety of datasets demonstrate the effectiveness of our prop osed method in determining the causal or anticausal direction of the data genera tive process.
****
Lukas Tatzel,Bálint Mucsányi,Osane Hackel,Philipp Hennig
Debiasing Mini-Batch Quadratics for Applications in Deep Learning
Quadratic approximations form a fundamental building block of machine learning m ethods. E.g., second-order optimizers try to find the Newton step into the minim um of a local quadratic proxy to the objective function; and the second-order ap proximation of a network's loss function can be used to quantify the uncertainty of its outputs via the Laplace approximation. When computations on the entire t raining set are intractable - typical for deep learning - the relevant quantitie s are computed on mini-batches. This, however, distorts and biases the shape of the associated *stochastic* quadratic approximations in an intricate way with de trimental effects on applications. In this paper, we (i) show that this bias int roduces a systematic error, (ii) provide a theoretical explanation for it, (iii) explain its relevance for second-order optimization and uncertainty quantificat ion via the Laplace approximation in deep learning, and (iv) develop and evaluat e debiasing strategies.
****
Xiaohao Xu,Tianyi Zhang,Shibo Zhao,Xiang Li,Sibo Wang,Yongqi Chen,Ye Li,Bhiksha Raj,Matthew Johnson-Roberson,Sebastian Scherer,Xiaonan Huang
Scalable Benchmarking and Robust Learning for Noise-Free Ego-Motion and 3D Recon struction from Noisy Video
We aim to redefine robust ego-motion estimation and photorealistic 3D reconstruc tion by addressing a critical limitation: the reliance on noise-free data in exi sting models. While such sanitized conditions simplify evaluation, they fail to capture the unpredictable, noisy complexities of real-world environments. Dynami c motion, sensor imperfections, and synchronization perturbations lead to sharp performance declines when these models are deployed in practice, revealing an ur gent need for frameworks that embrace and excel under real-world noise.
To bridge this gap, we tackle three core challenges: scalable data generation, c omprehensive benchmarking, and model robustness enhancement. First, we introduce a scalable noisy data synthesis pipeline that generates diverse datasets simula

ting complex motion, sensor imperfections, and synchronization errors. Second, w
e leverage this pipeline to create Robust-Ego3D, a benchmark rigorously designed
 to expose noise-induced performance degradation, highlighting the limitations o
f current learning-based methods in ego-motion accuracy and 3D reconstruction qu
ality. Third, we propose Correspondence-guided Gaussian Splatting (CorrGS), a no
vel method that progressively refines an internal clean 3D representation by ali
gning noisy observations with rendered RGB-D frames from clean 3D map, enhancing
 geometric alignment and appearance restoration through visual correspondence.
Extensive experiments on synthetic and real-world data demonstrate that CorrGS c
onsistently outperforms prior state-of-the-art methods, particularly in scenario
s involving rapid motion and dynamic illumination. We will release our code and
benchmark to advance robust 3D vision, setting a new standard for ego-motion est
imation and high-fidelity reconstruction in noisy environments.
****
Ling Yang,Zhaochen Yu,Tianjun Zhang,Minkai Xu,Joseph E. Gonzalez,Bin CUI,Shuiche
ng YAN
SuperCorrect: Advancing Small LLM Reasoning with Thought Template Distillation a
nd Self-Correction
Large language models (LLMs) like GPT-4, DeepSeek-R1, and ReasonFlux have shown
significant improvements in various reasoning tasks. However, smaller LLMs still
 struggle with complex mathematical reasoning because they fail to effectively i
dentify and correct reasoning errors. Recent reflection-based methods aim to add
ress these issues by enabling self-reflection and self-correction, but they stil
l face challenges in independently detecting errors in their reasoning steps. To
 overcome these limitations, we propose SuperCorrect, a novel two-stage framewor
k that uses a large teacher model to supervise and correct both the reasoning an
d reflection processes of a smaller student model. In the first stage, we extrac
t hierarchical high-level and detailed thought templates from the teacher model
to guide the student model in eliciting more fine-grained reasoning thoughts. In
 the second stage, we introduce cross-model collaborative direct preference opti
mization (DPO) to enhance the self-correction abilities of the student model by
following the teacher's correction traces during training. This cross-model DPO
approach teaches the student model to effectively locate and resolve erroneous t
houghts with error-driven insights from the teacher model, breaking the bottlene
ck of its thoughts and acquiring new skills and knowledge to tackle challenging
problems. Extensive experiments consistently demonstrate our superiority over pr
evious methods. Notably, our SuperCorrect-7B model significantly surpasses power
ful DeepSeekMath-7B by 7.8\%/5.3\% and Qwen2.5-Math-7B by 15.1\%/6.3\% on MATH/G
SM8K benchmarks, achieving new SOTA performance among all 7B models. Code is ava
ilable at: https://github.com/YangLing0818/SuperCorrect-llm
****
Haotian Sun,Tao Lei,Bowen Zhang,Yanghao Li,Haoshuo Huang,Ruoming Pang,Bo Dai,Nan
 Du
EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing
Diffusion transformers have been widely adopted for text-to-image synthesis. Whi
le scaling these models up to billions of parameters shows promise, the effectiv
eness of scaling beyond current sizes remains underexplored and challenging. By
explicitly exploiting the computational heterogeneity of image generations, we d
evelop a new family of Mixture-of-Experts (MoE) models (EC-DIT) for diffusion tr
ansformers with expert-choice routing. EC-DIT learns to adaptively optimize the
compute allocated to understand the input texts and generate the respective imag
e patches, enabling heterogeneous computation aligned with varying text-image co
mplexities. This heterogeneity provides an efficient way of scaling EC-DIT up to
 97 billion parameters and achieving significant improvements in training conver
gence, text-to-image alignment, and overall generation quality over dense models
 and conventional MoE models. Through extensive ablations, we show that EC-DIT d
emonstrates superior scalability and adaptive compute allocation by recognizing
varying textual importance through end-to-end training. Notably, in text-to-imag
e alignment evaluation, our largest models achieve a state-of-the-art GenEval sc
ore of 71.68% and still maintain competitive inference speed with intuitive inte

rpretability.
****
Ruikun Li,Huandong Wang,Qingmin Liao,Yong Li
Predicting the Energy Landscape of Stochastic Dynamical System via Physics-informed Self-supervised Learning
Energy landscapes play a crucial role in shaping dynamics of many real-world complex systems. System evolution is often modeled as particles moving on a landscape under the combined effect of energy-driven drift and noise-induced diffusion, where the energy governs the long-term motion of the particles.
Estimating the energy landscape of a system has been a longstanding interdisciplinary challenge, hindered by the high operational costs or the difficulty of obtaining supervisory signals. Therefore, the question of how to infer the energy landscape in the absence of true energy values is critical. In this paper, we propose a physics-informed self-supervised learning method to learn the energy landscape from the evolution trajectories of the system. It first maps the system state from the observation space to a discrete landscape space by an adaptive codebook, and then explicitly integrates energy into the graph neural Fokker-Planck equation, enabling the joint learning of energy estimation and evolution prediction. Experimental results across interdisciplinary systems demonstrate that our estimated energy has a correlation coefficient above 0.9 with the ground truth, and evolution prediction accuracy exceeds the baseline by an average of 17.65\%. The code is available at https://github.com/tsinghua-fib-lab/PESLA.
****
Yuankai Luo,Xiao-Ming Wu,Hao Zhu
Beyond Random Masking: When Dropout meets Graph Convolutional Networks
Graph Convolutional Networks (GCNs) have emerged as powerful tools for learning on graph-structured data, yet the behavior of dropout in these models remains poorly understood. This paper presents a comprehensive theoretical analysis of dropout in GCNs, revealing that its primary role differs fundamentally from standard neural networks - preventing oversmoothing rather than co-adaptation. We demonstrate that dropout in GCNs creates dimension-specific stochastic sub-graphs, leading to a form of structural regularization not present in standard neural networks. Our analysis shows that dropout effects are inherently degree-dependent, resulting in adaptive regularization that considers the topological importance of nodes. We provide new insights into dropout's role in mitigating oversmoothing and derive novel generalization bounds that account for graph-specific dropout effects. Furthermore, we analyze the synergistic interaction between dropout and batch normalization in GCNs, uncovering a mechanism that enhances overall regularization. Our theoretical findings are validated through extensive experiments on both node-level and graph-level tasks across 14 datasets. Notably, GCN with dropout and batch normalization outperforms state-of-the-art methods on several benchmarks, demonstrating the practical impact of our theoretical insights.
****
Yuheng Zhang,Dian Yu,Baolin Peng,Linfeng Song,Ye Tian,Mingyue Huo,Nan Jiang,Haitao Mi,Dong Yu
Iterative Nash Policy Optimization: Aligning LLMs with General Preferences via No-Regret Learning
Reinforcement Learning with Human Feedback (RLHF) has achieved great success in aligning large language models (LLMs) with human preferences. Prevalent RLHF approaches are reward-based, following the Bradley-Terry (BT) model assumption, which may not fully capture the complexity of human preferences. In this paper, we explore RLHF under a general preference framework and approach it from a game-theoretic perspective. Specifically, we formulate the problem as a two-player game and propose a novel online algorithm, iterative Nash policy optimization (INPO). The key idea is to let the policy play against itself via no-regret learning, thereby approximating the Nash policy. Unlike previous methods, INPO bypasses the need for estimating the expected win rate for individual responses, which typically incurs high computational or annotation costs. Instead, we introduce a new loss objective that is directly minimized over a preference

dataset. We provide theoretical analysis for our approach and demonstrate its effectiveness through experiments on various representative benchmarks. With an LLaMA-3-8B-based SFT model, INPO achieves a 42.6% length-controlled win rate on AlpacaEval 2.0 and a 3.7.8% win rate on Arena-Hard, showing substantial improvement over the state-of-the-art online RLHF algorithms.
****
Stefan Nielsen,Rachel Teo,Laziz Abdullaev,Tan Minh Nguyen
Tight Clusters Make Specialized Experts
Sparse Mixture-of-Experts (MoE) architectures have emerged as a promising approach to decoupling model capacity from computational cost. At the core of the MoE model is the router, which learns the underlying clustering structure of the input distribution in order to send input tokens to appropriate experts. However, latent clusters may be unidentifiable in high dimension, which causes slow convergence, susceptibility to data contamination, and overall degraded representations as the router is unable to perform appropriate token-expert matching. We examine the router through the lens of clustering optimization and derive optimal feature weights that maximally identify the latent clusters. We use these weights to compute the token-expert routing assignments in an adaptively transformed space that promotes well-separated clusters, which helps identify the best-matched expert for each token. In particular, for each expert cluster, we compute a set of weights that scales features according to whether that expert clusters tightly along that feature. We term this novel router the Adaptive Clustering (AC) router. Our AC router enables the MoE model to obtain three connected benefits: 1) faster convergence, 2) better robustness to data corruption, and 3) overall performance improvement, as experts are specialized in semantically distinct regions of the input space. We empirically demonstrate the advantages of our AC router over baseline routing methods when applied on a variety of MoE backbones for language modeling and image recognition tasks in both clean and corrupted settings.
****
Zeinab Navidi,Jun Ma,Esteban Miglietta,Le Liu,Anne E Carpenter,Beth A Cimini,Benjamin Haibe-Kains,BO WANG
MorphoDiff: Cellular Morphology Painting with Diffusion Models
Understanding cellular responses to external stimuli is critical for parsing biological mechanisms and advancing therapeutic development. High-content image-based assays provide a cost-effective approach to examine cellular phenotypes induced by diverse interventions, which offers valuable insights into biological processes and cellular states. We introduce MorphoDiff, a generative pipeline to predict high-resolution cell morphological responses under different conditions based on perturbation encoding. To the best of our knowledge, MorphoDiff is the first framework capable of producing guided, high-resolution predictions of cell morphology that generalize across both chemical and genetic interventions. The model integrates perturbation embeddings as guiding signals within a 2D latent diffusion model. The comprehensive computational, biological, and visual validations across three open-source Cell Painting datasets show that MorphoDiff can generate high-fidelity images and produce meaningful biology signals under various interventions. We envision the model will facilitate efficient in silico exploration of perturbational landscapes towards more effective drug discovery studies.
****
Jincheng Zhong,XiangCheng Zhang,Jianmin Wang,Mingsheng Long
Domain Guidance: A Simple Transfer Approach for a Pre-trained Diffusion Model
Recent advancements in diffusion models have revolutionized generative modeling. However, the impressive and vivid outputs they produce often come at the cost of significant model scaling and increased computational demands. Consequently, building personalized diffusion models based on off-the-shelf models has emerged as an appealing alternative. In this paper, we introduce a novel perspective on conditional generation for transferring a pre-trained model. From this viewpoint, we propose *Domain Guidance*, a straightforward transfer approach that leverages pre-trained knowledge to guide the sampling process toward the target domain. Domain Guidance shares a formulation similar to advanced classifier-free guidance, facilitating better domain alignment and higher-quality generations. We prov

ide both empirical and theoretical analyses of the mechanisms behind Domain Guid
ance. Our experimental results demonstrate its substantial effectiveness across
various transfer benchmarks, achieving over a 19.6\% improvement in FID and a 23
.4\% improvement in FD$_\text{DINOv2}$ compared to standard fine-tuning. Notably
, existing fine-tuned models can seamlessly integrate Domain Guidance to leverag
e these benefits, without additional training.
****
Laurent Condat,Arto Maranjyan,Peter Richtárik
LoCoDL: Communication-Efficient Distributed Learning with Local Training and Com
pression
In Distributed optimization and Learning, and even more in the modern framew
ork of federated learning, communication, which is slow and costly, is critical.
 We introduce LoCoDL, a communication-efficient algorithm that leverages the two
 popular and effective techniques of Local training, which reduces the communi
cation frequency, and Compression, in which short bitstreams are sent instead
of full-dimensional vectors of floats. LoCoDL works with a large class of unbias
ed compressors that includes widely-used sparsification and quantization methods
. LoCoDL provably benefits from local training and compression and enjoys a doub
ly-accelerated communication complexity, with respect to the condition number of
 the functions and the model dimension, in the general heterogeneous regime with
 strongly convex functions. This is confirmed in practice, with LoCoDL outperfor
ming existing algorithms.
****
Xinze Li,Sen Mei,Zhenghao Liu,Yukun Yan,Shuo Wang,Shi Yu,Zheni Zeng,Hao Chen,Ge
Yu,Zhiyuan Liu,Maosong Sun,Chenyan Xiong
RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rew
ards
Retrieval-Augmented Generation (RAG) has proven its effectiveness in mitigating
hallucinations in Large Language Models (LLMs) by retrieving knowledge from exte
rnal resources. To adapt LLMs for the RAG systems, current approaches use instru
ction tuning to optimize LLMs, improving their ability to utilize retrieved know
ledge. This supervised fine-tuning (SFT) approach focuses on equipping LLMs to h
andle diverse RAG tasks using different instructions. However, it trains RAG mod
ules to overfit training signals and overlooks the varying data preferences amon
g agents within the RAG system. In this paper, we propose a Differentiable Data
Rewards (DDR) method, which end-to-end trains RAG systems by aligning data prefe
rences between different RAG modules. DDR works by collecting the rewards to opt
imize each agent in the RAG system with the rollout method, which prompts agents
 to sample some potential responses as perturbations, evaluates the impact of th
ese perturbations on the whole RAG system, and subsequently optimizes the agent
to produce outputs that improve the performance of the RAG system. Our experimen
ts on various knowledge-intensive tasks demonstrate that DDR significantly outpe
rforms the SFT method, particularly for LLMs with smaller-scale parameters that
depend more on the retrieved knowledge. Additionally, DDR exhibits a stronger ca
pability to align the data preference between RAG modules. The DDR method makes
the generation module more effective in extracting key information from document
s and mitigating conflicts between parametric memory and external knowledge. All
 codes are available at https://github.com/OpenMatch/RAG-DDR.
****
Zhangchen Xu,Fengqing Jiang,Luyao Niu,Yuntian Deng,Radha Poovendran,Yejin Choi,B
ill Yuchen Lin
Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Not
hing
High-quality instruction data is critical for aligning large language models (LL
Ms). Although some models, such as Llama-3-Instruct, have open weights, their al
ignment data remain private, which hinders the democratization of AI. High human
 labor costs and a limited, predefined scope for prompting prevent existing open
-source data creation methods from scaling effectively, potentially limiting the
 diversity and quality of public alignment datasets. Is it possible to synthesiz
e high-quality instruction data at scale by extracting it directly from an align

ed LLM? We present a self-synthesis method for generating large-scale alignment data named Magpie. Our key observation is that aligned LLMs like Llama-3-Instruct can generate a user query when we input only the pre-query templates up to the position reserved for user messages, thanks to their auto-regressive nature. We use this method to prompt Llama-3-Instruct and generate 4 million instructions along with their corresponding responses. We further introduce extensions of Magpie for filtering, generating multi-turn, preference optimization, domain-specific and multilingual datasets. We perform a comprehensive analysis of the Magpie-generated data. To compare Magpie-generated data with other public instruction datasets (e.g., ShareGPT, WildChat, Evol-Instruct, UltraChat, OpenHermes, Tulu-V2-Mix, GenQA), we fine-tune Llama-3-8B-Base with each dataset and evaluate the performance of the fine-tuned models. Our results indicate that using Magpie for supervised fine-tuning (SFT) solely can surpass the performance of previous public datasets utilized for both SFT and preference optimization, such as direct preference optimization with UltraFeedback. We also show that in some tasks, models supervised fine-tuned with Magpie perform comparably to the official Llama-3-8B-Instruct, despite the latter being enhanced with 10 million data points through SFT and subsequent preference optimization. This advantage is evident on alignment benchmarks such as AlpacaEval, ArenaHard, and WildBench.
****
Yong Guo,Shulian Zhang,Haolin Pan,Jing Liu,Yulun Zhang,Jian Chen
Gap Preserving Distillation by Building Bidirectional Mappings with A Dynamic Teacher
Knowledge distillation aims to transfer knowledge from a large teacher model to a compact student counterpart, often coming with a significant performance gap between them. Interestingly, we find that a too-large performance gap can hamper the training process.
To alleviate this, we propose a **Gap Preserving Distillation (GPD)** method that trains an additional dynamic teacher model from scratch along with the student to maintain a reasonable performance gap. To further strengthen distillation, we develop a hard strategy by enforcing both models to share parameters. Besides, we also build the soft bidirectional mappings between them through ***Inverse Reparameterization (IR)*** and ***Channel-Branch Reparameterization (CBR)***.
IR initializes a larger dynamic teacher with approximately the same accuracy as the student to avoid a too large gap in early stage of training. CBR enables direct extraction of an effective student model from the dynamic teacher without post-training.
In experiments, GPD significantly outperforms existing distillation methods on top of both CNNs and transformers, achieving up to 1.58\% accuracy improvement. Interestingly, GPD also generalizes well to the scenarios without a pre-trained teacher, including training from scratch and fine-tuning, yielding a large improvement of 1.80\% and 0.89\% on ResNet18, respectively.
****
XuDong Wang,Shaolun Zhang,Shufan Li,Kehan Li,Konstantinos Kallidromitis,Yusuke Kato,Kazuki Kozuka,Trevor Darrell
SegLLM: Multi-round Reasoning Segmentation with Large Language Models
We present SegLLM, a novel multi-round interactive reasoning segmentation model that enhances LLM-based segmentation by exploiting conversational memory of both visual and textual outputs. By leveraging a mask-aware multimodal LLM, SegLLM re-integrates previous segmentation results into its input stream, enabling it to reason about complex user intentions and segment objects in relation to previously identified entities, including positional, interactional, and hierarchical relationships, across multiple interactions. This capability allows SegLLM to respond to visual and text queries in a chat-like manner. Evaluated on the newly curated MRSeg benchmark, SegLLM outperforms existing methods in multi- round interactive reasoning segmentation by over 20%. Additionally, we observed that training on multi-round reasoning segmentation data enhances performance on standard single-round referring segmentation and localization tasks, resulting in a 5.5% increase in cIoU for referring expression segmentation and a 4.5% improvement in Acc@0.5 for referring expression localization.

****
Yan Ru Pei
Let SSMs be ConvNets: State-space Modeling with Optimal Tensor Contractions
We introduce Centaurus, a class of networks composed of generalized state-space model (SSM) blocks, where the SSM operations can be treated as tensor contractions during training. The optimal order of tensor contractions can then be systematically determined for every SSM block to maximize training efficiency. This allows more flexibility in designing SSM blocks beyond the depthwise-separable configuration commonly implemented. The new design choices will take inspiration from classical convolutional blocks including group convolutions, full convolutions, and bottleneck blocks. We architect the Centaurus network with a mixture of these blocks, to balance between network size and performance, as well as memory and computational efficiency during both training and inference. We show that this heterogeneous network design outperforms its homogeneous counterparts in raw audio processing tasks including keyword spotting, speech denoising, and automatic speech recognition (ASR). For ASR, Centaurus is the first network with competitive performance that can be made fully state-space based, without using any non linear recurrence (LSTMs), explicit convolutions (CNNs), or (surrogate) attention mechanism.
****
Sedrick Keh,Jean Mercat,Samir Yitzhak Gadre,Kushal Arora,Igor Vasiljevic,Benjamin Burchfiel,Shuran Song,Russ Tedrake,Thomas Kollar,Ludwig Schmidt,Achal Dave
Should VLMs be Pre-trained with Image Data?
Pre-trained LLMs that are further trained with image data perform well on vision-language tasks.
While adding images during a second training phase effectively unlocks this capability, it is unclear how much of a gain or loss this two-step pipeline gives over VLMs which integrate images earlier into the training process.
To investigate this, we train models spanning various datasets, scales, image-text ratios, and amount of pre-training done before introducing vision tokens.
We then fine-tune these models and evaluate their downstream performance on a suite of vision-language and text-only tasks.
We find that pre-training with a mixture of image and text data allows models to perform better on vision-language tasks while maintaining strong performance on text-only evaluations.
On an average of 6 diverse tasks, we find that for a 1B model, introducing visual tokens 80\% of the way through pre-training results in a 2\% average improvement over introducing visual tokens to a fully pre-trained model.
****
Yunchao Zhang,Guandao Yang,Leonidas Guibas,Yanchao Yang
InfoGS: Efficient Structure-Aware 3D Gaussians via Lightweight Information Shaping
3D Gaussians, as an explicit scene representation, typically involve thousands to millions of elements per scene. This makes it challenging to control the scene in ways that reflect the underlying semantics, where the number of independent entities is typically much smaller. Especially, if one wants to animate or edit objects in the scene, as this requires coordination among the many Gaussians involved in representing each object. To address this issue, we develop a mutual information shaping technique that enforces resonance and coordination between correlated Gaussians via a Gaussian attribute decoding network. Such correlations can be learned from putative 2D object masks in different views. By approximating the mutual information with the gradients concerning the network parameters, our method ensures consistency between scene elements and enables efficient scene editing by operating on network parameters rather than massive Gaussians. In particular, we develop an effective learning pipeline named ***InfoGS*** with lightweight optimization to shape the attribute decoding network ,while ensuring that the shaping (consistency) is maintained during continuous edits, avoiding re-shaping after parameter changes. Notably, our training only touches a small fraction of all Gaussians in the scene yet attains the desired correlated behavior according to the underlying scene structure. The proposed technique is evaluated

on challenging scenes and demonstrates significant performance improvements in 3D object segmentation and promoting scene interactions, while inducing low computation and memory requirements. Our code is available at: https://github.com/StylesZhang/InfoGS.
****
Jiuqi Wang,Ethan Blaser,Hadi Daneshmand,Shangtong Zhang
Transformers Can Learn Temporal Difference Methods for In-Context Reinforcement Learning
Traditionally, reinforcement learning (RL) agents learn to solve new tasks by updating their neural network parameters through interactions with the task environment. However, recent works demonstrate that some RL agents, after certain pretraining procedures, can learn to solve unseen new tasks without parameter updates, a phenomenon known as in-context reinforcement learning (ICRL). The empirical success of ICRL is widely attributed to the hypothesis that the forward pass of the pretrained agent neural network implements an RL algorithm. In this paper, we support this hypothesis by showing, both empirically and theoretically, that when a transformer is trained for policy evaluation tasks, it can discover and learn to implement temporal difference learning in its forward pass.
****
Daniel Ward,Mark Beaumont,Matteo Fasiolo
SoftCVI: Contrastive variational inference with self-generated soft labels
Estimating a distribution given access to its unnormalized density is pivotal in Bayesian inference, where the posterior is generally known only up to an unknown normalizing constant. Variational inference and Markov chain Monte Carlo methods are the predominant tools for this task; however, both are often challenging to apply reliably, particularly when the posterior has complex geometry. Here, we introduce Soft Contrastive Variational Inference (SoftCVI), which allows a family of variational objectives to be derived through a contrastive estimation framework. The approach parameterizes a classifier in terms of a variational distribution, reframing the inference task as a contrastive estimation problem aiming to identify a single true posterior sample among a set of samples. Despite this framing, we do not require positive or negative samples, but rather learn by sampling the variational distribution and computing ground truth soft classification labels from the unnormalized posterior itself. The objectives have zero variance gradient when the variational approximation is exact, without the need for specialized gradient estimators. We empirically investigate the performance on a variety of Bayesian inference tasks, using both simple (e.g. normal) and expressive (normalizing flow) variational distributions. We find that SoftCVI can be used to form objectives which are stable to train and mass-covering, frequently outperforming inference with other variational approaches.
****
Long Wei,Haodong Feng,Yuchen Yang,Ruiqi Feng,Peiyan Hu,Xiang Zheng,Tao Zhang,Dixia Fan,Tailin Wu
CL-DiffPhyCon: Closed-loop Diffusion Control of Complex Physical Systems
The control problems of complex physical systems have broad applications in science and engineering. Previous studies have shown that generative control methods based on diffusion models offer significant advantages for solving these problems. However, existing generative control approaches face challenges in both performance and efficiency when extended to the closed-loop setting, which is essential for effective control. In this paper, we propose an efficient Closed-Loop Diffusion method for Physical systems Control (CL-DiffPhyCon). By employing an asynchronous denoising framework for different physical time steps, CL-DiffPhyCon generates control signals conditioned on real-time feedback from the system with significantly reduced computational cost during sampling. Additionally, the control process could be further accelerated by incorporating fast sampling techniques, such as DDIM. We evaluate CL-DiffPhyCon on two tasks: 1D Burgers' equation control and 2D incompressible fluid control. The results demonstrate that CL-DiffPhyCon achieves superior control performance with significant improvements in sampling efficiency. The code can be found at https://github.com/AI4Science-WestlakeU/CL_DiffPhyCon.

****
Amaia Cardiel,Eloi Zablocki,Elias Ramzi,Oriane Siméoni,Matthieu Cord
LLM-wrapper: Black-Box Semantic-Aware Adaptation of Vision-Language Models for R
eferring Expression Comprehension
Vision Language Models (VLMs) have demonstrated remarkable capabilities in vario
us open-vocabulary tasks, yet their zero-shot performance lags behind task-speci
fic fine-tuned models, particularly in complex tasks like Referring Expression C
omprehension (REC). Fine-tuning usually requires 'white-box' access to the model
's architecture and weights, which is not always feasible due to proprietary or
privacy concerns. In this work, we propose LLM-wrapper, a method for 'black-box'
 adaptation of VLMs for the REC task using Large Language Models (LLMs). LLM-wra
pper capitalizes on the reasoning abilities of LLMs, improved with a light fine-
tuning, to select the most relevant bounding box to match the referring expressi
on, from candidates generated by a zero-shot black-box VLM. Our approach offers
several advantages: it enables the adaptation of closed-source models without ne
eding access to their internal workings, it is versatile and works with any VLM,
 transfers to new VLMs, and it allows for the adaptation of an ensemble of VLMs.
 We evaluate LLM-wrapper on multiple datasets using different VLMs and LLMs, dem
onstrating significant performance improvements and highlighting the versatility
 of our method. While LLM-wrapper is not meant to directly compete with standard
 white-box fine-tuning, it offers a practical and effective alternative for blac
k-box VLM adaptation. The code will be open-sourced.
****
Alexandre Galashov,Valentin De Bortoli,Arthur Gretton
Deep MMD Gradient Flow without adversarial training
We propose a gradient flow procedure for generative modeling by transporting par
ticles from an initial source distribution to a target distribution, where the g
radient field on the particles is given by a noise-adaptive Wasserstein Gradient
 of the Maximum Mean Discrepancy (MMD). The noise adaptive MMD is trained on dat
a distributions corrupted by increasing levels of noise, obtained via a forward
diffusion process, as commonly used in denoising diffusion probabilistic models.
 The result is a generalization of MMD Gradient Flow, which we call Diffusion-MM
D-Gradient Flow or DMMD. The divergence training procedure is related to discrim
inator training in Generative Adversarial Networks (GAN), but does not require a
dversarial training. We obtain competitive empirical performance in unconditiona
l image generation on CIFAR10, MNIST, CELEB-A (64 x64) and LSUN Church (64 x 64)
. Furthermore, we demonstrate the validity of the approach when MMD is replaced
by a lower bound on the KL divergence.
****
Seamus Somerstep,Felipe Maia Polo,Moulinath Banerjee,Yaacov Ritov,Mikhail Yuroch
kin,Yuekai Sun
A transfer learning framework for weak to strong generalization
Modern large language model (LLM) alignment techniques rely on human feedback, b
ut it is unclear whether the techniques fundamentally limit the capabilities of
aligned LLMs. In particular, it is unclear whether it is possible to align (stro
nger) LLMs with superhuman capabilities with (weaker) human feedback *without de
grading their capabilities*. This is an instance of the weak-to-strong generaliz
ation problem: using weaker (less capable) feedback to train a stronger (more ca
pable) model. We prove that weak-to-strong generalization is possible by eliciti
ng latent knowledge from pre-trained LLMs. In particular, we cast the weak-to-st
rong generalization problem as a transfer learning problem in which we wish to t
ransfer a latent concept from a weak model to a strong pre-trained model. We pro
ve that a naive fine-tuning approach suffers from fundamental limitations, but a
n alternative refinement-based approach suggested by the problem structure prova
bly overcomes the limitations of fine-tuning. Finally, we demonstrate the practi
cal applicability of the refinement approach in multiple LLM alignment tasks.
****
Benoit Dufumier,Javiera Castillo Navarro,Devis Tuia,Jean-Philippe Thiran
What to align in multimodal contrastive learning?
Humans perceive the world through multisensory integration, blending the informa

tion of different modalities to adapt their behavior.
Contrastive learning offers an appealing solution for multimodal self-supervised learning. Indeed, by considering each modality as a different view of the same entity, it learns to align features of different modalities in a shared representation space. However, this approach is intrinsically limited as it only learns shared or redundant information between modalities, while multimodal interactions can arise in other ways. In this work, we introduce CoMM, a Contrastive Multimodal learning strategy that enables the communication between modalities in a single multimodal space. Instead of imposing cross- or intra- modality constraints, we propose to align multimodal representations by maximizing the mutual information between augmented versions of these multimodal features. Our theoretical analysis shows that shared, synergistic and unique terms of information naturally emerge from this formulation, allowing us to estimate multimodal interactions beyond redundancy. We test CoMM both in a controlled and in a series of real-world settings: in the former, we demonstrate that CoMM effectively captures redundant, unique and synergistic information between modalities. In the latter, CoMM learns complex multimodal interactions and achieves state-of-the-art results on seven multimodal tasks.
****
Kalliopi Basioti,Pritish Sahu,Tony Qingze Liu,Zihao Xu,Hao Wang,Vladimir Pavlovic
GenVP: Generating Visual Puzzles with Contrastive Hierarchical VAEs
Raven’s Progressive Matrices (RPMs) is an established benchmark to examine the ability to perform high-level abstract visual reasoning (AVR). Despite the current success of algorithms that solve this task, humans can generalize beyond a given puzzle and create new puzzles given a set of rules, whereas machines remain locked in solving a fixed puzzle from a curated choice list. We propose Generative Visual Puzzles (GenVP), a framework to model the entire RPM generation process, a substantially more challenging task. Our model’s capability spans from generating multiple solutions for one specific problem prompt to creating complete new puzzles out of the desired set of rules. Experiments on five different datasets indicate that GenVP achieves state-of-the-art (SOTA) performance both in puzzle-solving accuracy and out-of-distribution (OOD) generalization in 22 out of 24 OOD scenarios. Further, compared to SOTA generative approaches, which struggle to solve RPMs when the feasible solution space increases, GenVP efficiently generalizes to these challenging scenarios. Moreover, our model demonstrates the ability to produce a wide range of complete RPMs given a set of abstract rules by effectively capturing the relationships between abstract rules and visual object properties.
****
Ziyue Qiao,Junren Xiao,Qingqiang Sun,Meng Xiao,Xiao Luo,Hui Xiong
Towards Continuous Reuse of Graph Models via Holistic Memory Diversification
This paper addresses the challenge of incremental learning in growing graphs with increasingly complex tasks. The goal is to continuously train a graph model to handle new tasks while retaining proficiency in previous tasks via memory replay. Existing methods usually overlook the importance of memory diversity, limiting in selecting high-quality memory from previous tasks and remembering broad previous knowledge within the scarce memory on graphs. To address that, we introduce a novel holistic Diversified Memory Selection and Generation (DMSG) framework for incremental learning in graphs, which first introduces a buffer selection strategy that considers both intra-class and inter-class diversities, employing an efficient greedy algorithm for sampling representative training nodes from graphs into memory buffers after learning each new task. Then, to adequately rememorize the knowledge preserved in the memory buffer when learning new tasks, a diversified memory generation replay method is introduced. This method utilizes a variational layer to generate the distribution of buffer node embeddings and sample synthesized ones for replaying. Furthermore, an adversarial variational embedding learning method and a reconstruction-based decoder are proposed to maintain the integrity and consolidate the generalization of the synthesized node embeddings, respectively. Extensive experimental results on publicly accessible dataset

s demonstrate the superiority of DMSG over state-of-the-art methods.
****

Junli Cao,Vidit Goel,Chaoyang Wang,Anil Kag,Ju Hu,Sergei Korolev,Chenfanfu Jiang,Sergey Tulyakov,Jian Ren

Lightweight Predictive 3D Gaussian Splats

Recent approaches representing 3D objects and scenes using Gaussian splats show increased rendering speed across a variety of platforms and devices. While rendering such representations is indeed extremely efficient, storing and transmitting them is often prohibitively expensive. To represent large-scale scenes, one often needs to store millions of 3D Gaussian, which can occupy up to gigabytes of storage. This creates a significant practical barrier, preventing widespread adoption on resource-constrained devices.

In this work, we propose a new representation that dramatically reduces the hard drive footprint while featuring similar or improved quality when compared to the standard 3D Gaussian splats. This representation leverages the inherent feature sharing among splats in the close proximity using a hierarchical tree structure, with which only the parent splats need to be stored. We present a method for constructing tree structures from naturally unstructured point clouds. Additionally, we propose the adaptive tree manipulation to prune the redundant trees in the space, while spawn new ones from the significant children splats during the optimization process. On the benchmark datasets, we achieve 20x storage reduction in hard-drive footprint with improved fidelity compared to the vanilla 3DGS and 2-5x reduction compared to the exiting compact solutions. More importantly, we demonstrate the practical application of our method in real-world rendering on mobile devices and AR glasses.
****

JiaKui Hu,Lujia Jin,Zhengjian Yao,Yanye Lu

Universal Image Restoration Pre-training via Degradation Classification

This paper proposes the Degradation Classification Pre-Training (DCPT), which enables models to learn how to classify the degradation type of input images for universal image restoration pre-training. Unlike the existing self-supervised pre-training methods, DCPT utilizes the degradation type of the input image as an extremely weak supervision, which can be effortlessly obtained, even intrinsic in all image restoration datasets. DCPT comprises two primary stages. Initially, image features are extracted from the encoder. Subsequently, a lightweight decoder, such as ResNet18, is leveraged to classify the degradation type of the input image solely based on the features extracted in the first stage, without utilizing the input image. The encoder is pre-trained with a straightforward yet potent DCPT, which is used to address universal image restoration and achieve outstanding performance. Following DCPT, both convolutional neural networks (CNNs) and transformers demonstrate performance improvements, with gains of up to 2.55 dB in the 10D all-in-one restoration task and 6.53 dB in the mixed degradation scenarios. Moreover, previous self-supervised pretraining methods, such as masked image modeling, discard the decoder after pre-training, while our DCPT utilizes the pre-trained parameters more effectively. This superiority arises from the degradation classifier acquired during DCPT, which facilitates transfer learning between models of identical architecture trained on diverse degradation types. Source code and models are available at \url{https://github.com/MILab-PKU/dcpt}.
****

Kyungsu Lee,Haeyun Lee,Jae Youn Hwang

Connectome Mapping: Shape-Memory Network via Interpretation of Contextual Semantic Information

Contextual semantic information plays a pivotal role in the brain's visual interpretation of the surrounding environment. When processing visual information, electrical signals within synapses facilitate the dynamic activation and deactivation of synaptic connections, guided by the contextual semantic information associated with different objects. In the realm of Artificial Intelligence (AI), neural networks have emerged as powerful tools to emulate complex signaling systems, enabling tasks such as classification and segmentation by understanding visual information. However, conventional neural networks have limitations in simulatin

g the conditional activation and deactivation of synapses, collectively known as the connectome, a comprehensive map of neural connections in the brain. Additionally, the pixel-wise inference mechanism of conventional neural networks failed to account for the explicit utilization of contextual semantic information in the prediction process. To overcome these limitations, we developed a novel neural network, dubbed the Shape Memory Network (SMN), which excels in two key areas: (1) faithfully emulating the intricate mechanism of the brain's connectome, and (2) explicitly incorporating contextual semantic information during the inference process. The SMN memorizes the structure suitable for contextual semantic information and leverages this structure at the inference phase. The structural transformation emulates the conditional activation and deactivation of synaptic connections within the connectome. Rigorous experimentation carried out across a range of semantic segmentation benchmarks demonstrated the outstanding performance of the SMN, highlighting its superiority and effectiveness. Furthermore, our pioneering network on connectome emulation reveals the immense potential of the SMN for next-generation neural networks.
****
Qingkai Fang,Shoutao Guo,Yan Zhou,Zhengrui Ma,Shaolei Zhang,Yang Feng
LLaMA-Omni: Seamless Speech Interaction with Large Language Models
Models like GPT-4o enable real-time interaction with large language models (LLMs) through speech, significantly enhancing user experience compared to traditional text-based interaction. However, there is still a lack of exploration on how to build speech interaction models based on open-source LLMs. To address this, we propose LLaMA-Omni, a novel model architecture designed for low-latency and high-quality speech interaction with LLMs. LLaMA-Omni integrates a pretrained speech encoder, a speech adaptor, an LLM, and a streaming speech decoder. It eliminates the need for speech transcription, and can simultaneously generate text and speech responses directly from speech instructions with extremely low latency. We build our model based on the latest Llama-3.1-8B-Instruct model. To align the model with speech interaction scenarios, we construct a dataset named InstructS2S-200K, which includes 200K speech instructions and corresponding speech responses. Experimental results show that compared to previous speech-language models, LLaMA-Omni provides better responses in both content and style, with a response latency as low as 226ms. Additionally, training LLaMA-Omni takes less than 3 days on just 4 GPUs, paving the way for the efficient development of speech-language models in the future.
****
Md Awsafur Rahman,Zaber Ibn Abdul Hakim,Najibul Haque Sarker,Bishmoy Paul,Shaikh Anowarul Fattah
SONICS: Synthetic Or Not - Identifying Counterfeit Songs
The recent surge in AI-generated songs presents exciting possibilities and challenges. These innovations necessitate the ability to distinguish between human-composed and synthetic songs to safeguard artistic integrity and protect human musical artistry. Existing research and datasets in fake song detection only focus on singing voice deepfake detection (SVDD), where the vocals are AI-generated but the instrumental music is sourced from real songs. However, these approaches are inadequate for detecting contemporary end-to-end artificial songs where all components (vocals, music, lyrics, and style) could be AI-generated. Additionally, existing datasets lack music-lyrics diversity, long-duration songs, and open-access fake songs. To address these gaps, we introduce SONICS, a novel dataset for end-to-end Synthetic Song Detection (SSD), comprising over 97k songs (4,751 hours) with over 49k synthetic songs from popular platforms like Suno and Udio. Furthermore, we highlight the importance of modeling long-range temporal dependencies in songs for effective authenticity detection, an aspect entirely overlooked in existing methods. To utilize long-range patterns, we introduce SpecTTTra, a novel architecture that significantly improves time and memory efficiency over conventional CNN and Transformer-based models. For long songs, our top-performing variant outperforms ViT by 8% in F1 score, is 38% faster, and uses 26% less memory, while also surpassing ConvNeXt with a 1% F1 score gain, 20% speed boost, and 67% memory reduction.

****
Gust Verbruggen,Ashish Tiwari,Mukul Singh,Vu Le,Sumit Gulwani
Execution-guided within-prompt search for programming-by-example
Large language models (LLMs) can generate code from examples without being limit
ed to a DSL, but they lack search, as sampled programs are independent.
In this paper, we use an LLM as a policy that generates lines of code and then j
oin these lines of code to let the LLM implicitly estimate the value of each of
these lines in its next iteration.
We further guide the policy and value estimation by executing each line and anno
tating it with its results on the given examples.
This allows us to search for programs within a single (expanding) prompt until a
 sound program is found, by letting the policy reason in both the syntactic (cod
e) and semantic (execution) space.
We evaluate within-prompt search on straight-line Python code generation using f
ive benchmarks across different domains (strings, lists, and arbitrary Python pr
ogramming problems).
We show that the model uses the execution results to guide the search and that w
ithin-prompt search performs well at low token budgets.
We also analyze how the model behaves as a policy and value, show that it can pa
rallelize the search, and that it can implicitly backtrack over earlier generati
ons.
****
Li Huaqiu,HuXiaowan,Haoqian Wang
Interpretable Unsupervised Joint Denoising and Enhancement for Real-World low-li
ght Scenarios
Real-world low-light images often suffer from complex degradations such as local
 overexposure, low brightness, noise, and uneven illumination. Supervised method
s tend to overfit to specific scenarios, while unsupervised methods, though bett
er at generalization, struggle to model these degradations due to the lack of re
ference images. To address this issue, we propose an interpretable, zero-referen
ce joint denoising and low-light enhancement framework tailored for real-world s
cenarios. Our method derives a training strategy based on paired sub-images with
 varying illumination and noise levels, grounded in physical imaging principles
and retinex theory. Additionally, we leverage the Discrete Cosine Transform (DCT
) to perform frequency domain decomposition in the sRGB space, and introduce an
implicit-guided hybrid representation strategy that effectively separates intric
ate compounded degradations. In the backbone network design, we develop retinal
decomposition network guided by implicit degradation representation mechanisms.
Extensive experiments demonstrate the superiority of our method. Code will be av
ailable at https://github.com/huaqlili/unsupervised-light-enhance-ICLR2025.
****
Shuo Xie,Mohamad Amin Mohamadi,Zhiyuan Li
Adam Exploits ℓ_∞-geometry of Loss Landscape via Coordinate-wise Adapt
ivity
Adam outperforms SGD when training language models. Yet this advantage is not we
ll-understood theoretically -- previous convergence analysis for Adam and SGD m
ainly focuses on the number of steps T and is already minimax-optimal in non-c
onvex cases, which are both $\widetilde{O}(T^{-1/4})$. In this work, we argue th
at the exploitation of nice ℓ_∞-geometry is the key advantage of Adam
over SGD. More specifically, we give a new convergence analysis for Adam under n
ovel assumptions that loss is smooth under ℓ_∞-geometry rather than th
e more common ℓ_2-geometry, which yields a much better empirical smoothness
 constant for GPT-2 and ResNet models. Our experiments confirm that Adam perform
s much worse when the favorable ℓ_∞-geometry is changed while SGD prov
ably remains unaffected. We also extend the convergence analysis to blockwise Ad
am under novel blockwise smoothness assumptions.
****
Heejun Lee,Geon Park,Youngwan Lee,Jaduk Suh,Jina Kim,Wonyong Jeong,Bumsik Kim,Hy
emin Lee,Myeongjae Jeon,Sung Ju Hwang
A Training-Free Sub-quadratic Cost Transformer Model Serving Framework with Hier

archically Pruned Attention
In modern large language models (LLMs), increasing the context length is crucial for improving comprehension and coherence in long-context, multi-modal, and retrieval-augmented language generation.
While many recent transformer models attempt to extend their context length over a million tokens, they remain impractical due to the quadratic time and space complexities.
Although recent works on linear and sparse attention mechanisms can achieve this goal, their real-world applicability is often limited by the need to re-train from scratch and significantly worse performance. In response, we propose a novel approach, Hierarchically Pruned Attention (HiP), which reduces the time complexity of the attention mechanism to $O(T \log T)$ and the space complexity to $O(T)$, where T is the sequence length.
We notice a pattern in the attention scores of pretrained LLMs where tokens close together tend to have similar scores, which we call "attention locality". Based on this observation, we utilize a novel tree-search-like algorithm that estimates the top-k key tokens for a given query on the fly, which is mathematically guaranteed to have better performance than random attention pruning. In addition to improving the time complexity of the attention mechanism, we further optimize GPU memory usage by implementing KV cache offloading, which stores only $O(\log T)$ tokens on the GPU while maintaining similar decoding throughput. Experiments on benchmarks show that HiP, with its training-free nature, significantly reduces both prefill and decoding latencies, as well as memory usage, while maintaining high-quality generation with minimal degradation.
HiP enables pretrained LLMs to scale up to millions of tokens on commodity GPUs, potentially unlocking long-context LLM applications previously deemed infeasible.
****
Jingjing Gong,Yu Pei,Siyu Long,Yuxuan Song,Zhe Zhang,Wenhao Huang,Ziyao Cao,Shuyi Zhang,Hao Zhou,Wei-Ying Ma
Steering Protein Family Design through Profile Bayesian Flow
Protein family design emerges as a promising alternative by combining the advantages of de novo protein design and mutation-based directed evolution.In this paper, we propose ProfileBFN, the Profile Bayesian Flow Networks, for specifically generative modeling of protein families. ProfileBFN extends the discrete Bayesian Flow Network from an MSA profile perspective, which can be trained on single protein sequences by regarding it as a degenerate profile, thereby achieving efficient protein family design by avoiding large-scale MSA data construction and training. Empirical results show that ProfileBFN has a profound understanding of proteins. When generating diverse and novel family proteins, it can accurately capture the structural characteristics of the family. The enzyme produced by this method is more likely than the previous approach to have the corresponding function, offering better odds of generating diverse proteins with the desired functionality.
****
Max Klabunde,Tassilo Wald,Tobias Schumacher,Klaus Maier-Hein,Markus Strohmaier,Florian Lemmerich
ReSi: A Comprehensive Benchmark for Representational Similarity Measures
Measuring the similarity of different representations of neural architectures is a fundamental task and an open research challenge for the machine learning community. This paper presents the first comprehensive benchmark for evaluating representational similarity measures based on well-defined groundings of similarity. The representational similarity (ReSi) benchmark consists of (i) six carefully designed tests for similarity measures, (ii) 24 similarity measures, (iii) 14 neural network architectures, and (iv) seven datasets, spanning over the graph, language, and vision domains. The benchmark opens up several important avenues of research on representational similarity that enable novel explorations and applications of neural architectures. We demonstrate the utility of the ReSi benchmark by conducting experiments on various neural network architectures, real world datasets and similarity measures. All components of the benchmark are publicly a

vailable and thereby facilitate systematic reproduction and production of resear
ch results. The benchmark is extensible, future research can build on and furthe
r expand it. We believe that the ReSi benchmark can serve as a sound platform ca
talyzing future research that aims to systematically evaluate existing and explo
re novel ways of comparing representations of neural architectures.
ReSi is available at https://github.com/mklabunde/resi.
****
Sucheng Ren,Xianhang Li,Haoqin Tu,Feng Wang,Fangxun Shu,Lei Zhang,Jieru Mei,Linj
ie Yang,Peng Wang,Heng Wang,Alan Yuille,Cihang Xie
Autoregressive Pretraining with Mamba in Vision
The vision community has started to build with the recently developed state spac
e model, Mamba, as the new backbone for a range of tasks. This paper shows that
Mamba's visual capability can be significantly enhanced through autoregressive p
retraining, a direction not previously explored. Efficiency-wise, the autoregres
sive nature can well capitalize on the Mamba's unidirectional recurrent structur
e, enabling faster overall training speed compared to other training strategies
like mask modeling. Performance-wise, autoregressive pretraining equips the Mamb
a architecture with markedly higher accuracy over its supervised-trained counter
parts and, more importantly, successfully unlocks its scaling potential to large
 and even huge model sizes. For example, with autoregressive pretraining, a base
-size Mamba attains 83.2\% ImageNet accuracy, outperforming its supervised count
erpart by 2.0\%; our huge-size Mamba, the largest Vision Mamba to date, attains
85.0\% ImageNet accuracy (85.5\% when finetuned with 384×384 inputs), not
ably surpassing all other Mamba variants in vision. The code is available at \ur
l{https://github.com/OliverRensu/ARM}.
****
Quan Dao,Khanh Doan,Di Liu,Trung Le,Dimitris N. Metaxas
Improved Training Technique for Latent Consistency Models
Consistency models are a new family of generative models capable of producing hi
gh-quality samples in either a single step or multiple steps. Recently, consiste
ncy models have demonstrated impressive performance, achieving results on par wi
th diffusion models in the pixel space. However, the success of scaling consiste
ncy training to large-scale datasets, particularly for text-to-image and video g
eneration tasks, is determined by performance in the latent space. In this work,
 we analyze the statistical differences between pixel and latent spaces, discove
ring that latent data often contains highly impulsive outliers, which significan
tly degrade the performance of iCT in the latent space. To address this, we repl
ace Pseudo-Huber losses with Cauchy losses, effectively mitigating the impact of
 outliers. Additionally, we introduce a diffusion loss at early timesteps and em
ploy optimal transport (OT) coupling to further enhance performance. Lastly, we
introduce the adaptive scaling-c scheduler to manage the robust training proce
ss and adopt Non-scaling LayerNorm in the architecture to better capture the sta
tistics of the features and reduce outlier impact. With these strategies, we suc
cessfully train latent consistency models capable of high-quality sampling with
one or two steps, significantly narrowing the performance gap between latent con
sistency and diffusion models. The implementation is released here: \url{https:/
/github.com/quandao10/sLCT/}
****
Juntao Dai,Taiye Chen,Yaodong Yang,Qian Zheng,Gang Pan
Mitigating Reward Over-Optimization in RLHF via Behavior-Supported Regularizatio
n
Reinforcement learning from human feedback (RLHF) is an effective method for ali
gning large language models (LLMs) with human values. However, reward over-optim
ization remains an open challenge leading to discrepancies between the performan
ce of LLMs under the reward model and the true human objectives. A primary contr
ibutor to reward over-optimization is the extrapolation error that arises when t
he reward model evaluates out-of-distribution (OOD) responses. However, current
methods still fail to prevent the increasing frequency of OOD response generatio
n during the reinforcement learning (RL) process and are not effective at handli
ng extrapolation errors from OOD responses. In this work, we propose the *Behavi

or-Supported Policy Optimization* (BSPO) method to mitigate the reward over-optimization issue. Specifically, we define *behavior policy* as the next token distribution of the reward training dataset to model the in-distribution (ID) region of the reward model. Building on this, we introduce the behavior-supported Bellman operator to regularize the value function, penalizing all OOD values without impacting the ID ones. Consequently, BSPO reduces the generation of OOD responses during the RL process, thereby avoiding overestimation caused by the reward model's extrapolation errors. Theoretically, we prove that BSPO guarantees a monotonic improvement of the supported policy until convergence to the optimal behavior-supported policy. Empirical results from extensive experiments show that BSPO outperforms baselines in preventing reward over-optimization due to OOD evaluation and finding the optimal ID policy.
****
Dan MacKinlay,Russell Tsuchida,Daniel Edward Pagendam,Petra Kuhnert
Gaussian Ensemble Belief Propagation for Efficient Inference in High-Dimensional, Black-box Systems
Efficient inference in high-dimensional models is a central challenge in machine learning.
We introduce the Gaussian Ensemble Belief Propagation (GEnBP) algorithm, which combines the strengths of the Ensemble Kalman Filter (EnKF) and Gaussian Belief Propagation (GaBP) to address this challenge.
GEnBP updates ensembles of prior samples into posterior samples by passing low-rank local messages over the edges of a graphical model, enabling efficient handling of high-dimensional states, parameters, and complex, noisy, black-box generative processes.
By utilizing local message passing within a graphical model structure, GEnBP effectively manages complex dependency structures and remains computationally efficient even when the ensemble size is much smaller than the inference dimension --- a common scenario in spatiotemporal modeling, image processing, and physical model inversion.
We demonstrate that GEnBP can be applied to various problem structures, including data assimilation, system identification, and hierarchical models, and show through experiments that it outperforms existing belief propagation methods in terms of accuracy and computational efficiency.

Supporting code is available at https://github.com/danmackinlay/GEnBP}{github.com/danmackinlay/GEnBP
****
Xierui Wang,Siming Fu,Qihan Huang,Wanggui He,Hao Jiang
MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance
Recent advancements in text-to-image generation models have dramatically enhanced the generation of photorealistic images from textual prompts, leading to an increased interest in personalized text-to-image applications, particularly in multi-subject scenarios. However, these advances are hindered by two main challenges: firstly, the need to accurately maintain the details of each referenced subject in accordance with the textual descriptions; and secondly, the difficulty in achieving a cohesive representation of multiple subjects in a single image without introducing inconsistencies. To address these concerns, our research introduces the MS-Diffusion framework for layout-guided zero-shot image personalization with multi-subjects. This innovative approach integrates grounding tokens with the feature resampler to maintain detail fidelity among subjects. With the layout guidance, MS-Diffusion further improves the cross-attention to adapt to the multi-subject inputs, ensuring that each subject condition acts on specific areas. The proposed multi-subject cross-attention orchestrates harmonious inter-subject compositions while preserving the control of texts. Comprehensive quantitative and qualitative experiments affirm that this method surpasses existing models in both image and text fidelity, promoting the development of personalized text-to-image generation.
****
Tiago Silva,Amauri H Souza,Omar Rivasplata,Vikas Garg,Samuel Kaski,Diego Mesquit

a
Generalization and Distributed Learning of GFlowNets
Conventional wisdom attributes the success of Generative Flow Networks (GFlowNet
s) to their ability to exploit the compositional structure of the sample space f
or learning generalizable flow functions (Bengio et al., 2021). Despite the abun
dance of empirical evidence, formalizing this belief with verifiable non-vacuous
 statistical guarantees has remained elusive. We address this issue with the fir
st data-dependent generalization bounds for GFlowNets. We also elucidate the neg
ative impact of the state space size on the generalization performance of these
models via Azuma-Hoeffding-type oracle PAC-Bayesian inequalities. We leverage ou
r theoretical insights to design a novel distributed learning algorithm for GFlo
wNets, which we call *Subgraph Asynchronous Learning* (SAL). In a nutshell, SAL
utilizes a divide-and-conquer strategy: multiple GFlowNets are trained in parall
el on smaller subnetworks of the flow network, and then aggregated with an addit
ional GFlowNet that allocates appropriate flow to each subnetwork. Our experime
nts with synthetic and real-world problems demonstrate the benefits of SAL over
centralized training in terms of mode coverage and distribution matching.
****
Tianchun Wang,Yuanzhou Chen,Zichuan Liu,Zhanwen Chen,Haifeng Chen,Xiang Zhang,We
i Cheng
Humanizing the Machine: Proxy Attacks to Mislead LLM Detectors
The advent of large language models (LLMs) has revolutionized the field of text
generation, producing outputs that closely mimic human-like writing. Although ac
ademic and industrial institutions have developed detectors to prevent the malic
ious usage of LLM-generated texts, other research has doubt about the robustness
 of these systems. To stress test these detectors, we introduce a humanized prox
y-attack (HUMPA) strategy that effortlessly compromises LLMs, causing them to pr
oduce outputs that align with human-written text and mislead detection systems.
Our method attacks the source model by leveraging a reinforcement learning (RL)
fine-tuned humanized small language model (SLM) in the decoding phase. Through a
n in-depth analysis, we demonstrate that our attack strategy is capable of gener
ating responses that are indistinguishable to detectors, preventing them from di
fferentiating between machine-generated and human-written text. We conduct syste
matic evaluations on extensive datasets using proxy-attacked open-source models,
 including Llama2-13B, Llama3-70B, and Mixtral-8x7B in both white- and black-box
 settings. Our findings show that the proxy-attack strategy effectively deceives
 the leading detectors, resulting in an average AUROC drop of 70.4% across multi
ple datasets, with a maximum drop of 95.0% on a single dataset. Furthermore, in
cross-discipline scenarios, our strategy also bypasses these detectors, leading
to a significant relative decrease of up to 90.9%, while in cross-language scena
rio, the drop reaches 91.3%. Despite our proxy-attack strategy successfully bypa
ssing the detectors with such significant relative drops, we find that the gener
ation quality of the attacked models remains preserved, even within a modest uti
lity budget, when compared to the text produced by the original, unattacked sour
ce model.
****
Tianci Liu,Ruirui Li,Yunzhe Qi,Hui Liu,Xianfeng Tang,Tianqi Zheng,Qingyu Yin,Mon
ica Xiao Cheng,Jun Huan,Haoyu Wang,Jing Gao
Unlocking Efficient, Scalable, and Continual Knowledge Editing with Basis-Level
Representation Fine-Tuning
Large language models (LLMs) have achieved remarkable performance on vari-
ous natural language tasks. However, they are trained on static corpora and thei
r
knowledge can become outdated quickly in the fast-changing world. This moti-
vates the development of knowledge editing methods designed to update certain
knowledge in LLMs without changing unrelated others. To make selective edits,
previous efforts often sought to update a small amount of parameters in some spe
-
cific layer(s) of a LLM. Nonetheless, in challenging scenarios, they still fall
short

in making successful edits while preserving knowledge irrelevant to the updates simultaneously, resulting in a notable editing-locality trade-off. In this work, we question if the trade-offs are caused by the fact that parameter-based updates have a global effect, i.e., edited parameters affect all inputs indiscriminately. In light of this, we explore the feasibility of representation fine-tuning, which applied some linear update to a few representations in a learned subspace, for knowledge editing. While being effective to enhance an LLM’s general ability as demonstrated in the previous work, we theoretically show that this linear update imposes a tension in editing-locality trade-off. Subsequently, BaFT is proposed to break the linearity. BaFT computes a weight for each basis that spans a dimension of the subspace based on the input representation. This input-dependent weighting mechanism allows BaFT to manage different types of knowledge in an adaptive way, thereby achieving a better editing-locality trade-off. Experiments on three LLMs with five editing benchmarks in diverse scenarios show the superiority of our method.
****

Xi Xie,Yuebo Luo,Hongwu Peng,Caiwen Ding

RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs

Abstract Top-k selection algorithms are fundamental in a wide range of applications, including high-performance computing, information retrieval, big data processing, and neural network model training. In this paper, we present RTop-K, a highly efficient parallel row-wise top-k selection algorithm specifically designed for GPUs. RTop-K leverages a binary search-based approach to optimize row-wise top-k selection, providing a scalable and accelerated solution.
We conduct a detailed analysis of early stopping in our algorithm, showing that it effectively maintains the testing accuracy of neural network models while substantially improving performance. Our GPU implementation of RTop-K demonstrates superior performance over state-of-the-art row-wise top-k GPU implementations, achieving an average speed-up of up to 11.49× with early stopping and 7.29× without early stopping. Moreover, RTop-K accelerates the overall training workflow of MaxK-GNNs, delivering speed-ups ranging from 11.97% to 33.29% across different models and datasets.
****

Yu Ying Chiu,Liwei Jiang,Yejin Choi

DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life

As users increasingly seek guidance from LLMs for decision-making in daily life, many of these decisions are not clear-cut and depend significantly on the personal values and ethical standards of people. We present DailyDilemmas, a dataset of 1,360 moral dilemmas encountered in everyday life. Each dilemma presents two possible actions, along with affected parties and relevant human values for each action. Based on these dilemmas, we gather a repository of human values covering diverse everyday topics, such as interpersonal relationships, workplace, and environmental issues. With DailyDilemmas, we evaluate LLMs on these dilemmas to determine what action they will choose and the values represented by these action choices. Then, we analyze values through the lens of five theoretical frameworks inspired by sociology, psychology, and philosophy, including the World Values Survey, Moral Foundations Theory, Maslow's Hierarchy of Needs, Aristotle's Virtues, and Plutchik's Wheel of Emotions. For instance, we find LLMs are most aligned with self-expression over survival in World Values Survey and care over loyalty in Moral Foundations Theory. Interestingly, we find substantial preference dif

ferences in models for some core values. For example, for truthfulness, Mixtral-8x7B neglects it by 9.7% while GPT-4-turbo selects it by 9.4%. We also study the recent guidance released by OpenAI (ModelSpec), and Anthropic (Constitutional AI) to understand how their designated principles reflect their models' actual value prioritization when facing nuanced moral reasoning in daily-life settings. Finally, we find that end users cannot effectively steer such prioritization using system prompts.
****
Rui Yang,Jie Wang,Qijie Peng,Ruibo Guo,Guoping Wu,Bin Li
Learning Robust Representations with Long-Term Information for Generalization in Visual Reinforcement Learning
Generalization in visual reinforcement learning (VRL) aims to learn agents that can adapt to test environments with unseen visual distractions. Despite advances in robust representations learning, many methods do not take into account the essential downstream task of sequential decision-making. This leads to representations that lack critical long-term information, impairing decision-making abilities in test environments. To tackle this problem, we propose a novel robust action-value representation learning (ROUSER) under the information bottleneck (IB) framework. ROUSER learns robust representations to capture long-term information from the decision-making objective (i.e., action values). Specifically, ROUSER uses IB to encode robust representations by maximizing their mutual information with action values for long-term information, while minimizing mutual information with state-action pairs to discard irrelevant features. As action values are unknown, ROUSER proposes to decompose robust representations of state-action pairs into one-step rewards and robust representations of subsequent pairs. Thus, it can use known rewards to compute the loss for robust representation learning. Moreover, we show that ROUSER accurately estimates action values using learned robust representations, making it applicable to various VRL algorithms. Experiments demonstrate that ROUSER outperforms several state-of-the-art methods in eleven out of twelve tasks, across both unseen background and color distractions.
****
Mingzhi Wang,Chengdong Ma,Qizhi Chen,Linjian Meng,Yang Han,Jiancong Xiao,Zhaowei Zhang,Jing Huo,Weijie J Su,Yaodong Yang
Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Model Alignment
Self-play methods have demonstrated remarkable success in enhancing model capabilities across various domains. In the context of Reinforcement Learning from Human Feedback (RLHF), self-play not only boosts Large Language Model (LLM) performance but also overcomes the limitations of traditional Bradley-Terry (BT) model assumptions by finding the Nash equilibrium (NE) of a preference-based, two-player constant-sum game. However, existing methods either guarantee only average-iterate convergence, incurring high storage and inference costs, or converge to the NE of a regularized game, failing to accurately reflect true human preferences. In this paper, we introduce Magnetic Preference Optimization (MPO), a novel approach capable of achieving last-iterate convergence to the NE of the original game, effectively overcoming the limitations of existing methods. Building upon Magnetic Mirror Descent (MMD), MPO attains a linear convergence rate, making it particularly suitable for fine-tuning LLMs. To ensure our algorithm is both theoretically sound and practically viable, we present a simple yet effective implementation that adapts the theoretical insights to the RLHF setting. Empirical results demonstrate that MPO can significantly enhance the performance of LLMs, highlighting the potential of self-play methods in alignment.
****
Buqing Nie,Yangqing Fu,Yue Gao
Select before Act: Spatially Decoupled Action Repetition for Continuous Control
Reinforcement Learning (RL) has achieved remarkable success in various continuous control tasks, such as robot manipulation and locomotion.
Different to mainstream RL which makes decisions at individual steps, recent studies have incorporated action repetition into RL, achieving enhanced action persistence with improved sample efficiency and superior performance.

However, existing methods treat all action dimensions as a whole during repetition, ignoring variations among them.
This constraint leads to inflexibility in decisions, which reduces policy agility with inferior effectiveness.
In this work, we propose a novel repetition framework called SDAR, which implements Spatially Decoupled Action Repetition through performing closed-loop act-or-repeat selection for each action dimension individually.
SDAR achieves more flexible repetition strategies, leading to an improved balance between action persistence and diversity.
Compared to existing repetition frameworks, SDAR is more sample efficient with higher policy performance and reduced action fluctuation.
Experiments are conducted on various continuous control scenarios,
demonstrating the effectiveness of spatially decoupled repetition designed in this work.
****
Ching Lam Choi,Alexandre Duplessis,Serge Belongie
Unlearning-based Neural Interpretations
Gradient-based interpretations often require an anchor point of comparison to avoid saturation in computing feature importance. We show that current baselines defined using static functions—constant mapping, averaging or blurring—inject harmful colour, texture or frequency assumptions that deviate from model behaviour. This leads to accumulation of irregular gradients, resulting in attribution maps that are biased, fragile and manipulable. Departing from the static approach, we propose $\texttt{UNI}$ to compute an (un)learnable, debiased and adaptive baseline by perturbing the input towards an $\textit{unlearning direction}$ of steepest ascent. Our method discovers reliable baselines and succeeds in erasing salient features, which in turn locally smooths the high-curvature decision boundaries. Our analyses point to unlearning as a promising avenue for generating faithful, efficient and robust interpretations.
****
Eunseop Yoon,Hee Suk Yoon,Mark A. Hasegawa-Johnson,Chang D. Yoo
Can Video LLMs Refuse to Answer? Alignment for Answerability in Video Large Language Models
In the broader context of deep learning, Multimodal Large Language Models have achieved significant breakthroughs by leveraging powerful Large Language Models as a backbone to align different modalities into the language space. A prime exemplification is the development of Video Large Language Models (Video-LLMs). While numerous advancements have been proposed to enhance the video understanding capabilities of these models, they are predominantly trained on questions generated directly from video content. However, in real-world scenarios, users often pose questions that extend beyond the informational scope of the video, highlighting the need for Video-LLMs to assess the relevance of the question. We demonstrate that even the best-performing Video-LLMs fail to reject unfit questions-not necessarily due to a lack of video understanding, but because they have not been trained to identify and refuse such questions. To address this limitation, we propose alignment for answerability, a framework that equips Video-LLMs with the ability to evaluate the relevance of a question based on the input video and appropriately decline to answer when the question exceeds the scope of the video, as well as an evaluation framework with a comprehensive set of metrics designed to measure model behavior before and after alignment. Furthermore, we present a pipeline for creating a dataset specifically tailored for alignment for answerability, leveraging existing video-description paired datasets.
****
Dani Valevski,Yaniv Leviathan,Moab Arar,Shlomi Fruchter
Diffusion Models Are Real-Time Game Engines
We present GameNGen, the first game engine powered entirely by a neural model that also enables real-time interaction with a complex environment over long trajectories at high quality. When trained on the classic game DOOM, GameNGen extracts gameplay and uses it to generate a playable environment that can interactively simulate new trajectories. GameNGen runs at 20 frames per second on a single TP

U and remains stable over extended multi-minute play sessions. Next frame predic
tion achieves a PSNR of 29.4, comparable to lossy JPEG compression. Human raters
 are only slightly better than random chance at distinguishing short clips of th
e game from clips of the simulation, even after 5 minutes of auto-regressive gen
eration. GameNGen is trained in two phases: (1) an RL-agent learns to play the g
ame and the training sessions are recorded, and (2) a diffusion model is trained
 to produce the next frame, conditioned on the sequence of past frames and actio
ns. Conditioning augmentations help ensure stable auto-regressive generation ove
r long trajectories, and decoder fine-tuning improves the fidelity of visual det
ails and text.
****

Nguyen Hoang Khoi Do,Truc Nguyen,Malik Hassanaly,raed alharbi,Jung Taek Seo,My T
. Thai
Swift Hydra: Self-Reinforcing Generative Framework for Anomaly Detection with M
ultiple Mamba Models
Despite a plethora of anomaly detection models developed over the years, their a
bility to generalize to unseen anomalies remains an issue, particularly in criti
cal systems. This paper aims to address this challenge by introducing Swift Hydr
a, a new framework for training an anomaly detection method based on generative
AI and reinforcement learning (RL). Through featuring an RL policy that operates
 on the latent variables of a generative model, the framework synthesizes novel
and diverse anomaly samples that are capable of bypassing a detection model. The
se generated synthetic samples are, in turn, used to augment the detection model
, further improving its ability to handle challenging anomalies. Swift Hydra als
o incorporates Mamba models structured as a Mixture of Experts (MoE) to enable s
calable adaptation of the number of Mamba experts based on data complexity, effe
ctively capturing diverse feature distributions without increasing the model’s i
nference time. Empirical evaluations on ADBench benchmark demonstrate that Swift
 Hydra outperforms other state-of-the-art anomaly detection models while mainta
ining a relatively short inference time. From these results, our research highli
ghts a new and auspicious paradigm of integrating RL and generative AI for advan
cing anomaly detection.
****

Yehui Tang,Mabiao Long,Junchi Yan
QuaDiM: A Conditional Diffusion Model For Quantum State Property Estimation
Quantum state property estimation (QPE) is a fundamental challenge in quantum ma
ny-body problems in physics and chemistry, involving the prediction of character
istics such as correlation and entanglement entropy through statistical analysis
 of quantum measurement data. Recent advances in deep learning have provided pow
erful solutions, predominantly using auto-regressive models. These models genera
lly assume an intrinsic ordering among qubits, aiming to approximate the classic
al probability distribution through sequential training. However, unlike natural
 language, the entanglement structure of qubits lacks an inherent ordering, hurt
ing the motivation of such models. In this paper, we introduce a novel, non-auto
regressive generative model called \textbf{\model}, designed for \underline{\tex
tbf{Qua}}ntum state property estimation using \underline{\textbf{Di}}ffusion \un
derline{\textbf{M}}odels. \model progressively denoises Gaussian noise into the
distribution corresponding to the quantum state, encouraging equal, unbiased tre
atment of all qubits. \model learns to map physical variables to properties of t
he ground state of the parameterized Hamiltonian during offline training. Afterw
ards one can sample from the learned distribution conditioned on previously unse
en physical variables to collect measurement records and employ post-processing
to predict properties of unknown quantum states. We evaluate \model on large-sca
le QPE tasks using classically simulated data on the 1D anti-ferromagnetic Heise
nberg model with the system size up to 100 qubits. Numerical results demonstrate
 that \model outperforms baseline models, particularly auto-regressive approache
s, under conditions of limited measurement data during training and reduced samp
le complexity during inference.
****

Xinran Liu,Rocio Diaz Martin,Yikun Bai,Ashkan Shahbazi,Matthew Thorpe,Akram Aldr

oubi,Soheil Kolouri
Expected Sliced Transport Plans
The optimal transport (OT) problem has gained significant traction in modern machine learning for its ability to: (1) provide versatile metrics, such as Wasserstein distances and their variants, and (2) determine optimal couplings between probability measures. To reduce the computational complexity of OT solvers, methods like entropic regularization and sliced optimal transport have been proposed. The sliced OT framework improves efficiency by comparing one-dimensional projections (slices) of high-dimensional distributions. However, despite their computational efficiency, sliced-Wasserstein approaches lack a transportation plan between the input measures, limiting their use in scenarios requiring explicit coupling. In this paper, we address two key questions: Can a transportation plan be constructed between two probability measures using the sliced transport framework? If so, can this plan be used to define a metric between the measures? We propose a ‘lifting’ operation to extend one-dimensional optimal transport plans back to the original space of the measures. By computing the expectation of these lifted plans, we derive a new transportation plan, termed expected sliced transport (EST) plans. We further prove that using the EST plan to weight the sum of the individual Euclidean costs $\|x - y\|^p$ for moving from x to y results in a valid metric between the input discrete probability measures. Finally, we demonstrate the connection between our approach and the recently proposed min-SWGG, along with illustrative numerical examples that support our theoretical findings.
****
Charlie Blake,Constantin Eichenberg,Josef Dean,Lukas Balles,Luke Yuri Prince,Björn Deiseroth,Andres Felipe Cruz-Salinas,Carlo Luschi,Samuel Weinbach,Douglas Orr
u-μP: The Unit-Scaled Maximal Update Parametrization
The Maximal Update Parametrization (μP) aims to make the optimal hyperparameters (HPs) of a model independent of its size, allowing them to be swept using a cheap proxy model rather than the full-size target model. We present a new scheme, u-μP, which improves upon μP by combining it with Unit Scaling, a method for designing models that makes them easy to train in low-precision. The two techniques have a natural affinity: μP ensures that the scale of activations is independent of model size, and Unit Scaling ensures that activations, weights and gradients begin training with a scale of one. This synthesis opens the door to a simpler scheme, whose default values are near-optimal. This in turn facilitates a more efficient sweeping strategy, with u-μP models reaching a lower loss than comparable μP models and working out-of-the-box in FP8.
****
Yair Davidson,Nadav Dym
On the Hölder Stability of Multiset and Graph Neural Networks
Extensive research efforts have been put into characterizing and constructing maximally separating multiset and graph neural networks.
However, recent empirical evidence suggests the notion of separation itself doesn't capture several interesting phenomena. On the one hand, the quality of this separation may be very weak, to the extent that the embeddings of "separable" objects might even be considered identical when using fixed finite precision. On the other hand, architectures which aren't capable of separation in theory, somehow achieve separation when taking the network to be wide enough.

In this work, we address both of these issues, by proposing a novel pair-wise separation quality analysis framework which is based on an adaptation of Lipschitz and Hölder stability to parametric functions. The proposed framework, which we name Hölder in expectation, allows for separation quality analysis, without restricting the analysis to embeddings that can separate all the input space simultaneously. We prove that common sum-based models are lower-Hölder in expectation, with an exponent
that decays rapidly with the network's depth . Our analysis leads to adversarial examples of graphs which can be separated by three 1-WL iterations, but cannot be separated in practice by standard maximally powerful Message Passing Neural Networks (MPNNs). To remedy this, we propose two novel MPNNs with improved sepa

ration quality, one of which is lower Lipschitz in expectation. We show these MP
NNs can easily classify our adversarial examples, and compare favorably with sta
ndard MPNNs on standard graph learning tasks.
****

Weihao Zeng,Yuzhen Huang,Lulu Zhao,Yijun Wang,Zifei Shan,Junxian He
B-STaR: Monitoring and Balancing Exploration and Exploitation in Self-Taught Rea
soners
In the absence of extensive human-annotated data for complex reasoning tasks, se
lf-improvement -- where models are trained on their own outputs -- has emerged a
s a primary method for enhancing performance. Recently, the approach to self-imp
rovement has shifted toward a more dynamic, online fashion through iterative tra
ining processes. However, the critical factors underlying the mechanism of these
 self-improving methods remain poorly understood, such as under what conditions
self-improvement is effective, and what are the bottlenecks in the current itera
tions.
In this work, we identify and propose methods to monitor two pivotal factors in
this iterative process: (1) the model's ability to explore and generate high-qua
lity responses among multiple candidates (exploration); and (2) the reliability
of external rewards in selecting the best responses from the generated outputs (
exploitation).
These factors are inherently moving targets throughout the self-improvement cycl
es, yet their dynamics are rarely discussed in prior research -- It remains uncl
ear what impedes continual model enhancement after only a few iterations.
Using mathematical reasoning as a case study, we begin with a quantitative analy
sis to track the dynamics of exploration and exploitation, discovering that a mo
del's exploratory capabilities rapidly deteriorate over iterations, and the effe
ctiveness of exploiting external rewards diminishes as well due to shifts in dis
tribution from the original policy.
Motivated by these findings, we introduce B-STaR, a Self-Taught Reasoning framew
ork that autonomously adjusts configurations across iterations to Balance explor
ation and exploitation, thereby optimizing the self-teaching effectiveness based
 on the current policy model and available rewards.
Our experiments in mathematical reasoning demonstrate that B-STaR not only enhan
ces the model's exploratory capabilities throughout training but also achieves a
 more effective balance between exploration and exploitation, leading to superio
r performance. Crucially, this work deconstructs the opaque nature of self-train
ing algorithms, elucidating the interpretable dynamics throughout the process an
d highlighting current limitations for future research to address.
****

Wei Guo,Molei Tao,Yongxin Chen
Provable Benefit of Annealed Langevin Monte Carlo for Non-log-concave Sampling
We consider the outstanding problem of sampling from an unnormalized density tha
t may be non-log-concave and multimodal. To enhance the performance of simple Ma
rkov chain Monte Carlo (MCMC) methods, techniques of annealing type have been wi
dely used. However, quantitative theoretical guarantees of these techniques are
under-explored. This study takes a first step toward providing a non-asymptotic
analysis of annealed MCMC. Specifically, we establish, for the first time, an or
acle complexity of $\widetilde{O}\left(\frac{d\beta^2{\cal A}^2}{\varepsilon^6}\
right)$ for the simple annealed Langevin Monte Carlo algorithm to achieve $\vare
psilon^2$ accuracy in Kullback-Leibler divergence to the target distribution $\p
i\propto{\rm e}^{-V}$ on $\mathbb{R}^d$ with β-smooth potential V. Here,
 ${\cal A}$ represents the action of a curve of probability measures interpolati
ng the target distribution π and a readily sampleable distribution.
****

Baoqi Pei,Yifei Huang,Jilan Xu,Guo Chen,Yuping He,Lijin Yang,Yali Wang,Weidi Xie
,Yu Qiao,Fei Wu,Limin Wang
Modeling Fine-Grained Hand-Object Dynamics for Egocentric Video Representation L
earning
In egocentric video understanding, the motion of hands and objects as well as th
eir interactions play a significant role by nature.

However, existing egocentric video representation learning methods mainly focus on aligning video representation with high-level narrations, overlooking the intricate dynamics between hands and objects.
In this work, we aim to integrate the modeling of fine-grained hand-object dynamics into the video representation learning process.
Since no suitable data is available, we introduce HOD, a novel pipeline employing a hand-object detector and a large language model to generate high-quality narrations with detailed descriptions of hand-object dynamics.
To learn these fine-grained dynamics, we propose EgoVideo, a model with a new lightweight motion adapter to capture fine-grained hand-object motion information.

Through our co-training strategy, EgoVideo effectively and efficiently leverages the fine-grained hand-object dynamics in the HOD data.
Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple egocentric downstream tasks, including improvements of 6.3% in EK-100 multi-instance retrieval, 5.7% in EK-100 classification, and 16.3% in EGTEA classification in zero-shot settings. Furthermore, our model exhibits robust generalization capabilities in hand-object interaction and robot manipulation tasks.
****
Botao Ye,Sifei Liu,Haofei Xu,Xueting Li,Marc Pollefeys,Ming-Hsuan Yang,Songyou Peng
No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images
We introduce NoPoSplat, a feed-forward model capable of reconstructing 3D scenes parameterized by 3D Gaussians from unposed sparse multi-view images. Our model, trained exclusively with photometric loss, achieves real-time 3D Gaussian reconstruction during inference. To eliminate the need for accurate pose input during reconstruction, we anchor one input view's local camera coordinates as the canonical space and train the network to predict Gaussian primitives for all views within this space. This approach obviates the need to transform Gaussian primitives from local coordinates into a global coordinate system, thus avoiding errors associated with per-frame Gaussians and pose estimation. To resolve scale ambiguity, we design and compare various intrinsic embedding methods, ultimately opting to convert camera intrinsics into a token embedding and concatenate it with image tokens as input to the model, enabling accurate scene scale prediction. We utilize the reconstructed 3D Gaussians for novel view synthesis and pose estimation tasks and propose a two-stage coarse-to-fine pipeline for accurate pose estimation. Experimental results demonstrate that our pose-free approach can achieve superior novel view synthesis quality compared to pose-required methods, particularly in scenarios with limited input image overlap. For pose estimation, our method, trained without ground truth depth or explicit matching loss, significantly outperforms the state-of-the-art methods with substantial improvements. This work makes significant advances in pose-free generalizable 3D reconstruction and demonstrates its applicability to real-world scenarios. Code and trained models are available at https://noposplat.github.io/.
****
Justin Xu,Jack Gallifant,ALISTAIR JOHNSON,Matthew B.A. McDermott
ACES: Automatic Cohort Extraction System for Event-Stream Datasets
Reproducibility remains a significant challenge in machine learning (ML) for healthcare. Datasets, model pipelines, and even task or cohort definitions are often private in this field, leading to a significant barrier in sharing, iterating, and understanding ML results on electronic health record (EHR) datasets. We address a significant part of this problem by introducing the Automatic Cohort Extraction System (ACES) for event-stream data. This library is designed to simultaneously simplify the development of tasks and cohorts for ML in healthcare and also enable their reproduction, both at an exact level for single datasets and at a conceptual level across datasets. To accomplish this, ACES provides: (1) a highly intuitive and expressive domain-specific configuration language for defining both dataset-specific concepts and dataset-agnostic inclusion or exclusion crit

eria, and (2) a pipeline to automatically extract patient records that meet thes
e defined criteria from real-world data. ACES can be automatically applied to an
y dataset in either the Medical Event Data Standard (MEDS) or Event Stream GPT (
ESGPT) formats, or to *any* dataset in which the necessary task-specific predica
tes can be extracted in an event-stream form. ACES has the potential to signific
antly lower the barrier to entry for defining ML tasks in representation learnin
g, redefine the way researchers interact with EHR datasets, and significantly im
prove the state of reproducibility for ML studies using this modality. ACES is a
vailable at: https://github.com/justin13601/aces.
****

Yushi LAN,Shangchen Zhou,Zhaoyang Lyu,Fangzhou Hong,Shuai Yang,Bo Dai,Xingang Pa
n,Chen Change Loy
GaussianAnything: Interactive Point Cloud Flow Matching for 3D Generation
Recent advancements in diffusion models and large-scale datasets have revolution
ized image and video generation, with increasing focus on 3D content generation.
 While existing methods show promise, they face challenges in input formats, lat
ent space structures, and output representations. This paper introduces a novel
3D generation framework that addresses these issues, enabling scalable and high-
quality 3D generation with an interactive Point Cloud-structured Latent space. O
ur approach utilizes a VAE with multi-view posed RGB-D-N renderings as input, fe
atures a unique latent space design that preserves 3D shape information, and inc
orporates a cascaded latent flow-based model for improved shape-texture disentan
glement. The proposed method, GaussianAnything, supports multi-modal conditional
 3D generation, allowing for point cloud, caption, and single-view image inputs.
 Experimental results demonstrate superior performance on various datasets, adva
ncing the state-of-the-art in 3D content generation.
****

Suqin Yuan,Runqi Lin,Lei Feng,Bo Han,Tongliang Liu
Instance-dependent Early Stopping
In machine learning practice, early stopping has been widely used to regularize
models and can save computational costs by halting the training process when the
 model's performance on a validation set stops improving. However, conventional
early stopping applies the same stopping criterion to all instances without cons
idering their individual learning statuses, which leads to redundant computation
s on instances that are already well-learned. To further improve the efficiency,
 we propose an Instance-dependent Early Stopping (IES) method that adapts the ea
rly stopping mechanism from the entire training set to the instance level, based
 on the core principle that once the model has mastered an instance, the trainin
g on it should stop. IES considers an instance as mastered if the second-order d
ifferences of its loss value remain within a small range around zero. This offer
s a more consistent measure of an instance's learning status compared with direc
tly using the loss value, and thus allows for a unified threshold to determine w
hen an instance can be excluded from further backpropagation. We show that exclu
ding mastered instances from backpropagation can increase the gradient norms, th
ereby accelerating the decrease of the training loss and speeding up the trainin
g process. Extensive experiments on benchmarks demonstrate that IES method can r
educe backpropagation instances by 10%-50% while maintaining or even slightly im
proving the test accuracy and transfer learning performance of a model.
****

Zhepei Wei,Wei-Lin Chen,Yu Meng
InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rat
ionales
Retrieval-augmented generation (RAG) has shown promising potential to enhance th
e accuracy and factuality of language models (LMs). However, imperfect retriever
s or noisy corpora can introduce misleading or even erroneous information to the
 retrieved contents, posing a significant challenge to the generation quality. E
xisting RAG methods typically address this challenge by directly predicting fina
l answers despite potentially noisy inputs, resulting in an implicit denoising p
rocess that is difficult to interpret and verify. On the other hand, the acquisi
tion of explicit denoising supervision is often costly, involving significant hu

man efforts. In this work, we propose InstructRAG, where LMs explicitly learn th
e denoising process through self-synthesized rationales --- First, we instruct t
he LM to explain how the ground-truth answer is derived from retrieved documents
. Then, these rationales can be used either as demonstrations for in-context lea
rning of explicit denoising or as supervised fine-tuning data to train the model
. Compared to standard RAG approaches, InstructRAG requires no additional superv
ision, allows for easier verification of the predicted answers, and effectively
improves generation accuracy. Experiments show InstructRAG consistently outperfo
rms existing RAG methods in both training-free and trainable scenarios, achievin
g a relative improvement of 8.3% over the best baseline method on average across
 five knowledge-intensive benchmarks. Extensive analysis indicates that Instruct
RAG scales well with increased numbers of retrieved documents and consistently e
xhibits robust denoising ability even in out-of-domain datasets, demonstrating s
trong generalizability.
****
Ziming Liu,Yixuan Wang,Sachin Vaidya,Fabian Ruehle,James Halverson,Marin Soljaci
c,Thomas Y. Hou,Max Tegmark
KAN: Kolmogorov–Arnold Networks
Inspired by the Kolmogorov-Arnold representation theorem, we propose Kolmogorov-
Arnold Networks (KANs) as promising alternatives to Multi-Layer Perceptrons (MLP
s). While MLPs have fixed activation functions on nodes ("neurons''), KANs have
learnable activation functions on edges ("weights''). KANs have no linear weight
s at all -- every weight parameter is replaced by a univariate function parametr
ized as a spline. We show that this seemingly simple change makes KANs outperfor
m MLPs in terms of accuracy and interpretability, on small-scale AI + Science ta
sks. For accuracy, smaller KANs can achieve comparable or better accuracy than l
arger MLPs in function fitting tasks. Theoretically and empirically, KANs posses
s faster neural scaling laws than MLPs. For interpretability, KANs can be intuit
ively visualized and can easily interact with human users. Through two examples
in mathematics and physics, KANs are shown to be useful ``collaborators'' helpin
g scientists (re)discover mathematical and physical laws. In summary, KANs are p
romising alternatives for MLPs. Despite the slow training of KANs, their improve
d accuracy and interpretability show the potential to improve today's deep learn
ing models which rely heavily on MLPs. More research is necessary to make KANs'
training more efficient.
****
Jiarui Lu,Xiaoyin Chen,Stephen Zhewen Lu,Chence Shi,Hongyu Guo,Yoshua Bengio,Jia
n Tang
Structure Language Models for Protein Conformation Generation
Proteins adopt multiple structural conformations to perform their diverse biolog
ical functions, and understanding these conformations is crucial for advancing d
rug discovery. Traditional physics-based simulation methods often struggle with
sampling equilibrium conformations and are computationally expensive. Recently,
deep generative models have shown promise in generating protein conformations as
 a more efficient alternative. However, these methods predominantly rely on the
diffusion process within a 3D geometric space, which typically centers around th
e vicinity of metastable states and is often inefficient in terms of runtime. In
 this paper, we introduce Structure Language Modeling (SLM) as a novel framework
 for efficient protein conformation generation. Specifically, the protein struct
ures are first encoded into a compact latent space using a discrete variational
auto-encoder, followed by conditional language modeling that effectively capture
s sequence-specific conformation distributions. This enables a more efficient a
nd interpretable exploration of diverse ensemble modes compared to existing meth
ods. Based on this general framework, we instantiate SLM with various popular LM
 architectures as well as proposing the ESMDiff, a novel BERT-like structure lan
guage model fine-tuned from ESM3 with masked diffusion. We verify our approach i
n various scenarios, including the equilibrium dynamics of BPTI, conformational
change pairs, and intrinsically disordered proteins. SLM provides a highly effic
ient solution, offering a 20-100x speedup than existing methods in generating di
verse conformations, shedding light on promising avenues for future research.

****
Lingwei Zhu,Haseeb Shah,Han Wang,Yukie Nagai,Martha White
q-exponential family for policy optimization
Policy optimization methods benefit from a simple and tractable policy parametrization, usually the Gaussian for continuous action spaces. In this paper, we consider a broader policy family that remains tractable: the q-exponential family.
This family of policies is flexible, allowing the specification of both heavy-tailed policies ($q>1$) and light-tailed policies ($q<1$). This paper examines the interplay between q-exponential policies for several actor-critic algorithms conducted on both online and offline problems. We find that heavy-tailed policies are more effective in general and can consistently improve on Gaussian.
In particular, we find the Student's t-distribution to be more stable than the Gaussian across settings and that a heavy-tailed q-Gaussian for Tsallis Advantage Weighted Actor-Critic consistently performs well in offline benchmark problems.
In summary, we find that the Student's t policy a strong candidate for drop-in replacement to the Gaussian.
Our code is available at \url{https://github.com/lingweizhu/qexp}.
****
Peiye Zhuang,Songfang Han,Chaoyang Wang,Aliaksandr Siarohin,Jiaxu Zou,Michael Vasilkovsky,Vladislav Shakhrai,Sergei Korolev,Sergey Tulyakov,Hsin-Ying Lee
GTR: Improving Large 3D Reconstruction Models through Geometry and Texture Refinement
We propose a novel approach for 3D mesh reconstruction from multi-view images. We improve upon the large reconstruction model LRM that use a transformer-based triplane generator and a Neural Radiance Field (NeRF) model trained on multi-view images. We introduce three key components to significantly enhance the 3D reconstruction quality. First of all, we examine the original LRM architecture and find several shortcomings. Subsequently, we introduce respective modifications to the LRM architecture, which lead to improved multi-view image representation and more computationally efficient training. Second, in order to improve geometry reconstruction and enable supervision at full image resolution, we extract meshes from the NeRF in a differentiable manner and fine-tune the NeRF model through mesh rendering. These modifications allow us to achieve state-of-the-art performance on both 2D and 3D evaluation metrics on Google Scanned Objects (GSO) dataset and OmniObject3D dataset. Finally, we introduce a lightweight per-instance texture refinement procedure to better reconstruct complex textures, such as text and portraits on assets. To address this, we introduce a lightweight per-instance texture refinement procedure. This procedure fine-tunes the triplane representation and the NeRF's color estimation model on the mesh surface using the input multi-view images in just 4 seconds. This refinement achieves faithful reconstruction of complex textures. Additionally, our approach enables various downstream applications, including text/image-to-3D generation.
****
Kanchana Ranasinghe,Xiang Li,Kumara Kahatapitiya,Michael S Ryoo
Understanding Long Videos with Multimodal Language Models
Large Language Models (LLMs) have allowed recent LLM-based approaches to achieve excellent performance on long-video understanding benchmarks. We investigate how extensive world knowledge and strong reasoning skills of underlying LLMs influence this strong performance. Surprisingly, we discover that LLM-based approaches can yield surprisingly good accuracy on long-video tasks with limited video information, sometimes even with no video-specific information. Building on this, we explore injecting video-specific information into an LLM-based framework. We utilize off-the-shelf vision tools to extract three object-centric information modalities from videos, and then leverage natural language as a medium for fusing this information. Our resulting Multimodal Video Understanding (MVU) framework demonstrates state-of-the-art performance across multiple video understanding benchmarks. Strong performance also on robotics domain tasks establishes its strong generality. Code: github.com/kahnchana/mvu

****
Maximilian Weiherer,Bernhard Egger
Matérn Kernels for Tunable Implicit Surface Reconstruction
We propose to use the family of Matérn kernels for implicit surface reconstructi
on, building upon the recent success of kernel methods for 3D reconstruction of
oriented point clouds. As we show from a theoretical and practical perspective,
Matérn kernels have some appealing properties which make them particularly well
suited for surface reconstruction---outperforming state-of-the-art methods based
 on the arc-cosine kernel while being significantly easier to implement, faster
to compute, and scalable. Being stationary, we demonstrate that Matérn kernels a
llow for tunable surface reconstruction in the same way as Fourier feature mappi
ngs help coordinate-based MLPs overcome spectral bias. Moreover, we theoreticall
y analyze Matérn kernels' connection to SIREN networks as well as their relation
 to previously employed arc-cosine kernels. Finally, based on recently introduce
d Neural Kernel Fields, we present data-dependent Matérn kernels and conclude th
at especially the Laplace kernel (being part of the Matérn family) is extremely
competitive, performing almost on par with state-of-the-art methods in the noise
-free case while having a more than five times shorter training time.
****
Daning Huang,Hanyang He,John Harlim,Yan Li
Learning vector fields of differential equations on manifolds with geometrically
 constrained operator-valued kernels
We address the problem of learning ordinary differential equations (ODEs) on man
ifolds. Existing machine learning methods, particularly those using neural netwo
rks, often struggle with high computational demands. To overcome this issue, we
introduce a geometrically constrained operator-valued kernel that allows us to r
epresent vector fields on tangent bundles of smooth manifolds. The construction
of the kernel imposes the geometric constraints that are estimated from the data
 and ensures the computational feasibility for learning high dimensional systems
 of ODEs. Once the vector fields are estimated, e.g., by the kernel ridge regres
sion, we need an ODE solver that guarantees the solution to stay on (or close to
) the manifold. To overcome this issue, we propose a geometry-preserving ODE sol
ver that approximates the exponential maps corresponding to the ODE solutions.
We deduce a theoretical error bound for the proposed solver that guarantees the
approximate solutions to lie on the manifold in the limit of large data. We veri
fy the effectiveness of the proposed approach on high-dimensional dynamical syst
ems, including the cavity flow problem, the beating and travelling waves in Kura
moto-Sivashinsky equations, and the reaction-diffusion dynamics.
****
Changlong Wu,Ananth Grama,Wojciech Szpankowski
No Free Lunch: Fundamental Limits of Learning Non-Hallucinating Generative Model
s
Generative models have shown impressive capabilities in synthesizing high-qualit
y outputs across various domains. However, a persistent challenge is the occurre
nce of "hallucinations," where the model produces outputs that are not grounded
in the underlying facts. While empirical strategies have been explored to mitiga
te this issue, a rigorous theoretical understanding remains elusive. In this pap
er, we develop a theoretical framework to analyze the *learnability* of non-hall
ucinating generative models from a learning-theoretic perspective. Our results r
eveal that non-hallucinating learning is statistically *impossible* when relying
 solely on the training dataset, even for a hypothesis class of size two and whe
n the entire training set is truthful. To overcome these limitations, we show th
at incorporating *inductive biases* aligned with the actual facts into the learn
ing process is essential. We provide a systematic approach to achieve this by re
stricting the fact set to a concept class of finite VC-dimension and demonstrate
 its effectiveness under various learning paradigms. Although our findings are p
rimarily conceptual, they represent a first step towards a principled approach t
o addressing hallucinations in learning generative models.
****
Tianzhu Ye,Li Dong,Yuqing Xia,Yutao Sun,Yi Zhu,Gao Huang,Furu Wei

Differential Transformer
Transformer tends to overallocate attention to irrelevant context. In this work, we introduce Diff Transformer, which amplifies attention to the relevant context while canceling noise. Specifically, the differential attention mechanism calculates attention scores as the difference between two separate softmax attention maps. The subtraction cancels noise, promoting the emergence of sparse attention patterns. Experimental results on language modeling show that Diff Transformer outperforms Transformer in various settings of scaling up model size and training tokens. More intriguingly, it offers notable advantages in practical applications, such as long-context modeling, key information retrieval, hallucination mitigation, in-context learning, and reduction of activation outliers. By being less distracted by irrelevant context, Diff Transformer can mitigate hallucination in question answering and text summarization. For in-context learning, Diff Transformer not only enhances accuracy but is also more robust to order permutation, which was considered as a chronic robustness issue. The results position Diff Transformer as a highly effective and promising architecture for large language models.
****
Tommaso Cesari,Roberto Colomboni
An Online Learning Theory of Trading-Volume Maximization
We explore brokerage between traders in an online learning framework.
At any round t, two traders meet to exchange an asset, provided the exchange is mutually beneficial.
The broker proposes a trading price, and each trader tries to sell their asset or buy the asset from the other party, depending on whether the price is higher or lower than their private valuations.
A trade happens if one trader is willing to sell and the other is willing to buy at the proposed price.
Previous work provided guidance to a broker aiming at enhancing traders' total earnings by maximizing the *gain from trade*, defined as the sum of the traders' net utilities after each interaction.
This classical notion of reward can be highly unfair to traders with small profit margins, and far from the real-life utility of the broker.
For these reasons, we investigate how the broker should behave to maximize the trading volume, i.e., the *total number of trades*.
We model the traders' valuations as an i.i.d. process with an unknown distribution.
If the traders' valuations are revealed after each interaction (full-feedback), and the traders' valuations cumulative distribution function (cdf) is continuous, we provide an algorithm achieving logarithmic regret and show its optimality up to constants.
If only their willingness to sell or buy at the proposed price is revealed after each interaction (2-bit feedback), we provide an algorithm achieving poly-logarithmic regret when the traders' valuations cdf is Lipschitz and show its near-optimality.
We complement our results by analyzing the implications of dropping the regularity assumptions on the unknown traders' valuations cdf.
If we drop the continuous cdf assumption, the regret rate degrades to $\Theta(\sqrt{T})$ in the full-feedback case, where T is the time horizon.
If we drop the Lipschitz cdf assumption, learning becomes impossible in the 2-bit feedback case.
****
Shu Yu,Chaochao Lu
ADAM: An Embodied Causal Agent in Open-World Environments
In open-world environments like Minecraft, existing agents face challenges in continuously learning structured knowledge, particularly causality. These challenges stem from the opacity inherent in black-box models and an excessive reliance on prior knowledge during training, which impair their interpretability and generalization capability. To this end, we introduce ADAM, An emboDied causal Agent in Minecraft, which can autonomously navigate the open world, perceive multimoda

l context, learn causal world knowledge, and tackle complex tasks through lifelo
ng learning. ADAM is empowered by four key components: 1) an interaction module,
 enabling the agent to execute actions while recording the interaction processes
; 2) a causal model module, tasked with constructing an ever-growing causal grap
h from scratch, which enhances interpretability and reduces reliance on prior kn
owledge; 3) a controller module, comprising a planner, an actor, and a memory po
ol, using the learned causal graph to accomplish tasks; 4) a perception module,
powered by multimodal large language models, enabling ADAM to perceive like a hu
man player. Extensive experiments show that ADAM constructs a nearly perfect cau
sal graph from scratch, enabling efficient task decomposition and execution with
 strong interpretability. Notably, in the modified Minecraft game where no prior
 knowledge is available, ADAM excels with remarkable robustness and generalizati
on capability. ADAM pioneers a novel paradigm that integrates causal methods and
 embodied agents synergistically. Our project page is at https://opencausalab.gi
thub.io/ADAM.
****

Emmanuel Abbe,Elisabetta Cornacchia,Jan H■z■a,Donald Kougang-Yombi
Learning High-Degree Parities: The Crucial Role of the Initialization
Parities have become a standard benchmark for evaluating learning algorithms. Re
cent works show that regular neural networks trained by gradient descent can eff
iciently learn degree k parities on uniform inputs for constant k, but fail
to do so when k and $d-k$ grow with d (here d is the ambient dimension). H
owever, the case where $k=d-O_d(1)$, including the degree d parity (the \texti
t{full parity}), has remained unsettled. This paper shows that for gradient desc
ent on regular neural networks, learnability depends on the initial weight distr
ibution. On one hand, the discrete Rademacher initialization enables efficient l
earning of full parities, while on the other hand, its Gaussian perturbation wit
h large enough constant standard deviation σ prevents it. The positive re
sult for full parity is shown to hold up to $\sigma=O(d^{-1})$, pointing to ques
tions about a sharper threshold phenomenon. Unlike statistical query (SQ) learni
ng, where a singleton function class like the full parity is trivially learnabl
e, our negative result applies to a fixed function and relies on an \emph{initia
l gradient alignment} measure of potential broader relevance to neural networks
learning.
****

Shengjie Zhou,Xin Cheng,Haiyang Xu,Ming Yan,Tao Xiang,Feng Liu,Lei Feng
Endowing Visual Reprogramming with Adversarial Robustness
Visual reprogramming (VR) leverages well-developed pre-trained models (e.g., a p
re-trained classifier on ImageNet) to tackle target tasks (e.g., a traffic sign
recognition task), without the need for training from scratch. Despite the effec
tiveness of previous VR methods, all of them did not consider the adversarial ro
bustness of reprogrammed models against adversarial attacks, which could lead to
 unpredictable problems in safety-crucial target tasks. In this paper, we empiri
cally find that reprogramming pre-trained models with adversarial robustness and
 incorporating adversarial samples from the target task during reprogramming can
 both improve the adversarial robustness of reprogrammed models. Furthermore, we
 propose a theoretically guaranteed adversarial robustness risk upper bound for
VR, which validates our empirical findings and could provide a theoretical found
ation for future research. Extensive experiments demonstrate that by adopting th
e strategies revealed in our empirical findings, the adversarial robustness of r
eprogrammed models can be enhanced.
****

Yun-Wei Chu,Dong-Jun Han,Seyyedali Hosseinalipour,Christopher Brinton
Unlocking the Potential of Model Calibration in Federated Learning
Over the past several years, various federated learning (FL) methodologies have
been developed to improve model accuracy, a primary performance metric in machin
e learning. However, to utilize FL in practical decision-making scenarios, beyon
d considering accuracy, the trained model must also have a reliable confidence i
n each of its predictions, an aspect that has been largely overlooked in existin
g FL research. Motivated by this gap, we propose Non-Uniform Calibration for Fed

erated Learning (NUCFL), a generic framework that integrates FL with the concept of model calibration. The inherent data heterogeneity in FL environments makes model calibration particularly difficult, as it must ensure reliability across diverse data distributions and client conditions. Our NUCFL addresses this challenge by dynamically adjusting the model calibration objectives based on statistical relationships between each client's local model and the global model in FL. In particular, NUCFL assesses the similarity between local and global model relationships, and controls the penalty term for the calibration loss during client-side local training. By doing so, NUCFL effectively aligns calibration needs for the global model in heterogeneous FL settings while not sacrificing accuracy. Extensive experiments show that NUCFL offers flexibility and effectiveness across various FL algorithms, enhancing accuracy as well as model calibration.
****

Ruichen Shao,Bei Li,Gangao Liu,Yang Chen,ZhouXiang,Jingang Wang,Xunliang Cai,Peng Li

Earlier Tokens Contribute More: Learning Direct Preference Optimization From Temporal Decay Perspective

Direct Preference Optimization (DPO) has gained attention as an efficient alternative to reinforcement learning from human feedback (RLHF) for aligning large language models (LLMs) with human preferences. Despite its advantages, DPO suffers from a length bias, generating responses longer than those from the reference model. Existing solutions like SimPO and SamPO address this issue but uniformly treat the contribution of rewards across sequences, overlooking temporal dynamics. To this end, we propose an enhanced preference optimization method that incorporates a temporal decay factor controlled by a gamma parameter. This dynamic weighting mechanism adjusts the influence of each reward based on its position in the sequence, prioritizing earlier tokens that are more critical for alignment. By adaptively focusing on more relevant feedback, our approach mitigates overfitting to less pertinent data and remains responsive to evolving human preferences. Experimental results on several benchmarks show that our approach consistently outperforms vanilla DPO by 5.9-8.8 points on AlpacaEval 2 and 3.3-9.7 points on Arena-Hard across different model architectures and sizes.
Furthermore, additional experiments on mathematical and reasoning benchmarks (MMLU, GSM8K, and MATH) confirm that our method enhances performance without compromising general capabilities. Our codebase would be available at \url{https://github.com/LotuSrc/D2PO}.
****

Chang Ma,Haiteng Zhao,Junlei Zhang,Junxian He,Lingpeng Kong

Non-myopic Generation of Language Models for Reasoning and Planning

Large Language Models (LLMs) have demonstrated remarkable abilities in reasoning and planning. Despite their success in various domains, such as mathematical problem-solving and coding, LLMs face challenges in ensuring reliable and optimal planning due to the inherent myopic nature of autoregressive decoding. This paper revisits LLM reasoning from an optimal control perspective, proposing a novel method, Predictive-Decoding, that leverages Model Predictive Control to enhance planning accuracy. By reweighting LLM distributions based on foresight trajectories, Predictive-Decoding aims to mitigate early errors and promote non-myopic planning. Our experiments show significant improvements across a wide range of tasks in math, coding, and agent-based scenarios. Furthermore, Predictive-Decoding demonstrates computational efficiency, outperforming search baselines while utilizing inference compute more effectively. This study provides insights into optimizing LLM planning capabilities.
****

Kim Yong Tan,Yueming Lyu,Ivor Tsang,Yew-Soon Ong

Fast Direct: Query-Efficient Online Black-box Guidance for Diffusion-model Target Generation

Guided diffusion-model generation is a promising direction for customizing the generation process of a pre-trained diffusion-model to address the specific downstream tasks. Existing guided diffusion models either rely on training of the guidance model with pre-collected datasets or require the objective functions to be

differentiable. However, for most real-world tasks, the offline datasets are often unavailable, and their objective functions are often not differentiable, such as image generation with human preferences, molecular generation for drug discovery, and material design. Thus, we need an **online** algorithm capable of collecting data during runtime and supporting a **black-box** objective function. Moreover, the **query efficiency** of the algorithm is also critical because the objective evaluation of the query is often expensive in the real-world scenarios. In this work, we propose a novel and simple algorithm, **Fast Direct**, for query-efficient online black-box target generation. Our Fast Direct builds a pseudo-target on the data manifold to update the noise sequence of the diffusion model with a universal direction, which is promising to perform query-efficient guided generation. Extensive experiments on twelve high-resolution ($\small {1024 \times 1024}$) image target generation tasks and six 3D-molecule target generation tasks show $\textbf{6}\times$ up to $\textbf{10}\times$ query efficiency improvement and $\textbf{11}\times$ up to $\textbf{44}\times$ query efficiency improvement, respectively.

****

Jarrid Rector-Brooks,Mohsin Hasan,Zhangzhi Peng,Cheng-Hao Liu,Sarthak Mittal,Nouha Dziri,Michael M. Bronstein,Pranam Chatterjee,Alexander Tong,Joey Bose

Steering Masked Discrete Diffusion Models via Discrete Denoising Posterior Prediction

Generative modeling of discrete data underlies important applications spanning text-based agents like ChatGPT to the design of the very building blocks of life in protein sequences. However, application domains need to exert control over the generated data by steering the generative process—typically via RLHF—to satisfy a specified property, reward, or affinity metric. In this paper, we study the problem of steering Masked Diffusion Models (MDMs), a recent class of discrete diffusion models that offer a compelling alternative to traditional autoregressive models. We introduce Discrete Denoising Posterior Prediction (DDPP), a novel framework that casts the task of steering pretrained MDMs as a problem of probabilistic inference by learning to sample from a target Bayesian posterior. Our DDPP framework leads to a family of three novel objectives that are all simulation-free, and thus scalable while applying to general non-differentiable reward functions. Empirically, we instantiate DDPP by steering MDMs to perform class-conditional pixel-level image modeling, RLHF-based alignment of MDMs using text based rewards, and finetuning protein language models to generate more diverse secondary structures and shorter proteins. We substantiate our designs via wet-lab validation, where we observe transient expression of reward-optimized protein sequences.

****

Kevin Frans,Danijar Hafner,Sergey Levine,Pieter Abbeel

One Step Diffusion via Shortcut Models

Diffusion models and flow matching models have enabled generating diverse and realistic images by learning to transfer noise to data. However, sampling from these models involves iterative denoising over many neural network passes, making generation slow and expensive. Previous approaches for speeding up sampling require complex training regimes, such as multiple training phases, multiple networks, or fragile scheduling. We introduce Shortcut Models, a family of generative models that use a single network and training phase to produce high-quality samples in a single or multiple sampling steps. Shortcut models condition the network not only on the current noise level but also on the desired step size, allowing the model to skip ahead in the generation process. Across a wide range of sampling step budgets, shortcut models consistently produce higher quality samples than previous approaches, such as consistency models and reflow. Compared to distillation, shortcut models reduce complexity to a single network and training phase and additionally allow varying step budgets at inference time.

****

Jianqun Zhou,Yuanlei Zheng,Wei Chen,Qianqian Zheng,Shang Zeyuan,Wei Zhang,Rui Meng,Xiaoyu Shen

Beyond Content Relevance: Evaluating Instruction Following in Retrieval Models

Instruction-following capabilities in large language models (LLMs) have progress
ed significantly, enabling more complex user interactions through detailed promp
ts. However, retrieval systems have not matched these advances, most of them sti
ll relies on traditional lexical and semantic matching techniques that fail to f
ully capture user intent. Recent efforts have introduced instruction-aware retri
eval models, but these primarily focus on intrinsic content relevance, which neg
lects the importance of customized preferences for broader document-level attrib
utes. This study evaluates the instruction-following capabilities of various ret
rieval models beyond content relevance, including LLM-based dense retrieval and
reranking models. We develop InfoSearch, a novel retrieval evaluation benchmark
spanning six document-level attributes: Audience, Keyword, Format, Language, Len
gth, and Source, and introduce novel metrics -- Strict Instruction Compliance Ra
tio (SICR) and Weighted Instruction Sensitivity Evaluation (WISE) to accurately
assess the models' responsiveness to instructions. Our findings indicate that a
lthough fine-tuning models on instruction-aware retrieval datasets and increasin
g model size enhance performance, most models still fall short of instruction co
mpliance. We release our dataset and code on https://github.com/EIT-NLP/InfoSear
ch.
****

Chinmaya Kausik,Mirco Mutti,Aldo Pacchiano,Ambuj Tewari

A Theoretical Framework for Partially-Observed Reward States in RLHF

The growing deployment of reinforcement learning from human feedback (RLHF) call
s for a deeper theoretical investigation of its underlying models. The prevalent
 models of RLHF do not account for neuroscience-backed, partially-observed "inte
rnal states'' that can affect human feedback, nor do they accommodate intermedia
te feedback during an interaction. Both of these can be instrumental in speeding
 up learning and improving alignment. To address these limitations, we model RLH
F as reinforcement learning with partially observed reward-states (PORRL). We ac
commodate two kinds of feedback — cardinal and dueling feedback. We first
demonstrate that PORRL subsumes a wide class of RL problems, including tradition
al RL, RLHF, and reward machines. For cardinal feedback, we present two model-ba
sed methods (POR-UCRL, POR-UCBVI). We give both cardinal regret and sample compl
exity guarantees for the methods, showing that they improve over naive history-s
ummarization. We then discuss the benefits of a model-free method like GOLF with
 naive history-summarization in settings with recursive internal states and dens
e intermediate feedback. For this purpose, we define a new history aware version
 of the Bellman-eluder dimension and give a new guarantee for GOLF in our settin
g, which can be exponentially sharper in illustrative examples. For dueling feed
back, we show that a naive reduction to cardinal feedback fails to achieve subli
near dueling regret. We then present the first explicit reduction that converts
guarantees for cardinal regret to dueling regret. In both feedback settings, we
show that our models and guarantees generalize and extend existing ones.
****

Hoang V. Tran,Minh-Khoi Nguyen-Nhat,Huyen Trang Pham,Thanh Chu,Tam Le,Tan Minh N
guyen

Distance-Based Tree-Sliced Wasserstein Distance

To overcome computational challenges of Optimal Transport (OT), several variants
 of Sliced Wasserstein (SW) has been developed in the literature. These approach
es exploit the closed-form expression of the univariate OT by projecting measure
s onto one-dimensional lines. However, projecting measures onto low-dimensional
spaces can lead to a loss of topological information. Tree-Sliced Wasserstein di
stance on Systems of Lines (TSW-SL) has emerged as a promising alternative that
replaces these lines with a more intricate structure called tree systems. The tr
ee structures enhance the ability to capture topological information of the metr
ic while preserving computational efficiency. However, at the core of TSW-SL, th
e splitting maps, which serve as the mechanism for pushing forward measures onto
 tree systems, focus solely on the position of the measure supports while disreg
arding the projecting domains. Moreover, the specific splitting map used in TSW-
SL leads to a metric that is not invariant under Euclidean transformations, a ty
pically expected property for OT on Euclidean space. In this work, we propose a

novel class of splitting maps that generalizes the existing one studied in TSW-S
L enabling the use of all positional information from input measures, resulting
in a novel Distance-based Tree-Sliced Wasserstein (*Db-TSW*) distance. In additi
on, we introduce a simple tree sampling process better suited for Db-TSW, leadin
g to an efficient GPU-friendly implementation for tree systems, similar to the o
riginal SW. We also provide a comprehensive theoretical analysis of proposed cla
ss of splitting maps to verify the injectivity of the corresponding Radon Transf
orm, and demonstrate that Db-TSW is an Euclidean invariant metric. We empiricall
y show that Db-TSW significantly improves accuracy compared to recent SW variant
s while maintaining low computational cost via a wide range of experiments on gr
adient flows, image style transfer, and generative models.
****
Bruce W. Lee,Inkit Padhi,Karthikeyan Natesan Ramamurthy,Erik Miehling,Pierre Dog
nin,Manish Nagireddy,Amit Dhurandhar
Programming Refusal with Conditional Activation Steering
LLMs have shown remarkable capabilities, but precisely controlling their respons
e behavior remains challenging.
Existing activation steering methods alter LLM behavior indiscriminately, limiti
ng their practical applicability in settings where selective responses are essen
tial, such as content moderation or domain-specific assistants.
In this paper, we propose Conditional Activation Steering (CAST), which analyzes
 LLM activation patterns during inference to selectively apply or withhold activ
ation steering based on the input context.
Our method is based on the observation that different categories of prompts acti
vate distinct patterns in the model's hidden states.
Using CAST, one can systematically control LLM behavior with rules like "if inpu
t is about hate speech or adult content, then refuse" or "if input is not about
legal advice, then refuse."
This allows for selective modification of responses to specific content while ma
intaining normal responses to other content, all without requiring weight optimi
zation.
We release an open-source implementation of our framework.
****
Chenrui Tie,Yue Chen,Ruihai Wu,Boxuan Dong,Zeyi Li,Chongkai Gao,Hao Dong
ET-SEED: EFFICIENT TRAJECTORY-LEVEL SE(3) EQUIVARIANT DIFFUSION POLICY
Imitation learning, e.g., diffusion policy, has been proven effective in various
 robotic manipulation tasks.
However, extensive demonstrations are required for policy robustness and general
ization.
To reduce the demonstration reliance, we leverage spatial symmetry and propose E
T-SEED, an efficient trajectory-level SE(3) equivariant diffusion model for gene
rating action sequences in complex robot manipulation tasks.
Further, previous equivariant diffusion models require the per-step equivariance
 in the Markov process, making it difficult to learn policy under such strong co
nstraints.
We theoretically extend equivariant Markov kernels and simplify the condition of
 equivariant diffusion process, thereby significantly improving training efficie
ncy for trajectory-level SE(3) equivariant diffusion policy in an end-to-end man
ner.
We evaluate ET-SEED on representative robotic manipulation tasks, involving rigi
d body, articulated and deformable object.
Experiments demonstrate superior data efficiency and manipulation proficiency of
 our proposed method,
as well as its ability to generalize to unseen configurations with only a few de
monstrations. Website: https://et-seed.github.io/
****
Yui Oka,Taku Hasegawa,Kyosuke Nishida,Kuniko Saito
Wavelet-based Positional Representation for Long Context
In the realm of large-scale language models, a significant challenge arises when
 extrapolating sequences beyond the maximum allowable length.

This is because the model's position embedding mechanisms are limited to positio
ns encountered during training, thus preventing effective representation of posi
tions in longer sequences.
We analyzed conventional position encoding methods for long contexts and found t
he following characteristics.
(1) When the representation dimension is regarded as the time axis, Rotary Posit
ion Embedding (RoPE) can be interpreted as a restricted wavelet transform using
Haar-like wavelets.
However, because it uses only a fixed scale parameter, it does not fully exploit
 the advantages of wavelet transforms, which capture the fine movements of non-s
tationary signals using multiple scales (window sizes).
This limitation could explain why RoPE performs poorly in extrapolation.
(2)
Previous research as well as our own analysis indicates that Attention with Line
ar Biases (ALiBi) functions similarly to windowed attention, using windows of va
rying sizes.
However, it has limitations in capturing deep dependencies because it restricts
the receptive field of the model.
From these insights, we propose a new position representation method that captur
es multiple scales (i.e., window sizes) by leveraging wavelet transforms without
 limiting the model's attention field.
Experimental results show that this new method improves the performance of the m
odel in both short and long contexts.
In particular, our method allows extrapolation of position information without l
imiting the model's attention field.
****
Gonzalo Gonzalez-Pumariega,Leong Su Yean,Neha Sunkara,Sanjiban Choudhury
Robotouille: An Asynchronous Planning Benchmark for LLM Agents
Effective asynchronous planning, or the ability to efficiently reason and plan o
ver states and actions that must happen in parallel or sequentially, is essentia
l for agents that must account for time delays, reason over diverse long-horizon
 tasks, and collaborate with other agents. While large language model (LLM) agen
ts show promise in high-level task planning, current benchmarks focus primarily
on short-horizon tasks and do not evaluate such asynchronous planning capabiliti
es. We introduce Robotouille, a challenging benchmark environment designed to te
st LLM agents' ability to handle long-horizon asynchronous scenarios. Our synchr
onous and asynchronous datasets capture increasingly complex planning challenges
 that go beyond existing benchmarks, requiring agents to manage over-
lapping tasks and interruptions Our results show that ReAct (gpt-4o) achieves 47
% on synchronous tasks but only 11% on asynchronous tasks, highlighting signific
ant room for improvement. We further analyze failure modes, demonstrating the ne
ed for LLM agents to better incorporate long-horizon feedback and self-audit the
ir reasoning during task execution.
****
Guangtao Lyu,Chenghao Xu,Jiexi Yan,Muli Yang,Cheng Deng
Towards Unified Human Motion-Language Understanding via Sparse Interpretable Cha
racterization
Recently, the comprehensive understanding of human motion has been a prominent a
rea of research due to its critical importance in many fields. However, existing
 methods often prioritize specific downstream tasks and roughly align text and m
otion features within a CLIP-like framework. This results in a lack of rich sema
ntic information which restricts a more profound comprehension of human motions,
 ultimately leading to unsatisfactory performance.
Therefore, we propose a novel motion-language representation paradigm to enhance
 the interpretability of motion representations by constructing a universal moti
on-language space, where both motion and text features are concretely lexicalize
d, ensuring that each element of features carries specific semantic meaning.
Specifically, we introduce a multi-phase strategy mainly comprising Lexical Bott
lenecked Masked Language Modeling to enhance the language model's focus on high-
entropy words crucial for motion semantics, Contrastive Masked Motion Modeling t

o strengthen motion feature extraction by capturing spatiotemporal dynamics dire
ctly from skeletal motion, Lexical Bottlenecked Masked Motion Modeling to enable
 the motion model to capture the underlying semantic features of motion for impr
oved cross-modal understanding, and Lexical Contrastive Motion-Language Pretrain
ing to align motion and text lexicon representations, thereby ensuring enhanced
cross-modal coherence.
Comprehensive analyses and extensive experiments across multiple public datasets
 demonstrate that our model achieves state-of-the-art performance across various
 tasks and scenarios.
****
Xunhao Lai,Jianqiao Lu,Yao Luo,Yiyuan Ma,Xun Zhou
FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Seque
nce Inference
Large language models (LLMs) encounter computational challenges during long-sequ
ence inference, especially in the attention pre-filling phase, where the complex
ity grows quadratically with the prompt length. Previous efforts to mitigate the
se challenges have relied on fixed sparse attention patterns or identifying spar
se attention patterns based on limited cases. However, these methods lacked the
flexibility to efficiently adapt to varying input demands. In this paper, we int
roduce FlexPrefill, a Flexible sparse Pre-filling mechanism that dynamically adj
usts sparse attention patterns and computational budget in real-time to meet the
 specific requirements of each input and attention head. The flexibility of our
method is demonstrated through two key innovations: 1) Query-Aware Sparse Patter
n Determination: By measuring Jensen-Shannon divergence, this component adaptive
ly switches between query-specific diverse attention patterns and predefined att
ention patterns. 2) Cumulative-Attention Based Index Selection: This component d
ynamically selects query-key indexes to be computed based on different attention
 patterns, ensuring the sum of attention scores meets a predefined threshold.
FlexPrefill adaptively optimizes the sparse pattern and sparse ratio of each att
ention head based on the prompt, enhancing efficiency in long-sequence inference
 tasks. Experimental results show significant improvements in both speed and acc
uracy over prior methods, providing a more flexible and efficient solution for L
LM inference.
****
Yuwei Luo,Mohsen Bayati
Geometry-Aware Approaches for Balancing Performance and Theoretical Guarantees i
n Linear Bandits
This paper is motivated by recent research in the d-dimensional stochastic lin
ear bandit literature, which has revealed an unsettling discrepancy: algorithms
like Thompson sampling and Greedy demonstrate promising empirical performance, y
et this contrasts with their pessimistic theoretical regret bounds. The challeng
e arises from the fact that while these algorithms may perform poorly in certain
 problem instances, they generally excel in typical instances. To address this,
we propose a new data-driven technique that tracks the geometric properties of t
he uncertainty ellipsoid around the main problem parameter. This methodology ena
bles us to formulate a data-driven frequentist regret bound, which incorporates
the geometric information, for a broad class of base algorithms, including Greed
y, OFUL, and Thompson sampling. This result allows us to identify and ``course-c
orrect" problem instances in which the base algorithms perform poorly. The cours
e-corrected algorithms achieve the minimax optimal regret of order $\tilde{\math
cal{O}}(d\sqrt{T})$ for a T-period decision-making scenario, effectively maint
aining the desirable attributes of the base algorithms, including their empirica
l efficacy. We present simulation results to validate our findings using synthet
ic and real data.
****
Siyu Sun,Han Lu,Jiangtong Li,Yichen Xie,Tianjiao Li,Xiaokang Yang,Liqing Zhang,J
unchi Yan
Rethinking Classifier Re-Training in Long-Tailed Recognition: Label Over-Smooth
Can Balance
In the field of long-tailed recognition, the Decoupled Training paradigm has sho

wn exceptional promise by dividing training into two stages: representation lear
ning and classifier re-training. While previous work has tried to improve both
stages simultaneously, this complicates isolating the effect of classifier re-tr
aining. Recent studies reveal that simple regularization can produce strong fea
ture representations, highlighting the need to reassess classifier re-training m
ethods. In this study, we revisit classifier re-training methods based on a uni
fied feature representation and re-evaluate their performances.
We propose two new metrics, Logits Magnitude and Regularized Standard Deviation,
 to compare the differences and similarities between various methods.
Using these two newly proposed metrics, we demonstrate that when the Logits Magn
itude across classes is nearly balanced, further reducing its overall value can
effectively decrease errors and disturbances during training, leading to better
model performance.
Based on our analysis using these metrics, we observe that adjusting the logits
could improve model performance, leading us to develop a simple label over-smoot
hing approach to adjust the logits without requiring prior knowledge of class di
stribution.
This method softens the original one-hot labels by assigning a probability sligh
tly higher than $\frac{1}{K}$ to the true class and slightly lower than $\frac{1
}{K}$ to the other classes, where K is the number of classes.
Our method achieves state-of-the-art performance on various imbalanced datasets,
 including CIFAR100-LT, ImageNet-LT, and iNaturalist2018.
****
Mattia Segu,Luigi Piccinelli,Siyuan Li,Yung-Hsu Yang,Luc Van Gool,Bernt Schiele
Samba: Synchronized Set-of-Sequences Modeling for Multiple Object Tracking
Multiple object tracking in complex scenarios - such as coordinated dance perfor
mances, team sports, or dynamic animal groups - presents unique challenges. In t
hese settings, objects frequently move in coordinated patterns, occlude each oth
er, and exhibit long-term dependencies in their trajectories. However, it remain
s a key open research question on how to model long-range dependencies within tr
acklets, interdependencies among tracklets, and the associated temporal occlusio
ns. To this end, we introduce Samba, a novel linear-time set-of-sequences model
designed to jointly process multiple tracklets by synchronizing the multiple sel
ective state-spaces used to model each tracklet. Samba autoregressively predicts
 the future track query for each sequence while maintaining synchronized long-te
rm memory representations across tracklets. By integrating Samba into a tracking
-by-propagation framework, we propose SambaMOTR, the first tracker effectively a
ddressing the aforementioned issues, including long-range dependencies, tracklet
 interdependencies, and temporal occlusions. Additionally, we introduce an effec
tive technique for dealing with uncertain observations (MaskObs) and an efficien
t training recipe to scale SambaMOTR to longer sequences. By modeling long-range
 dependencies and interactions among tracked objects, SambaMOTR implicitly learn
s to track objects accurately through occlusions without any hand-crafted heuris
tics. Our approach significantly surpasses prior state-of-the-art on the DanceTr
ack, BFT, and SportsMOT datasets.
****
Marina Sheshukova,Denis Belomestny,Alain Oliviero Durmus,Eric Moulines,Alexey Na
umov,Sergey Samsonov
Nonasymptotic Analysis of Stochastic Gradient Descent with the Richardson–Romber
g Extrapolation
We address the problem of solving strongly convex and smooth minimization proble
ms using stochastic gradient descent (SGD) algorithm with a constant step size.
Previous works suggested to combine the Polyak-Ruppert averaging procedure with
 the Richardson-Romberg extrapolation to reduce the asymptotic bias of SGD at th
e expense of a mild increase of the variance. We significantly extend previous r
esults by providing an expansion of the mean-squared error of the resulting est
imator with respect to the number of iterations n. We show that the root mean-
squared error can be decomposed into the sum of two terms: a leading one of orde
r $\mathcal{O}(n^{-1/2})$ with explicit dependence on a minimax-optimal asympto
tic covariance matrix, and a second-order term of order $\mathcal{O}(n^{-3/4})$,

where the power $3/4$ is best known. We also extend this result to the higher-order moment bounds. Our analysis relies on the properties of the SGD iterates viewed as a time-homogeneous Markov chain. In particular, we establish that this chain is geometrically ergodic with respect to a suitably defined weighted Wasserstein semimetric.
****
Qing Wu,Chenhe Du,Xuanyu Tian,Jingyi Yu,Yuyao Zhang,Hongjiang Wei
Moner: Motion Correction in Undersampled Radial MRI with Unsupervised Neural Representation
Motion correction (MoCo) in radial MRI is a particularly challenging problem due to the unpredictability of subject movement. Current state-of-the-art (SOTA) MoCo algorithms often rely on extensive high-quality MR images to pre-train neural networks, which constrains the solution space and leads to outstanding image reconstruction results. However, the need for large-scale datasets significantly increases costs and limits model generalization. In this work, we propose Moner, an unsupervised MoCo method that jointly reconstructs artifact-free MR images and estimates accurate motion from undersampled, rigid motion-corrupted k-space data, without requiring any training data. Our core idea is to leverage the continuous prior of implicit neural representation (INR) to constrain this ill-posed inverse problem, facilitating optimal solutions. Specifically, we integrate a quasi-static motion model into the INR, granting its ability to correct subject's motion. To stabilize model optimization, we reformulate radial MRI reconstruction as a back-projection problem using the Fourier-slice theorem. Additionally, we propose a novel coarse-to-fine hash encoding strategy, significantly enhancing MoCo accuracy. Experiments on multiple MRI datasets show our Moner achieves performance comparable to SOTA MoCo techniques on in-domain data, while demonstrating significant improvements on out-of-domain data. The code is available at: https://github.com/iwuqing/Moner
****
Jiayi Liu,Denys Iliash,Angel X Chang,Manolis Savva,Ali Mahdavi Amiri
SINGAPO: Single Image Controlled Generation of Articulated Parts in Objects
We address the challenge of creating 3D assets for household articulated objects from a single image.
Prior work on articulated object creation either requires multi-view multi-state input, or only allows coarse control over the generation process.
These limitations hinder the scalability and practicality for articulated object modeling.
In this work, we propose a method to generate articulated objects from a single image.
Observing the object in a resting state from an arbitrary view, our method generates an articulated object that is visually consistent with the input image.
To capture the ambiguity in part shape and motion posed by a single view of the object, we design a diffusion model that learns the plausible variations of objects in terms of geometry and kinematics.
To tackle the complexity of generating structured data with attributes in multiple domains, we design a pipeline that produces articulated objects from high-level structure to geometric details in a coarse-to-fine manner, where we use a part connectivity graph and part abstraction as proxies.
Our experiments show that our method outperforms the state-of-the-art in articulated object creation by a large margin in terms of the generated object realism, resemblance to the input image, and reconstruction quality.
****
Jiale Chen,Dingling Yao,Adeel Pervez,Dan Alistarh,Francesco Locatello
Scalable Mechanistic Neural Networks
We propose Scalable Mechanistic Neural Network (S-MNN), an enhanced neural network framework designed for scientific machine learning applications involving long temporal sequences. By reformulating the original Mechanistic Neural Network (MNN) (Pervez et al., 2024), we reduce the computational time and space complexities from cubic and quadratic with respect to the sequence length, respectively, to linear. This significant improvement enables efficient modeling of long-term

dynamics without sacrificing accuracy or interpretability. Extensive experiments demonstrate that S-MNN matches the original MNN in precision while substantially reducing computational resources. Consequently, S-MNN can drop-in replace the original MNN in applications, providing a practical and efficient tool for integrating mechanistic bottlenecks into neural network models of complex dynamical systems. Source code is available at https://github.com/IST-DASLab/ScalableMNN.
****

Qian Long,Ruoyan Li,Minglu Zhao,Tao Gao,Demetri Terzopoulos

Inverse Attention Agents for Multi-Agent Systems

A major challenge for Multi-Agent Systems (MAS) is enabling agents to adapt dynamically to diverse environments in which opponents and teammates may continually change. Agents trained using conventional methods tend to excel only within the confines of their training cohorts; their performance drops significantly when confronting unfamiliar agents. To address this shortcoming, we introduce Inverse Attention Agents that adopt concepts from the Theory of Mind (ToM) implemented algorithmically using an attention mechanism trained in an end-to-end manner. Crucial to determining the final actions of these agents, the weights in their attention model explicitly represent attention to different goals. We furthermore propose an inverse attention network that deduces the ToM of agents based on observations and prior actions. The network infers the attentional states of other agents, thereby refining the attention weights to adjust the agent's final action. We conduct experiments in a continuous environment, tackling demanding tasks encompassing cooperation, competition, and a blend of both. They demonstrate that the inverse attention network successfully infers the attention of other agents, and that this information improves agent performance. Additional human experiments show that, compared to baseline agent models, our inverse attention agents exhibit superior cooperation with humans and better emulate human behaviors.
****

Jingxuan Chen,Derek Yuen,Bin Xie,Yuhao Yang,Gongwei Chen,Zhihao Wu,Li Yixing,Xurui Zhou,Weiwen Liu,Shuai Wang,Kaiwen Zhou,Rui Shao,Liqiang Nie,Yasheng Wang,Jianye HAO,Jun Wang,Kun Shao

SPA-BENCH: A COMPREHENSIVE BENCHMARK FOR SMARTPHONE AGENT EVALUATION

Smartphone agents are increasingly important for helping users control devices efficiently, with (Multimodal) Large Language Model (MLLM)-based approaches emerging as key contenders. Fairly comparing these agents is essential but challenging, requiring a varied task scope, the integration of agents with different implementations, and a generalisable evaluation pipeline to assess their strengths and weaknesses. In this paper, we present SPA-Bench, a comprehensive SmartPhone Agent Benchmark designed to evaluate (M)LLM-based agents in an interactive environment that simulates real-world conditions. SPA-Bench offers three key contributions: (1) A diverse set of tasks covering system and third-party apps in both English and Chinese, focusing on features commonly used in daily routines; (2) A plug-and-play framework enabling real-time agent interaction with Android devices, integrating over ten agents with the flexibility to add more; (3) A novel evaluation pipeline that automatically assesses agent performance across multiple dimensions, encompassing seven metrics related to task completion and resource consumption. Our extensive experiments across tasks and agents reveal challenges like interpreting mobile user interfaces, action grounding, memory retention, and execution costs. We propose future research directions to ease these difficulties, moving closer to real-world smartphone agent applications.
****

Angelo Porrello,Lorenzo Bonicelli,Pietro Buzzega,Monica Millunzi,Simone Calderara,Rita Cucchiara

A Second-Order Perspective on Model Compositionality and Incremental Learning

The fine-tuning of deep pre-trained models has revealed compositional properties, with multiple specialized modules that can be arbitrarily composed into a single, multi-task model. However, identifying the conditions that promote compositionality remains an open issue, with recent efforts concentrating mainly on linearized networks. We conduct a theoretical study that attempts to demystify compositionality in standard non-linear networks through the second-order Taylor appro

ximation of the loss function. The proposed formulation highlights the importanc
e of staying within the pre-training basin to achieve composable modules. Moreov
er, it provides the basis for two dual incremental training algorithms: the one
from the perspective of multiple models trained individually, while the other ai
ms to optimize the composed model as a whole. We probe their application in incr
emental classification tasks and highlight some valuable skills. In fact, the po
ol of incrementally learned modules not only supports the creation of an effecti
ve multi-task model but also enables unlearning and specialization in certain ta
sks. Code available at <https://github.com/aimagelab/mammoth>
****

Pedro Valdeira,Shiqiang Wang,Yuejie Chi

Vertical Federated Learning with Missing Features During Training and Inference

Vertical federated learning trains models from feature-partitioned datasets acro
ss multiple clients, who collaborate without sharing their local data. Standard
approaches assume that all feature partitions are available during both training
 and inference. Yet, in practice, this assumption rarely holds, as for many samp
les only a subset of the clients observe their partition. However, not utilizing
 incomplete samples during training harms generalization, and not supporting the
m during inference limits the utility of the model. Moreover, if any client leav
es the federation after training, its partition becomes unavailable, rendering t
he learned model unusable. Missing feature blocks are therefore a key challenge
limiting the applicability of vertical federated learning in real-world scenario
s. To address this, we propose LASER-VFL, a vertical federated learning method f
or efficient training and inference of split neural network-based models that is
 capable of handling arbitrary sets of partitions. Our approach is simple yet ef
fective, relying on the sharing of model parameters and on task-sampling to trai
n a family of predictors. We show that LASER-VFL achieves a $\mathcal{O}({1}/{\s
qrt{T}})$ convergence rate for nonconvex objectives and, under the Polyak-■ojasi
ewicz inequality, it achieves linear convergence to a neighborhood of the optimu
m. Numerical experiments show improved performance of LASER-VFL over the baselin
es. Remarkably, this is the case even in the absence of missing features. For ex
ample, for CIFAR-100, we see an improvement in accuracy of 19.3\% when each of
 four feature blocks is observed with a probability of 0.5 and of 9.5\% when a
ll features are observed. The code for this work is available at https://github.
com/Valdeira/LASER-VFL.
****

Hao Mark Chen,Fuwen Tan,Alexandros Kouris,Royson Lee,Hongxiang Fan,Stylianos Ven
ieris

Progressive Mixed-Precision Decoding for Efficient LLM Inference

In spite of the great potential of large language models (LLMs) across various t
asks, their deployment on resource-constrained devices remains challenging due t
o their excessive computational and memory demands. Quantization has emerged as
an effective solution by storing weights in reduced precision. However, utilizin
g low precisions (i.e.~2/3-bit) to substantially alleviate the memory-boundednes
s of LLM decoding, still suffers from prohibitive performance drop. In this work
,
we argue that existing approaches fail to explore the diversity in computational
 patterns, redundancy, and sensitivity to approximations of the different phases
 of LLM inference, resorting to a uniform quantization policy throughout.
Instead, we propose a novel phase-aware method that selectively allocates precis
ion during different phases of LLM inference, achieving both strong context extr
action during prefill and efficient memory bandwidth utilization during decoding
. To further address the memory-boundedness of the decoding phase, we introduce
Progressive Mixed-Precision Decoding (PMPD), a technique that enables the gradua
l lowering of precision deeper in the generated sequence, together with a spectr
um of precision-switching schedulers that dynamically drive the precision-loweri
ng decisions in either task-adaptive or prompt-adaptive manner.
Extensive evaluation across diverse language tasks shows that when targeting Nvi
dia GPUs, PMPD achieves 1.4$-$12.2$\times$ speedup in matrix-vector multiplicati
ons over fp16 models, while when targeting an LLM-optimized NPU, our approach de

livers a throughput gain of 3.8$-$8.0$\times$ over fp16 models and up to 1.54$\times$ over uniform quantization approaches while preserving the output quality.
****
Zun Wang,Jialu Li,Yicong Hong,Songze Li,Kunchang Li,Shoubin Yu,Yi Wang,Yu Qiao,Yali Wang,Mohit Bansal,Limin Wang
Bootstrapping Language-Guided Navigation Learning with Self-Refining Data Flywheel
Creating high-quality data for training robust language-instructed agents is a long-lasting challenge in embodied AI. In this paper, we introduce a Self-Refining Data Flywheel (SRDF) that generates high-quality and large-scale navigational instruction-trajectory pairs by iteratively refining the data pool through the collaboration between two models, the instruction generator and the navigator, without any human-in-the-loop annotation.
Specifically, SRDF starts with using a base generator to create an initial data pool for training a base navigator, followed by applying the trained navigator to filter the data pool. This leads to higher-fidelity data to train a better generator, which can, in turn, produce higher-quality data for training the next-round navigator. Such a flywheel establishes a data self-refining process, yielding a continuously improved and highly effective dataset for large-scale language-guided navigation learning. Our experiments demonstrate that after several flywheel rounds, the navigator elevates the performance boundary from 70\% to 78\% SPL on the classic R2R test set, surpassing human performance (76\%) for the first time.
Meanwhile, this process results in a superior generator, evidenced by a SPICE increase from 23.5 to 26.2, better than all previous VLN instruction generation methods. Finally, we demonstrate the scalability of our method through increasing environment and instruction diversity, and
the generalization ability of our pre-trained navigator across various downstream navigation tasks, surpassing state-of-the-art methods by a large margin in all cases.
****
Chao Li,Ziwei Deng,Chenxing Lin,Wenqi Chen,Yongquan Fu,Weiquan Liu,Chenglu Wen,Cheng Wang,Siqi Shen
DoF: A Diffusion Factorization Framework for Offline Multi-Agent Reinforcement Learning
Diffusion models have been widely adopted in image and language generation and are now being applied to reinforcement learning. However, the application of diffusion models in offline cooperative Multi-Agent Reinforcement Learning (MARL) remains limited. Although existing studies explore this direction, they suffer from scalability or poor cooperation issues due to the lack of design principles for diffusion-based MARL. The Individual-Global-Max (IGM) principle is a popular design principle for cooperative MARL. By satisfying this principle, MARL algorithms achieve remarkable performance with good scalability. In this work, we extend the IGM principle to the Individual-Global-identically-Distributed (IGD) principle. This principle stipulates that the generated outcome of a multi-agent diffusion model should be identically distributed as the collective outcomes from multiple individual-agent diffusion models. We propose DoF, a diffusion factorization framework for Offline MARL. It uses noise factorization function to factorize a centralized diffusion model into multiple diffusion models. We theoretically show that the noise factorization functions satisfy the IGD principle. Furthermore, DoF uses data factorization function to model the complex relationship among data generated by multiple diffusion models. Through extensive experiments, we demonstrate the effectiveness of DoF. The source code is available at https://github.com/xmu-rl-3dv/DoF.
****
Fan Wang,Juyong Jiang,Chansung Park,Sunghun Kim,Jing Tang
KaSA: Knowledge-Aware Singular-Value Adaptation of Large Language Models
The increasing sizes of large language models (LLMs) result in significant computational overhead and memory usage when adapting these models to specific tasks or domains. Various parameter-efficient fine-tuning (PEFT) methods have been dev

ised to mitigate these challenges by training a small set of parameters for the task-specific updates of the model weights. Among PEFT methods, LoRA stands out for its simplicity and efficiency, inspiring the development of a series of variants. However, LoRA and its successors disregard the knowledge that is noisy or irrelevant to the targeted task, detrimentally impacting model performance and leading to suboptimality. To address this limitation, we introduce Knowledge-aware Singular-value Adaptation (KaSA), a PEFT method that leverages singular value decomposition (SVD) with knowledge-aware singular values to dynamically activate knowledge based on its relevance to the task at hand. We conduct extensive experiments across a range of LLMs on tasks spanning natural language understanding (NLU), generation (NLG), instruction following, and commonsense reasoning. The experimental results demonstrate that KaSA consistently outperforms FFT and 14 popular PEFT baselines across 16 benchmarks and 4 synthetic datasets, underscoring our method's efficacy and adaptability. The source code of our method is available at https://github.com/juyongjiang/KaSA.
****
Ruibing Song,Chuan Liu,Chunshu Wu,Ang Li,Dongfang Liu,Ying Nian Wu,Tong Geng
DS-LLM: Leveraging Dynamical Systems to Enhance Both Training and Inference of Large Language Models
The training of large language models (LLMs) faces significant computational cost challenges, limiting their scalability toward artificial general intelligence (AGI) and broader adoption. With model sizes doubling approximately every 3.4 months and training costs escalating from 64 million USD for GPT-4 in 2020 to 191 million USD for Gemini Ultra in 2023, the economic burden has become unsustainable. While techniques such as quantization offer incremental improvements, they fail to address the fundamental computational bottleneck. In this work, we introduce DS-LLM, a novel framework that leverages dynamical system (DS)-based machines, which exploit Natural Annealing to rapidly converge to minimal energy states, yielding substantial efficiency gains. Unlike traditional methods, DS-LLM maps LLM components to optimization problems solvable via Hamiltonian configurations and utilizes continuous electric current flow in DS-machines for hardware-native gradient descent during training. We mathematically demonstrate the equivalence between conventional LLMs and DS-LLMs and present a method for transforming a trained LLM into a DS-LLM. Experimental evaluations across multiple model sizes demonstrate orders-of-magnitude improvements in speed and energy efficiency for both training and inference while maintaining consistent accuracy. Additionally, we provide an in-depth analysis of the challenges and potential solutions associated with this emerging computing paradigm, aiming to lay a solid foundation for future research.
****
Rodrigo González Laiz,Tobias Schmidt,Steffen Schneider
Self-supervised contrastive learning performs non-linear system identification
Self-supervised learning (SSL) approaches have brought tremendous success across many tasks and domains. It has been argued that these successes can be attributed to a link between SSL and identifiable representation learning: Temporal structure and auxiliary variables ensure that latent representations are related to the true underlying generative factors of the data. Here, we deepen this connection and show that SSL can perform system identification in latent space. We propose DynCL, a framework to uncover linear, switching linear and non-linear dynamics under a non-linear observation model, give theoretical guarantees and validate them empirically.
****
Jintao Zhang,Jia wei,Pengle Zhang,Jun Zhu,Jianfei Chen
SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration
The transformer architecture predominates across various models. As the heart of the transformer, attention has a computational complexity of $O(N^2)$, compared to $O(N)$ for linear transformations. When handling large sequence lengths, attention becomes the primary time-consuming component. Although quantization has proven to be an effective method for accelerating model inference, existing quantization methods primarily focus on optimizing the linear layer.

In response, we first analyze the feasibility of quantization in attention detailedly. Following that, we propose SageAttention, a highly efficient and accurate quantization method for attention. The OPS (operations per second) of our approach outperforms FlashAttention2 and xformers by about 2.1x and 2.7x, respectively. SageAttention also achieves superior accuracy performance over FlashAttention3. Comprehensive experiments confirm that our approach incurs almost no end-to-end metrics loss across diverse models—including those for large language processing, image generation, and video generation. The code is available at https://github.com/thu-ml/SageAttention.
****
Simon Dahan,Gabriel Bénédict,Logan Zane John Williams,Yourong Guo,Daniel Rueckert,Robert Leech,Emma Claire Robinson
SIM: Surface-based fMRI Analysis for Inter-Subject Multimodal Decoding from Movie-Watching Experiments
Current AI frameworks for brain decoding and encoding, typically train and test models within the same datasets. This limits their utility for cognitive training (neurofeedback) for which it would be useful to pool experiences across individuals to better simulate stimuli not sampled during training. A key obstacle to model generalisation is the degree of variability of inter-subject cortical organisation, which makes it difficult to align or compare cortical signals across participants. In this paper we address this through use of surface vision transformers, which build a generalisable model of cortical functional dynamics, through encoding the topography of cortical networks and their interactions as a moving image across a surface. This is then combined with tri-modal self-supervised contrastive (CLIP) alignment of audio, video, and fMRI modalities to enable the retrieval of visual and auditory stimuli from patterns of cortical activity (and vice-versa). We validate our approach on 7T task-fMRI data from 174 healthy participants engaged in the movie-watching experiment from the Human Connectome Project (HCP). Results show that it is possible to detect which movie clips an individual is watching purely from their brain activity, even for individuals and movies *not seen during training*. Further analysis of attention maps reveals that our model captures individual patterns of brain activity that reflect semantic and visual systems. This opens the door to future personalised simulations of brain function. Code \& pre-trained models will be made available at https://github.com/metrics-lab/sim.
****
Xingyao Wang,Boxuan Li,Yufan Song,Frank F. Xu,Xiangru Tang,Mingchen Zhuge,Jiayi Pan,Yueqi Song,Bowen Li,Jaskirat Singh,Hoang H. Tran,Fuqiang Li,Ren Ma,Mingzhang Zheng,Bill Qian,Yanjun Shao,Niklas Muennighoff,Yizhe Zhang,Binyuan Hui,Junyang Lin,Robert Brennan,Hao Peng,Heng Ji,Graham Neubig
OpenHands: An Open Platform for AI Software Developers as Generalist Agents
Software is one of the most powerful tools that we humans have at our disposal; it allows a skilled programmer to interact with the world in complex and profound ways. At the same time, thanks to improvements in large language models (LLMs), there has also been a rapid development in AI agents that interact with and effect change in their surrounding environments. In this paper, we introduce OpenHands, a platform for the development of powerful and flexible AI agents that interact with the world in similar ways to a human developer: by writing code, interacting with a command line, and browsing the web. We describe how the platform allows for the implementation of new agents, utilization of various LLMs, safe interaction with sandboxed environments for code execution, and incorporation of evaluation benchmarks. Based on our currently incorporated benchmarks, we perform an evaluation of agents over 13 challenging tasks, including software engineering (e.g., SWE-Bench) and web browsing (e.g., WebArena), amongst others. Released under the permissive MIT license, OpenHands is a community project spanning academia and industry with more than 2K contributions from over 186 contributors in less than six months of development, and will improve going forward.
****
Shih-Hsin Wang,Yuhao Huang,Justin M. Baker,Yuan-En Sun,Qi Tang,Bao Wang
A Theoretically-Principled Sparse, Connected, and Rigid Graph Representation of

Molecules
Graph neural networks (GNNs) -- learn graph representations by exploiting graph's sparsity, connectivity, and symmetries -- have become indispensable for learning geometric data like molecules. However, the most used graphs (e.g., radial cutoff graphs) in molecular modeling lack theoretical guarantees for achieving connectivity and sparsity simultaneously, which are essential for the performance and scalability of GNNs. Furthermore, existing widely used graph construction methods for molecules lack rigidity, limiting GNNs' ability to exploit graph nodes' spatial arrangement. In this paper, we introduce a new hyperparameter-free graph construction of molecules and beyond with sparsity, connectivity, and rigidity guarantees. Remarkably, our method consistently generates connected and sparse graphs with the edge-to-node ratio being bounded above by 3. Our graphs' rigidity guarantees that edge distances and dihedral angles are sufficient to uniquely determine the general spatial arrangements of atoms. We substantiate the effectiveness and efficiency of our proposed graphs in various molecular modeling benchmarks. Code is available at \url{https://github.com/Utah-Math-Data-Science/UnitSphere}.
****
Michael Sun,Alston Lo,Minghao Guo,Jie Chen,Connor W. Coley,Wojciech Matusik
Procedural Synthesis of Synthesizable Molecules
Designing synthetically accessible molecules and recommending analogs to unsynthesizable molecules are important problems for accelerating molecular discovery. We reconceptualize both problems using ideas from program synthesis. Drawing inspiration from syntax-guided synthesis approaches, we decouple the syntactic skeleton from the semantics of a synthetic tree to create a bilevel framework for reasoning about the combinatorial space of synthesis pathways. Given a molecule we aim to generate analogs for, we iteratively refine its skeletal characteristics via Markov Chain Monte Carlo simulations over the space of syntactic skeletons. Given a black-box oracle to optimize, we formulate a joint design space over syntactic templates and molecular descriptors and introduce evolutionary algorithms that optimize both syntactic and semantic dimensions synergistically. Our key insight is that once the syntactic skeleton is set, we can amortize over the search complexity of deriving the program's semantics by training policies to fully utilize the fixed horizon Markov Decision Process imposed by the syntactic template. We demonstrate performance advantages of our bilevel framework for synthesizable analog generation and synthesizable molecule design. Notably, our approach offers the user explicit control over the resources required to perform synthesis and biases the design space towards simpler solutions, making it particularly promising for autonomous synthesis platforms. Supporting code is at https://github.com/shiningsunnyday/SynthesisNet.
****
Wangjia Yu,Xiaomeng Fu,Qiao Li,Jizhong Han,Xiaodan Zhang
Resolution Attack: Exploiting Image Compression to Deceive Deep Neural Networks
Model robustness is essential for ensuring the stability and reliability of machine learning systems. Despite extensive research on various aspects of model robustness, such as adversarial robustness and label noise robustness, the exploration of robustness towards different resolutions, remains less explored. To address this gap, we introduce a novel form of attack: the resolution attack. This attack aims to deceive both classifiers and human observers by generating images that exhibit different semantics across different resolutions. To implement the resolution attack, we propose an automated framework capable of generating dual-semantic images in a zero-shot manner. Specifically, we leverage large-scale diffusion models for their comprehensive ability to construct images and propose a staged denoising strategy to achieve a smoother transition across resolutions. Through the proposed framework, we conduct resolution attacks against various off-the-shelf classifiers. The experimental results exhibit high attack success rate, which not only validates the effectiveness of our proposed framework but also reveals the vulnerability of current classifiers towards different resolutions. Additionally, our framework, which incorporates features from two distinct objects, serves as a competitive tool for applications such as face swapping and faci

al camouflage. The code is available at https://github.com/ywj1/resolution-attac
k.
****
Zuyan Liu,Yuhao Dong,Ziwei Liu,Winston Hu,Jiwen Lu,Yongming Rao
Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution
Visual data comes in various forms, ranging from small icons of just a few pixel
s to long videos spanning hours. Existing multi-modal LLMs usually standardize t
hese diverse visual inputs to fixed-resolution images or patches for visual enco
ders and yield similar numbers of tokens for LLMs. This approach is non-optimal
for multimodal understanding and inefficient for processing inputs with long and
 short visual contents. To solve the problem, we propose Oryx, a unified multimo
dal architecture for the spatial-temporal understanding of images, videos, and m
ulti-view 3D scenes. Oryx offers an on-demand solution to seamlessly and efficie
ntly process visual inputs with arbitrary spatial sizes and temporal lengths thr
ough two core innovations: 1) a pre-trained OryxViT model that can encode images
 at any resolution into LLM-friendly visual representations; 2) a dynamic compre
ssor module that supports 1x to 16x compression on visual tokens by request. The
se designs enable Oryx to accommodate extremely long visual contexts, such as vi
deos, with lower resolution and high compression while maintaining high recognit
ion precision for tasks like document understanding with native resolution and n
o compression. Beyond the architectural improvements, enhanced data curation and
 specialized training on long-context retrieval and spatial-aware data help Oryx
 achieve strong capabilities in image, video, and 3D multimodal understanding si
multaneously.
****
Shilin Yan,Ouxiang Li,Jiayin Cai,Yanbin Hao,Xiaolong Jiang,Yao Hu,Weidi Xie
A Sanity Check for AI-generated Image Detection
With the rapid development of generative models, discerning AI-generated content
 has evoked increasing attention from both industry and academia. In this paper,
 we conduct a sanity check on whether the task of AI-generated image detection h
as been solved. To start with, we present Chameleon dataset, consisting of AI-ge
nerated images that are genuinely challenging for human perception. To quantify
the generalization of existing methods, we evaluate 9 off-the-shelf AI-generated
 image detectors on Chameleon dataset. Upon analysis, almost all models misclass
ify AI-generated images as real ones. Later, we propose AIDE AI-generated Image
DEtector with Hybrid Features, which leverages multiple experts to simultaneousl
y extract visual artifacts and noise patterns. Specifically, to capture the high
-level semantics, we utilize CLIP to compute the visual embedding. This effectiv
ely enables the model to discern AI-generated images based on semantics and cont
extual information. Secondly, we select the highest and lowest frequency patches
 in the image, and compute the low-level patchwise features, aiming to detect AI
-generated images by low-level artifacts, for example, noise patterns, anti-alia
sing effects. While evaluating on existing benchmarks, for example, AIGCDetectBe
nchmark and GenImage, AIDE achieves +3.5% and +4.6% improvements to state-of-the
-art methods, and on our proposed challenging Chameleon benchmarks, it also achi
eves promising results, despite the problem of detecting AI-generated images rem
ains far from being solved.
****
Weibin Chen,Azhir Mahmood,Michel Tsamados,So Takao
Deep Random Features for Scalable Interpolation of Spatiotemporal Data
The rapid growth of earth observation systems calls for a scalable approach to i
nterpolate remote-sensing observations. These methods in principle, should acqui
re more information about the observed field as data grows. Gaussian processes (
GPs) are candidate model choices for interpolation. However, due to their poor s
calability, they usually rely on inducing points for inference, which restricts
their expressivity. Moreover, commonly imposed assumptions such as stationarity
prevents them from capturing complex patterns in the data. While deep GPs can ov
ercome this issue, training and making inference with them are difficult, again
requiring crude approximations via inducing points. In this work, we instead app
roach the problem through Bayesian deep learning, where spatiotemporal fields ar

e represented by deep neural networks, whose layers share the inductive bias of stationary GPs on the plane/sphere via random feature expansions. This allows on e to (1) capture high frequency patterns in the data, and (2) use mini-batched g radient descent for large scale training. We experiment on various remote sensin g data at local/global scales, showing that our approach produce competitive or superior results to existing methods, with well-calibrated uncertainties.
****

Seungyoon Woo,Junhyeog Yun,Gunhee Kim
Meta-Continual Learning of Neural Fields
Neural Fields (NF) have gained prominence as a versatile framework for complex d ata representation. This work unveils a new problem setting termed Meta-Continua l Learning of Neural Fields (MCL-NF) and introduces a novel strategy that employ s a modular architecture combined with optimization-based meta-learning. Focused on overcoming the limitations of existing methods for continual learning of neu ral fields, such as catastrophic forgetting and slow convergence, our strategy a chieves high-quality reconstruction with significantly improved learning speed. We further introduce Fisher Information Maximization loss for neural radiance fi elds (FIM-NeRF), which maximizes information gains at the sample level to enhanc e learning generalization, with proved convergence guarantee and generalization bound. We perform extensive evaluations across image, audio, video reconstructio n, and view synthesis tasks on six diverse datasets, demonstrating our method’s superiority in reconstruction quality and speed over existing MCL and CL-NF appr oaches. Notably, our approach attains rapid adaptation of neural fields for city -scale NeRF rendering with reduced parameter requirement.
****

Jingsheng Gao,Linxu Li,Ke Ji,Weiyuan Li,Yixin Lian,yuzhuo fu,Bin Dai
SmartRAG: Jointly Learn RAG-Related Tasks From the Environment Feedback
RAG systems consist of multiple modules to work together. However, these modules are usually separately trained. We argue that a system like RAG that incorporat es multiple modules should be jointly optimized to achieve optimal performance. To demonstrate this, we design a specific pipeline called SmartRAG that includes a policy network and a retriever. The policy network can serve as 1) a decision maker that decides when to retrieve, 2) a query rewriter to generate a query mo st suited to the retriever and 3) an answer generator that produces the final re sponse with/without the observations. We then propose to jointly optimize the wh ole system using a reinforcement learning algorithm, with the reward designed to encourage the system to achieve the highest performance with minimal retrieval cost. When jointly optimized, each module can be aware of how other modules are working and thus find the best way to work together as a complete system. Empiri cal results demonstrate that the jointly optimized system can achieve better per formance than separately optimized counterparts.
****

hongyi nie,Quanming Yao,Yang Liu,Zhen Wang,Yatao Bian
Erasing Concept Combination from Text-to-Image Diffusion Model
Advancements in the text-to-image diffusion model have raised security concerns due to their potential to generate images with inappropriate themes such as soci etal biases and copyright infringements. Current studies have made notable progr ess in preventing the model from generating images containing specific high-risk visual concepts. However, these methods neglect the issue that inappropriate th emes may also arise from the combination of benign visual concepts. A crucial ch allenge arises because the same image theme can be represented through multiple distinct visual concept combinations, and the model's ability to generate indivi dual concepts may become distorted when processing these combinations. Consequen tly, effectively erasing such visual concept combinations from the diffusion mod el remains a formidable challenge. To tackle this problem, we formalize the prob lem as the Concept Combination Erasing (CCE) problem and propose a Concept Graph -based high-level Feature Decoupling framework (CoGFD) to address CCE. CoGFD ide ntifies and decomposes visual concept combinations with a consistent image theme from an LLM-induced concept logic graph, and erases these combinations through decoupling co-occurrent high-level features. These techniques enable CoGFD to el

iminate undesirable visual concept combinations while minimizing adverse effects on the generative fidelity of related individual concepts, outperforming state-of-the-art baselines. Extensive experiments across diverse visual concept combination scenarios verify the effectiveness of CoGFD.
****

Adriano Guastella,Lorenzo Sani,Alex Iacob,Alessio Mora,Paolo Bellavista,Nicholas Donald Lane

SparsyFed: Sparse Adaptive Federated Learning

Sparse training is often adopted in cross-device federated learning (FL) environments where constrained devices collaboratively train a machine learning model on private data by exchanging pseudo-gradients across heterogeneous networks. Although sparse training methods can reduce communication overhead and computational burden in FL, they are often not used in practice for the following key reasons: (1) data heterogeneity makes it harder for clients to reach consensus on sparse models compared to dense ones, requiring longer training; (2) methods for obtaining sparse masks lack adaptivity to accommodate very heterogeneous data distributions, crucial in cross-device FL; and (3) additional hyperparameters are required, which are notably challenging to tune in FL. This paper presents SparsyFed, a practical federated sparse training method that critically addresses the problems above. Previous works have only solved one or two of these challenges at the expense of introducing new trade-offs, such as clients' consensus on masks versus sparsity pattern adaptivity. We show that SparsyFed simultaneously (1) can produce 95% sparse models, with negligible degradation in accuracy, while only needing a single hyperparameter, (2) achieves a per-round weight regrowth 200 times smaller than previous methods, and (3) allows the sparse masks to adapt to highly heterogeneous data distributions and outperform all baselines under such conditions.
****

Kwanyoung Park,Youngwoon Lee

Model-based Offline Reinforcement Learning with Lower Expectile Q-Learning

Model-based offline reinforcement learning (RL) is a compelling approach that addresses the challenge of learning from limited, static data by generating imaginary trajectories using learned models. However, these approaches often struggle with inaccurate value estimation from model rollouts. In this paper, we introduce a novel model-based offline RL method, Lower Expectile Q-learning (LEQ), which provides a low-bias model-based value estimation via lower expectile regression of λ-returns. Our empirical results show that LEQ significantly outperforms previous model-based offline RL methods on long-horizon tasks, such as the D4RL AntMaze tasks, matching or surpassing the performance of model-free approaches and sequence modeling approaches. Furthermore, LEQ matches the performance of state-of-the-art model-based and model-free methods in dense-reward environments across both state-based tasks (NeoRL and D4RL) and pixel-based tasks (V-D4RL), showing that LEQ works robustly across diverse domains. Our ablation studies demonstrate that lower expectile regression, λ-returns, and critic training on offline data are all crucial for LEQ.
****

Chongjie Si,Xuehui Wang,Xue Yang,Zhengqin Xu,Qingyun Li,Jifeng Dai,Yu Qiao,Xiaokang Yang,Wei Shen

Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning

Adapting pre-trained foundation models for various downstream tasks has been prevalent in artificial intelligence. Due to the vast number of tasks and high costs, adjusting all parameters becomes unfeasible. To mitigate this, several fine-tuning techniques have been developed to update the pre-trained model weights in a more resource-efficient manner, such as through low-rank adjustments. Yet, almost all of these methods focus on linear weights, neglecting the intricacies of parameter spaces in higher dimensions like 4D.

Alternatively, some methods can be adapted for high-dimensional parameter space by compressing changes in the original space into two dimensions and then employing low-rank matrix adaptations. However, these approaches destructs the structu

ral integrity of the involved high-dimensional spaces. To tackle the diversity o
f dimensional spaces across different foundation models and provide a more preci
se representation of the changes within these spaces, this paper introduces a ge
neralized parameter-efficient fine-tuning framework, designed for various dimens
ional parameter space. Specifically, our method asserts that changes in each dim
ensional parameter space are based on a low-rank core space which maintains the
consistent topological structure with the original space. It then models the cha
nges through this core space alongside corresponding weights to reconstruct alte
rations in the original space. It effectively preserves the structural integrity
 of the change of original N-dimensional parameter space, meanwhile models it vi
a low-rank tensor adaptation. Extensive experiments on computer vision, natural
language processing and multi-modal tasks validate the effectiveness of our meth
od.
****

Yaniv Nikankin,Anja Reusch,Aaron Mueller,Yonatan Belinkov
Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristi
cs
Do large language models (LLMs) solve reasoning tasks by learning robust general
izable algorithms, or do they memorize training data? To investigate this questi
on, we use arithmetic reasoning as a representative task. Using causal analysis,
 we identify a subset of the model (a circuit) that explains most of the model's
 behavior for basic arithmetic logic and examine its functionality. By zooming i
n on the level of individual circuit neurons, we discover a sparse set of import
ant neurons that implement simple heuristics. Each heuristic identifies a numeri
cal input pattern and outputs corresponding answers. We hypothesize that the com
bination of these heuristic neurons is the mechanism used to produce correct ari
thmetic answers. To test this, we categorize each neuron into several heuristic
types---such as neurons that activate when an operand falls within a certain ran
ge---and find that the unordered combination of these heuristic types is the mec
hanism that explains most of the model's accuracy on arithmetic prompts. Finally
, we demonstrate that this mechanism appears as the main source of arithmetic ac
curacy early in training. Overall, our experimental results across several LLMs
show that LLMs perform arithmetic using neither robust algorithms nor memorizati
on; rather, they rely on a ``bag of heuristics''.
****

Ouns El Harzli,Bernardo Cuenca Grau
Bayesian Treatment of the Spectrum of the Empirical Kernel in (Sub)Linear-Width
Neural Networks
We study Bayesian neural networks (BNNs) in the theoretical limits of infinitely
 increasing number of training examples, network width and input space dimension
. Our findings establish new bridges between kernel-theoretic approaches and tec
hniques derived from statistical mechanics through the correspondence between Me
rcer's eigenvalues and limiting spectral distributions of covariance matrices st
udied in random matrix theory.

Our theoretical contributions first consist in novel integral formulas that a
ccurately describe the predictors of BNNs in the asymptotic linear-width and sub
linear-width regimes. Moreover, we extend the recently developed renormalisation
 theory of deep linear neural networks, enabling a rigorous explanation of the m
ounting empirical evidence that hints at the theory's applicability to nonlinear
 BNNs with ReLU activations in the linear-width regime.

From a practical standpoint, our results introduce a novel technique for esti
mating the predictor statistics of a trained BNN that is applicable to the subli
near-width regime where the predictions of the renormalisation theory are inaccu
rate.
****

Xi Lin,Yilu Liu,Xiaoyuan Zhang,Fei Liu,Zhenkun Wang,Qingfu Zhang
Few for Many: Tchebycheff Set Scalarization for Many-Objective Optimization
Multi-objective optimization can be found in many real-world applications where
some conflicting objectives can not be optimized by a single solution. Existing
optimization methods often focus on finding a set of Pareto solutions with diffe

rent optimal trade-offs among the objectives. However, the required number of so
lutions to well approximate the whole Pareto optimal set could be exponentially
large with respect to the number of objectives, which makes these methods unsuit
able for handling many optimization objectives. In this work, instead of finding
 a dense set of Pareto solutions, we propose a novel Tchebycheff set scalarizati
on method to find a few representative solutions (e.g., 5) to cover a large numb
er of objectives (e.g., >100) in a collaborative and complementary manner. In
this way, each objective can be well addressed by at least one solution in the s
mall solution set. In addition, we further develop a smooth Tchebycheff set scal
arization approach for efficient optimization with good theoretical guarantees.
Experimental studies on different problems with many optimization objectives dem
onstrate the effectiveness of our proposed method.
****

Weibin Liao,Xu Chu,Yasha Wang

TPO: Aligning Large Language Models with Multi-branch & Multi-step Preference Tr
ees

In the domain of complex reasoning tasks, such as mathematical reasoning, recent
 advancements have proposed the use of Direct Preference Optimization (DPO) to s
uppress output of dispreferred responses, thereby enhancing the long-chain reaso
ning capabilities of large language models (LLMs). To this end, these studies em
ployed LLMs to generate preference trees via Tree-of-thoughts (ToT) and sample t
he paired preference responses required by the DPO algorithm. However, the DPO a
lgorithm based on binary preference optimization is unable to learn multiple res
ponses with varying degrees of preference/dispreference that provided by the pre
ference trees, resulting in incomplete preference learning. In this work, we int
roduce Tree Preference Optimization (TPO), that does not sample paired preferenc
e responses from the preference tree; instead, it directly learns from the entir
e preference tree during the fine-tuning. Specifically, TPO formulates the langu
age model alignment as a Preference List Ranking problem, where the policy can p
otentially learn more effectively from a ranked preference list of responses giv
en the prompt. In addition, to further assist LLMs in identifying discriminativ
e steps within long-chain reasoning and increase the relative reward margin in t
he preference list, TPO utilizes Adaptive Step Reward to adjust the reward value
s of each step in trajectory for performing fine-grained preference optimization
. We carry out extensive experiments on mathematical reasoning tasks to evaluate
 TPO. The experimental results indicate that TPO consistently outperforms DPO ac
ross five public large language models on four datasets.
****

Kaustubh Sridhar,Souradeep Dutta,Dinesh Jayaraman,Insup Lee

REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New En
vironments

Building generalist agents that can rapidly adapt to new environments is a key c
hallenge for deploying AI in the digital and real worlds. Is scaling current age
nt architectures the most effective way to build generalist agents? We propose a
 novel approach to pre-train relatively small policies on relatively small datas
ets and adapt them to unseen environments via in-context learning, without any f
inetuning. Our key idea is that retrieval offers a powerful bias for fast adapta
tion. Indeed, we demonstrate that even a simple retrieval-based 1-nearest neighb
or agent offers a surprisingly strong baseline for today's state-of-the-art gene
ralist agents. From this starting point, we construct a semi-parametric agent, R
EGENT, that trains a transformer-based policy on sequences of queries and retrie
ved neighbors. REGENT can generalize to unseen robotics and game-playing environ
ments via retrieval augmentation and in-context learning, achieving this with up
 to 3x fewer parameters and up to an order-of-magnitude fewer pre-training datap
oints, significantly outperforming today's state-of-the-art generalist agents.
****

Georg Manten,Cecilia Casolo,Emilio Ferrucci,Søren Wengel Mogensen,Cristopher Sal
vi,Niki Kilbertus

Signature Kernel Conditional Independence Tests in Causal Discovery for Stochast
ic Processes

Inferring the causal structure underlying stochastic dynamical systems from obse
rvational data holds great promise in domains ranging from science and health to
 finance. Such processes can often be accurately modeled via stochastic differen
tial equations (SDEs), which naturally imply causal relationships via `which var
iables enter the differential of which other variables'. In this paper, we devel
op conditional independence (CI) constraints on coordinate processes over select
ed intervals that are Markov with respect to the acyclic dependence graph (allow
ing self-loops) induced by a general SDE model. We then provide a sound and comp
lete causal discovery algorithm, capable of handling both fully and partially ob
served data, and uniquely recovering the underlying or induced ancestral graph b
y exploiting time directionality assuming a CI oracle. Finally, to make our algo
rithm practically usable, we also propose a flexible, consistent signature kerne
l-based CI test to infer these constraints from data. We extensively benchmark t
he CI test in isolation and as part of our causal discovery algorithms, outperfo
rming existing approaches in SDE models and beyond.
****

Ryotaro Kawata,Kazusato Oko,Atsushi Nitanda,Taiji Suzuki
Direct Distributional Optimization for Provable Alignment of Diffusion Models
We introduce a novel alignment method for diffusion models from distribution opt
imization perspectives while providing rigorous convergence guarantees.
We first formulate the problem as a generic regularized loss minimization over p
robability distributions and directly optimize the distribution using the Dual A
veraging method.
Next, we enable sampling from the learned distribution by approximating its scor
e function via Doob's h-transform technique.
The proposed framework is supported by rigorous convergence guarantees and an en
d-to-end bound on the sampling error, which imply that when the original distrib
ution's score is known accurately, the complexity of sampling from shifted distr
ibutions is independent of isoperimetric conditions.
This framework is broadly applicable to general distribution optimization proble
ms, including alignment tasks in Reinforcement Learning with Human Feedback (RLH
F), Direct Preference Optimization (DPO), and Kahneman-Tversky Optimization (KTO
). We empirically validate its performance on synthetic and image datasets using
 the DPO objective.
****

Matt Ricci,Guy Pelc,Zoe Piran,Noa Moriel,Mor Nitzan
TRENDy: Temporal Regression of Effective Nonlinear Dynamics
Spatiotemporal dynamics pervade the natural sciences, from the morphogen dynamic
s underlying patterning in animal pigmentation to the protein waves controlling
cell division. A central challenge lies in understanding how controllable parame
ters induce qualitative changes in system behavior called bifurcations. This end
eavor is particularly difficult in realistic settings where governing partial di
fferential equations (PDEs) are unknown and data is limited and noisy. To addres
s this challenge, we propose TRENDy (Temporal Regression of Effective Nonlinear
Dynamics), an equation-free approach to learning low-dimensional, predictive mod
els of spatiotemporal dynamics. TRENDy first maps input data to a low-dimensiona
l space of effective dynamics through a cascade of multiscale filtering operatio
ns. Our key insight is the recognition that these effective dynamics can be fit
by a neural ordinary differential equation (NODE) having the same parameter spac
e as the input PDE. The preceding filtering operations strongly regularize the p
hase space of the NODE, making TRENDy significantly more robust to noise compare
d to existing methods. We train TRENDy to predict the effective dynamics of synt
hetic and real data representing dynamics from across the physical and life scie
nces. We then demonstrate how we can automatically locate both Turing and Hopf b
ifurcations in unseen regions of parameter space. We finally apply our method to
 the analysis of spatial patterning of the ocellated lizard through development.
 We found that TRENDy's predicted effective state not only accurately predicts s
patial changes over time but also identifies distinct pattern features unique to
 different anatomical regions, such as the tail, neck, and body--an insight that
 highlights the potential influence of surface geometry on reaction-diffusion me

chanisms and their role in driving spatially varying pattern dynamics.
****
Junyi An,Chao Qu,Yun-Fei Shi,XinHao Liu,Qianwei Tang,Fenglei Cao,Yuan Qi
Equivariant Masked Position Prediction for Efficient Molecular Representation
Graph neural networks (GNNs) have shown considerable promise in computational chemistry. However, the limited availability of molecular data raises concerns regarding GNNs' ability to effectively capture the fundamental principles of physics and chemistry, which constrains their generalization capabilities. To address this challenge, we introduce a novel self-supervised approach termed Equivariant Masked Position Prediction (EMPP), grounded in intramolecular potential and force theory. Unlike conventional attribute masking techniques, EMPP formulates a nuanced position prediction task that is more well-defined and enhances the learning of quantum mechanical features. EMPP also bypasses the approximation of the Gaussian mixture distribution commonly used in denoising methods, allowing for more accurate acquisition of physical properties. Experimental results indicate that EMPP significantly enhances performance of advanced molecular architectures, surpassing state-of-the-art self-supervised approaches. Our code is released in https://github.com/ajy112/EMPP.
****
Yuhan Chen,Yihong Luo,Yifan Song,Pengwen Dai,Jing Tang,Xiaochun Cao
Decoupled Graph Energy-based Model for Node Out-of-Distribution Detection on Heterophilic Graphs
Despite extensive research efforts focused on Out-of-Distribution (OOD) detection on images, OOD detection on nodes in graph learning remains underexplored. The dependence among graph nodes hinders the trivial adaptation of existing approaches on images that assume inputs to be i.i.d. sampled, since many unique features and challenges specific to graphs are not considered, such as the heterophily issue. Recently, GNNSafe, which considers node dependence, adapted energy-based detection to the graph domain with state-of-the-art performance, however, it has two serious issues: 1) it derives node energy from classification logits without specifically tailored training for modeling data distribution, making it less effective at recognizing OOD data; 2) it highly relies on energy propagation, which is based on homophily assumption and will cause significant performance degradation on heterophilic graphs, where the node tends to have dissimilar distribution with its neighbors. To address the above issues, we suggest training Energy-based Models (EBMs) by Maximum Likelihood Estimation (MLE) to enhance data distribution modeling and removing energy propagation to overcome the heterophily issues. However, training EBMs via MLE requires performing Markov Chain Monte Carlo (MCMC) sampling on both node feature and node neighbors, which is challenging due to the node interdependence and discrete graph topology. To tackle the sampling challenge, we introduce Decoupled Graph Energy-based Model (DeGEM), which decomposes the learning process into two parts—a graph encoder that leverages topology information for node representations and an energy head that operates in latent space. Additionally, we propose a Multi-Hop Graph encoder (MH) and Energy Readout (ERo) to enhance node representation learning, Conditional Energy (CE) for improved EBM training, and Recurrent Update for the graph encoder and energy head to promote each other. This approach avoids sampling adjacency matrices and removes the need for energy propagation to extract graph topology information. Extensive experiments validate that DeGEM, without OOD exposure during training, surpasses previous state-of-the-art methods, achieving an average AUROC improvement of 6.71% on *homophilic* graphs and 20.29% on *heterophilic* graphs, and even outperform methods trained with OOD exposure. Our code is available at: https://github.com/draym28/DeGEM.
****
Junyi Chen,Di Huang,Weicai Ye,Wanli Ouyang,Tong He
Where Am I and What Will I See: An Auto-Regressive Model for Spatial Localization and View Prediction
Spatial intelligence is the ability of a machine to perceive, reason, and act in three dimensions within space and time.
Recent advancements in large-scale auto-regressive models have demonstrated rema

rkable capabilities across various reasoning tasks. However, these models often struggle with fundamental aspects of spatial reasoning, particularly in answerin g questions like "Where am I?" and "What will I see?". While some attempts have been done, existing approaches typically treat them as separate tasks, failing t o capture their interconnected nature. In this paper, we present **G**enerative **S**patial **T**ransformer (GST), a novel auto-regressive framework that jointl y addresses spatial localization and view prediction. Our model simultaneously e stimates the camera pose from a single image and predicts the view from a new ca mera pose, effectively bridging the gap between spatial awareness and visual pre diction. The proposed innovative camera tokenization method enables the model to learn the joint distribution of 2D projections and their corresponding spatial perspectives in an auto-regressive manner. This unified training paradigm demons trates that joint optimization of pose estimation and novel view synthesis leads to improved performance in both tasks, for the first time, highlighting the inh erent relationship between spatial awareness and visual prediction.
****

Prateek Garg,Lokesh Nagalapatti,Sunita Sarawagi
From Search to Sampling: Generative Models for Robust Algorithmic Recourse
Algorithmic Recourse provides recommendations to individuals who are adversely i mpacted by automated model decisions, on how to alter their profiles to achieve a favorable outcome. Effective recourse methods must balance three conflicting g oals: proximity to the original profile to minimize cost, plausibility for reali stic recourse, and validity to ensure the desired outcome. We show that existing methods train for these objectives separately and then search for recourse thro ugh a joint optimization over the recourse goals during inference, leading to po or recourse recommendations. We introduce GenRe, a generative recourse model des igned to train the three recourse objectives jointly. Training such generative m odels is non-trivial due to lack of direct recourse supervision. We propose effi cient ways to synthesize such supervision and further show that GenRe's training leads to a consistent estimator. Unlike most prior methods, that employ non-rob ust gradient descent based search during inference, GenRe simply performs a forw ard sampling over the generative model to produce minimum cost recourse, leading to superior performance across multiple metrics. We also demonstrate GenRe prov ides the best trade-off between cost, plausibility and validity, compared to sta te-of-art baselines. We release anonymized code at: https://anonymous.4open.scie nce/r/GenRe-BD71
****

Liangliang Shi,Yufeng Li,Kaipeng Zeng,Yihui Tu,Junchi Yan
Optimal Flow Transport and its Entropic Regularization: a GPU-friendly Matrix It erative Algorithm for Flow Balance Satisfaction
The Sinkhorn algorithm, based on Entropic Regularized Optimal Transport (OT), ha s garnered significant attention due to its computational efficiency enabled by GPU-friendly matrix-vector multiplications. However, vanilla OT primarily deals with computations between the source and target nodes in a bipartite graph, limi ting its practical application in real-world transportation scenarios.
In this paper, we introduce the concept of Optimal Flow Transport (OFT) as an ex tension, where we consider a more general graph setting and the marginal constra ints in vanilla OT are replaced by flow balance constraints.
To obtain solutions, we incorporate entropic regularization into the OFT and int roduce virtual flows for individual nodes to tackle the issue of potentially num erous isolated nodes lacking flow passages. Our proposition, the OFT-Sinkhorn al gorithm, utilizes GPU-friendly matrix iterations to maintain flow balance constr aints and minimize the objective function, and theoretical results for global co nvergence is also proposed in this paper.
Furthermore, we enhance OFT by introducing capacity constraints on nodes and edg es, transforming the OFT problem into a minimum-cost flow problem. We then prese nt the Capacity-Constrained EOFT-Sinkhorn algorithm and compare it with the trad itional Minimum cost flow (MCF) algorithm, showing that our algorithm is quite e fficient for calculation.
In particular, our EOFT-Sinkhorn is evaluated on high-precision and integer-prec

ision MCF problems with different scales from one hundred to five thousand size, exhibiting significant time efficiency and the ability to approximate optimal s olutions.
****

Priyanshu Kumar,Elaine Lau,Saranya Vijayakumar,Tu Trinh,Elaine T Chang,Vaughn Ro binson,Shuyan Zhou,Matt Fredrikson,Sean M. Hendryx,Summer Yue,Zifan Wang
Aligned LLMs Are Not Aligned Browser Agents
For safety reasons, large language models (LLMs) are trained to refuse harmful u ser instructions, such as assisting dangerous activities. We study an open quest ion in this work: does the desired safety refusal, typically enforced in chat co ntexts, generalize to non-chat and agentic use cases? Unlike chatbots, LLM agent s equipped with general-purpose tools, such as web browsers and mobile devices, can directly influence the real world, making it even more crucial to refuse har mful instructions. In this work, we primarily focus on red-teaming browser agents – LLMs that leverage information via web browsers. To this end, we introd uce Browser Agent Red teaming Toolkit (BrowserART), a comprehensive test suite d esigned specifically for red-teaming browser agents. BrowserART consists of 100 diverse browser-related harmful behaviors (including original behaviors and ones sourced from HarmBench (Mazeika et al., 2024) and AirBench 2024 (Zeng et al., 2 024b)) across both synthetic and real websites. Our empirical study on state-of-the-art browser agents reveals that while the backbone LLM refuses harmful instr uctions as a chatbot, the corresponding agent does not. Moreover, attack methods designed to jailbreak refusal-trained LLMs in the chat settings transfer effect ively to browser agents. With human rewrites, GPT-4o and o1-preview -based brows er agents pursued 98 and 63 harmful behaviors (out of 100), respectively. Theref ore, simply ensuring LLM’s refusal to harmful instruc-
tions in chats is not sufficient to ensure that the downstream agents are safe. We publicly release BrowserART and call on LLM developers, policymakers, and age nt developers to collaborate on improving agent safety.
****

Zhengxi Lu,Shizhuo Cheng,Tintin Jiang,Yan Zhang,Min Zhang
ProtPainter: Draw or Drag Protein via Topology-guided Diffusion
Recent advances in protein backbone generation have achieved promising results u nder structural, functional, or physical constraints. However, existing methods lack the flexibility for precise topology control, limiting navigation of the ba ckbone space. We present $\textbf{ProtPainter}$, a diffusion-based approach for generating protein backbones conditioned on 3D curves. ProtPainter follows a two -stage process: curve-based sketching and sketch-guided backbone generation. For the first stage, we propose $\textbf{CurveEncoder}$, which predicts secondary s tructure annotations from a curve to parametrize sketch generation. For the seco nd stage, the sketch guides the generative process in Denoising Diffusion Probab ilistic Modeling (DDPM) to generate backbones. During the process, we further in troduce a fusion scheduling scheme, Helix-Gating, to control the scaling factors . To evaluate, we propose the first benchmark for topology-conditioned protein g eneration, introducing Protein Restoration Task and a new metric, self-consisten cy Topology Fitness (scTF). Experiments demonstrate ProtPainter's ability to gen erate topology-fit (scTF $>$ 0.8) and designable (scTM $>$ 0.5) backbones, with drawing and dragging tasks showcasing its flexibility and versatility.
****

Tao Huang,Yanxiang Ma,Shan You,Chang Xu
Learning Mask Invariant Mutual Information for Masked Image Modeling
Masked autoencoders (MAEs) represent a prominent self-supervised learning paradi gm in computer vision. Despite their empirical success, the underlying mechanism s of MAEs remain insufficiently understood. Recent studies have attempted to elu cidate the functioning of MAEs through contrastive learning and feature represen tation analysis, yet these approaches often provide only implicit insights. In t his paper, we propose a new perspective for understanding MAEs by leveraging the information bottleneck principle in information theory. Our theoretical analyse s reveal that optimizing the latent features to balance relevant and irrelevant information is key to improving MAE performance. Building upon our proofs, we in

troduce MI-MAE, a novel method that optimizes MAEs through mutual information ma
ximization and minimization. By enhancing latent features to retain maximal rele
vant information between them and the output, and minimizing irrelevant informat
ion between them and the input, our approach achieves better performance. Extens
ive experiments on standard benchmarks show that MI-MAE significantly outperform
s MAE models in tasks such as image classification, object detection, and semant
ic segmentation. Our findings validate the theoretical framework and highlight t
he practical advantages of applying the information bottleneck principle to MAEs
, offering deeper insights for developing more powerful self-supervised learning
 models.
****
Joshua Southern,Francesco Di Giovanni,Michael M. Bronstein,Johannes F. Lutzeyer
Understanding Virtual Nodes: Oversquashing and Node Heterogeneity
While message passing neural networks (MPNNs) have convincing success in a range
 of applications, they exhibit limitations such as the oversquashing problem and
 their inability to capture long-range interactions. Augmenting MPNNs with a vir
tual node (VN) removes the locality constraint of the layer aggregation and has
been found to improve performance on a range of benchmarks. We provide a compreh
ensive theoretical analysis of the role of VNs and benefits thereof, through the
 lenses of oversquashing and sensitivity analysis. First, we characterize, preci
sely, how the improvement afforded by VNs on the mixing abilities of the network
 and hence in mitigating oversquashing, depends on the underlying topology. We t
hen highlight that, unlike Graph-Transformers (GTs), classical instantiations of
 the VN are often constrained to assign uniform importance to different nodes. C
onsequently, we propose a variant of VN with the same computational complexity,
which can have different sensitivity to nodes based on the graph structure. We s
how that this is an extremely effective and computationally efficient baseline f
or graph-level tasks.
****
Rong Tang,Lizhen Lin,Yun Yang
Conditional Diffusion Models are Minimax-Optimal and Manifold-Adaptive for Condi
tional Distribution Estimation
We consider a class of conditional forward-backward diffusion models for conditi
onal generative modeling, that is, generating new data given a covariate (or con
trol variable). To formally study the theoretical properties of these conditiona
l generative models, we adopt a statistical framework of distribution regression
 to characterize the large sample properties of the conditional distribution est
imators induced by these conditional forward-backward diffusion models. Here, th
e conditional distribution of data is assumed to smoothly change over the covari
ate. In particular, our derived convergence rate is minimax-optimal under the to
tal variation metric within the regimes covered by the existing literature. Addi
tionally, we extend our theory by allowing both the data and the covariate varia
ble to potentially admit a low-dimensional manifold structure. In this scenario,
 we demonstrate that the conditional forward-backward diffusion model can adapt
to both manifold structures, meaning that the derived estimation error bound (un
der the Wasserstein metric) depends only on the intrinsic dimensionalities of th
e data and the covariate.
****
Ruibo Chen,Yihan Wu,Yanshuo Chen,Chenxi Liu,Junfeng Guo,Heng Huang
A Watermark for Order-Agnostic Language Models
Statistical watermarking techniques are well-established for sequentially decode
d language models (LMs). However, these techniques cannot be directly applied to
 order-agnostic LMs, as the tokens in order-agnostic LMs are not generated seque
ntially. In this work, we introduce PATTERN-MARK, a pattern-based watermarking f
ramework specifically designed for order-agnostic LMs. We develop a
Markov-chain-based watermark generator that produces watermark key sequences wit
h high-frequency key patterns. Correspondingly, we propose a statistical pattern
-based detection algorithm that recovers the key sequence during detection and c
onducts statistical tests based on the count of high-frequency patterns. Our ext
ensive evaluations on order-agnostic LMs, such as ProteinMPNN and CMLM, demonstr

ate PATTERN-MARK’s enhanced detection efficiency, generation quality, and robust
ness, positioning it as a superior watermarking technique for order-agnostic LMs
.
****
Raiyan R. Khan,Philippe Chlenski,Itsik Pe'er
Hyperbolic Genome Embeddings
Current approaches to genomic sequence modeling often struggle to align the indu
ctive biases of machine learning models with the evolutionarily-informed structu
re of biological systems. To this end, we formulate a novel application of hyper
bolic CNNs that exploits this structure, enabling more expressive DNA sequence r
epresentations. Our strategy circumvents the need for explicit phylogenetic mapp
ing while discerning key properties of sequences pertaining to core functional a
nd regulatory behavior. Across 37 out of 43 genome interpretation benchmark data
sets, our hyperbolic models outperform their Euclidean equivalents. Notably, our
 approach even surpasses state-of-the-art performance on seven GUE benchmark dat
asets, consistently outperforming many DNA language models while using orders of
 magnitude fewer parameters and avoiding pretraining. Our results include a nove
l set of benchmark datasets - the Transposable Elements Benchmark - which explor
es a major but understudied component of the genome with deep evolutionary signi
ficance. We further motivate our work by exploring how our hyperbolic models rec
ognize genomic signal under various data-generating conditions and by constructi
ng an empirical method for interpreting the hyperbolicity of dataset embeddings.
 Throughout these assessments, we find persistent evidence highlighting the pote
ntial of our hyperbolic framework as a robust paradigm for genome representation
 learning. Our code and benchmark datasets are available at https://github.com/r
rkhan/HGE.
****
Liu Ziyin,Isaac L. Chuang,Tomer Galanti,Tomaso A Poggio
Formation of Representations in Neural Networks
Understanding neural representations will help open the black box of neural netw
orks and advance our scientific understanding of modern AI systems. However, how
 complex, structured, and transferable representations emerge in modern neural n
etworks has remained a mystery. Building on previous results, we propose the Can
onical Representation Hypothesis (CRH), which posits a set of six alignment rela
tions to universally govern the formation of representations in most hidden laye
rs of a neural network. Under the CRH, the latent representations (R), weights (
W), and neuron gradients (G) become mutually aligned during training. This align
ment implies that neural networks naturally learn compact representations, where
 neurons and weights are invariant to task-irrelevant transformations. We then s
how that the breaking of CRH leads to the emergence of reciprocal power-law rela
tions between R, W, and G, which we refer to as the Polynomial Alignment Hypothe
sis (PAH). We present a minimal-assumption theory proving that the balance betwe
en gradient noise and regularization is crucial for the emergence of the canonic
al representation. The CRH and PAH lead to an exciting possibility of unifying m
ajor key deep learning phenomena, including neural collapse and the neural featu
re ansatz, in a single framework.
****
Dhruv Gautam,Spandan Garg,Jinu Jang,Neel Sundaresan,Roshanak Zilouchian Moghadda
m
RefactorBench: Evaluating Stateful Reasoning in Language Agents Through Code
Recent advances in language model (LM) agents and function calling have enabled
autonomous, feedback-driven systems to solve problems across various digital dom
ains. To better understand the unique limitations of LM agents, we introduce Ref
actorBench, a benchmark consisting of 100 large handcrafted multi-file refactori
ng tasks in popular open-source repositories. Solving tasks within RefactorBench
 requires thorough exploration of dependencies across multiple files and strong
adherence to relevant instructions. Every task is defined by 3 natural language
instructions of varying specificity and is mutually exclusive, allowing for the
creation of longer combined tasks on the same repository. Baselines on RefactorB
ench reveal that current LM agents struggle with simple compositional tasks, sol

ving only 22\% of tasks with base instructions, in contrast to a human developer with short time constraints solving 87\%. Through trajectory analysis, we identify various unique failure modes of LM agents, and further explore the failure mode of tracking past actions. By adapting a baseline agent to condition on representations of state, we achieve a 43.9\% improvement in solving RefactorBench tasks. We further extend our state-aware approach to encompass entire digital environments and outline potential directions for future research. RefactorBench aims to support the study of LM agents by providing a set of real-world, multi-hop tasks within the realm of code.

****

Vincent Plassier,Alexander Fishkov,Mohsen Guizani,Maxim Panov,Eric Moulines

Probabilistic Conformal Prediction with Approximate Conditional Validity

We develop a new method for generating prediction sets that combines the flexibility of conformal methods with an estimate of the conditional distribution $\textup{P}_{Y \mid X}$. Existing methods, such as conformalized quantile regression and probabilistic conformal prediction, usually provide only a marginal coverage guarantee. In contrast, our approach extends these frameworks to achieve approximately conditional coverage, which is crucial for many practical applications. Our prediction sets adapt to the behavior of the predictive distribution, making them effective even under high heteroscedasticity. While exact conditional guarantees are infeasible without assumptions on the underlying data distribution, we derive non-asymptotic bounds that depend on the total variation distance of the conditional distribution and its estimate. Using extensive simulations, we show that our method consistently outperforms existing approaches in terms of conditional coverage, leading to more reliable statistical inference in a variety of applications.

****

Somesh Kumar Singh,Yaman Kumar Singla,Harini S I,Balaji Krishnamurthy

Measuring And Improving Persuasiveness Of Large Language Models

Large Language Models (LLMs) are increasingly being used in workflows involving generating content to be consumed by humans (*e.g.,* marketing) and also in directly interacting with humans (*e.g.,* through chatbots). The development of such systems that are capable of generating verifiably persuasive messages presents both opportunities and challenges for society. On the one hand, such systems could positively impact domains like advertising and social good, such as addressing drug addiction, and on the other, they could be misused for spreading misinformation and shaping political opinions. To channel LLMs' impact on society, we need to develop systems to measure and benchmark their persuasiveness. With this motivation, we introduce **PersuasionBench** and **PersuasionArena**, the first large-scale benchmark and arena containing a battery of tasks to automatically measure the simulative and generative persuasion abilities of large language models. We introduce **transsuasion** (trans = carrying across, suasion = the act of persuading), a novel task of transforming non-persuasive language into persuasive content while preserving other factors determining persuasiveness (sender, receiver, time, and channel). Our findings indicate that the simulative persuasion capabilities of LLMs are barely above random; however, their generative persuasion capabilities are much better. For instance, GPT-4o loses only 36% of the time when playing against the best human persuader. Further, we find that LLMs' persuasiveness correlates positively with model size, but smaller models can also be made to have a higher persuasiveness than much larger models. Notably, targeted training using synthetic and natural datasets significantly enhances smaller models' persuasive capabilities, challenging scale-dependent assumptions. Our findings carry key implications for both model developers and policymakers. For instance, while the EU AI Act and California's SB-1047 aim to regulate AI models based on the number of floating point operations, we demonstrate that simple metrics like this alone fail to capture the full scope of AI's societal impact. We invite the community to explore and contribute to PersuasionArena and PersuasionBench, available at behavior-in-the-wild.github.io/measure-persuasion, to advance our understanding of AI-driven persuasion and its societal implications.

****
Ziyue Li,Tian Li,Virginia Smith,Jeff Bilmes,Tianyi Zhou
Many-Objective Multi-Solution Transport
Optimizing the performance of many objectives (instantiated by tasks or clients) jointly with a few Pareto stationary solutions (models) is critical in machine learning. However, previous multi-objective optimization methods often focus on a few objectives and cannot scale to many objectives that outnumber the solutions, leading to either subpar performance or ignored objectives. We introduce ''Many-objective multi-solution Transport (MosT)'', a framework that finds multiple diverse solutions in the Pareto front of many objectives. Our insight is to seek multiple solutions, each performing as a domain expert and focusing on a specific subset of objectives while collectively covering all of them. MosT formulates the problem as a bi-level optimization of weighted objectives for each solution, where the weights are defined by an optimal transport between objectives and solutions. Our algorithm ensures convergence to Pareto stationary solutions for complementary subsets of objectives. On a range of applications in federated learning, multi-task learning, and mixture-of-prompt learning for LLMs, MosT distinctly outperforms strong baselines, delivering high-quality, diverse solutions that profile the entire Pareto frontier, thus ensuring balanced trade-offs across many objectives.
****
Hengzhe Zhang,Qi Chen,Bing XUE,Wolfgang Banzhaf,Mengjie Zhang
RAG-SR: Retrieval-Augmented Generation for Neural Symbolic Regression
Symbolic regression is a key task in machine learning, aiming to discover mathematical expressions that best describe a dataset. While deep learning has increased interest in using neural networks for symbolic regression, many existing approaches rely on pre-trained models. These models require significant computational resources and struggle with regression tasks involving unseen functions and variables. A pre-training-free paradigm is needed to better integrate with search-based symbolic regression algorithms. To address these limitations, we propose a novel framework for symbolic regression that integrates evolutionary feature construction with a neural network, without the need for pre-training. Our approach adaptively generates symbolic trees that align with the desired semantics in real-time using a language model trained via online supervised learning, providing effective building blocks for feature construction. To mitigate hallucinations from the language model, we design a retrieval-augmented generation mechanism that explicitly leverages searched symbolic expressions. Additionally, we introduce a scale-invariant data augmentation technique that further improves the robustness and generalization of the model. Experimental results demonstrate that our framework achieves state-of-the-art accuracy across 25 regression algorithms and 120 regression tasks.
****
Anthony Zhou,Zijie Li,Michael Schneier,John R Buchanan Jr,Amir Barati Farimani
Text2PDE: Latent Diffusion Models for Accessible Physics Simulation
Recent advances in deep learning have inspired numerous works on data-driven solutions to partial differential equation (PDE) problems. These neural PDE solvers can often be much faster than their numerical counterparts; however, each presents its unique limitations and generally balances training cost, numerical accuracy, and ease of applicability to different problem setups. To address these limitations, we introduce several methods to apply latent diffusion models to physics simulation. Firstly, we introduce a mesh autoencoder to compress arbitrarily discretized PDE data, allowing for efficient diffusion training across various physics. Furthermore, we investigate full spatiotemporal solution generation to mitigate autoregressive error accumulation. Lastly, we investigate conditioning on initial physical quantities, as well as conditioning solely on a text prompt to introduce text2PDE generation. We show that language can be a compact, interpretable, and accurate modality for generating physics simulations, paving the way for more usable and accessible PDE solvers. Through experiments on both uniform and structured grids, we show that the proposed approach is competitive with current neural PDE solvers in both accuracy and efficiency, with promising scaling

behavior up to $\sim$3 billion parameters. By introducing a scalable, accurate, and usable physics simulator, we hope to bring neural PDE solvers closer to practical use.
****
Yonatan Sverdlov,Nadav Dym
On the Expressive Power of Sparse Geometric MPNNs
Motivated by applications in chemistry and other sciences, we study the expressive power of message-passing neural networks for geometric graphs, whose node features correspond to 3-dimensional positions. Recent work has shown that such models can separate generic pairs of non-isomorphic geometric graphs, though they may fail to separate some rare and complicated instances. However, these results assume a fully connected graph, where each node possesses complete knowledge of all other nodes. In contrast, often, in application, every node only possesses knowledge of a small number of nearest neighbors.
This paper shows that generic pairs of non-isomorphic geometric graphs can be separated by message-passing networks with rotation equivariant features as long as the underlying graph is connected. When only invariant intermediate features are allowed, generic separation is guaranteed for generically globally rigid graphs. We introduce a simple architecture, EGENNET, which achieves our theoretical guarantees and compares favorably with alternative architecture on synthetic and chemical benchmarks
****
Xiangru Zhu,Penglei Sun,Yaoxian Song,Yanghua Xiao,Zhixu Li,Chengyu Wang,Jun Huang,Bei Yang,Xiaoxiao Xu
Evaluating Semantic Variation in Text-to-Image Synthesis: A Causal Perspective
Accurate interpretation and visualization of human instructions are crucial for text-to-image (T2I) synthesis.
However, current models struggle to capture semantic variations from word order changes, and existing evaluations, relying on indirect metrics like text-image similarity, fail to reliably assess these challenges. This often obscures poor performance on complex or uncommon linguistic patterns by the focus on frequent word combinations.
To address these deficiencies, we propose a novel metric called SemVarEffect and a benchmark named SemVarBench, designed to evaluate the causality between semantic variations in inputs and outputs in T2I synthesis.
Semantic variations are achieved through two types of linguistic permutations, while avoiding easily predictable literal variations.
Experiments reveal that the CogView-3-Plus and Ideogram 2 performed the best, achieving a score of 0.2/1. Semantic variations in object relations are less understood than attributes, scoring 0.07/1 compared to 0.17-0.19/1.
We found that cross-modal alignment in UNet or Transformers plays a crucial role in handling semantic variations, a factor previously overlooked by a focus on textual encoders.
Our work establishes an effective evaluation framework that advances the T2I synthesis community's exploration of human instruction understanding. Our benchmark and code are available at https://github.com/zhuxiangru/SemVarBench.
****
Divij Handa,Pavel Dolin,Shrinidhi Kumbhar,Tran Cao Son,Chitta Baral
ActionReasoningBench: Reasoning about Actions with and without Ramification Constraints
Reasoning about Actions and Change (RAC) has historically played a pivotal role in solving foundational AI problems, such as the frame problem. It has driven advancements in AI fields, such as non-monotonic and commonsense reasoning. RAC remains crucial for AI systems that operate in dynamic environments, engage in interactive scenarios, or rely on commonsense reasoning. Despite substantial advances made by Large Language Models (LLMs) in various AI domains, their performance in RAC remains underexplored. To address this gap, we introduce a new diagnosti

c benchmark, $\textbf{ActionReasoningBench}$, which encompasses 8 domains and in
cludes questions for up to 19 action sequences. This benchmark rigorously evalua
tes LLMs across six key RAC dimensions: $\textit{Fluent Tracking}$, $\textit{Sta
te Tracking}$, $\textit{Action Executability}$, $\textit{Effects of Actions}$, $
\textit{Numerical RAC}$, and $\textit{Composite Questions}$. LLMs demonstrate av
erage accuracy rates of 73.55%, 65.63%, 58.73%, and 62.38% on the former four di
mensions, which are frequently discussed in RAC literature. However, the perform
ance on the latter two dimensions, which introduce complex and novel reasoning q
uestions, the average performance of LLMs is lowered to 33.16% and 51.19%, respe
ctively, reflecting a 17.9% performance decline. We also introduce new ramificat
ion constraints to capture the indirect effects of actions, providing deeper ins
ights into RAC challenges. Our evaluation of state-of-the-art LLMs, including bo
th open-source and commercial models, reveals challenges across all RAC dimensio
ns, particularly in handling ramifications, with GPT-4o failing to solve any que
stion and o1-preview achieving a score of only 18.4%.
****
Kasia Kobalczyk,Mihaela van der Schaar
Towards Automated Knowledge Integration From Human-Interpretable Representations
A significant challenge in machine learning, particularly in noisy and low-data
environments, lies in effectively incorporating inductive biases to enhance data
 efficiency and robustness. Despite the success of informed machine learning met
hods, designing algorithms with explicit inductive biases remains largely a manu
al process. In this work, we explore how prior knowledge represented in its nati
ve formats, e.g. in natural language, can be integrated into machine learning mo
dels in an automated manner. Inspired by the learning to learn principles of met
a-learning, we consider the approach of learning to integrate knowledge via cond
itional meta-learning, a paradigm we refer to as informed meta-learning. We intr
oduce and motivate theoretically the principles of informed meta-learning enabli
ng automated and controllable inductive bias selection. To illustrate our claims
, we implement an instantiation of informed meta-learning--the Informed Neural P
rocess, and empirically demonstrate the potential benefits and limitations of in
formed meta-learning in improving data efficiency and generalisation.
****
Issar Tzachor,Boaz Lerner,Matan Levy,Michael Green,Tal Berkovitz Shalev,Gavriel
Habib,Dvir Samuel,Noam Korngut Zailer,Or Shimshi,Nir Darshan,Rami Ben-Ari
EffoVPR: Effective Foundation Model Utilization for Visual Place Recognition
The task of Visual Place Recognition (VPR) is to predict the location of a query
 image from a database of geo-tagged images. Recent studies in VPR have highligh
ted the significant advantage of employing pre-trained foundation models like DI
NOv2 for the VPR task. However, these models are often deemed inadequate for VPR
 without further fine-tuning on VPR-specific data.
In this paper, we present an effective approach to harness the potential of a fo
undation model for VPR. We show that features extracted from self-attention laye
rs can act as a powerful re-ranker for VPR, even in a zero-shot setting. Our met
hod not only outperforms previous zero-shot approaches but also introduces resul
ts competitive with several supervised methods.
We then show that a single-stage approach utilizing internal ViT layers for pool
ing can produce global features that achieve state-of-the-art performance, with
impressive feature compactness down to 128D. Moreover, integrating our local fou
ndation features for re-ranking further widens this performance gap. Our method
also demonstrates exceptional robustness and generalization, setting new state-o
f-the-art performance, while handling challenging conditions such as occlusion,
day-night transitions, and seasonal variations.
****
Jonas Hübotter,Sascha Bongni,Ido Hakimi,Andreas Krause
Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs
Recent efforts in fine-tuning language models often rely on automatic data selec
tion, commonly using Nearest Neighbors retrieval from large datasets.
However, we theoretically show that this approach tends to select redundant data
, limiting its effectiveness or even hurting performance.

To address this, we introduce SIFT, a data selection algorithm designed to reduce uncertainty about the model's response given a prompt, which unifies ideas from retrieval and active learning.
Whereas Nearest Neighbor retrieval typically fails in the presence of information duplication, SIFT accounts for information duplication and optimizes the overall information gain of the selected examples.
We focus our evaluations on fine-tuning at test-time for prompt-specific language modeling on the Pile dataset, and show that SIFT consistently outperforms Nearest Neighbor retrieval, with minimal computational overhead.
Moreover, we show that our uncertainty estimates can predict the performance gain of test-time fine-tuning, and use this to develop an adaptive algorithm that invests test-time compute proportional to realized performance gains.
We provide the `activeft` (Active Fine-Tuning) library which can be used as a drop-in replacement for Nearest Neighbor retrieval.
****
Jiacheng Ye,Jiahui Gao,Shansan Gong,Lin Zheng,Xin Jiang,Zhenguo Li,Lingpeng Kong
Beyond Autoregression: Discrete Diffusion for Complex Reasoning and Planning
Autoregressive language models, despite their impressive capabilities, struggle with complex reasoning and long-term planning tasks. We introduce discrete diffusion models as a novel solution to these challenges. Through the lens of subgoal imbalance, we demonstrate how diffusion models effectively learn difficult subgoals that elude autoregressive approaches. We propose Multi-Granularity Diffusion Modeling (MGDM), which prioritizes subgoals based on difficulty during learning. On complex tasks like Countdown, Sudoku, and Boolean Satisfiability Problems, MGDM significantly outperforms autoregressive models without using search techniques. For instance, MGDM achieves 91.5\% and 100\% accuracy on Countdown and Sudoku, respectively, compared to 45.8\% and 20.7\% for autoregressive models. Our work highlights the potential of diffusion-based approaches in advancing AI capabilities for sophisticated language understanding and problem-solving tasks. All associated codes are available at \href{https://github.com/HKUNLP/diffusion-vs-ar}{https://github.com/HKUNLP/diffusion-vs-ar}.
****
Khai Nguyen,Hai Nguyen,Nhat Ho
Towards Marginal Fairness Sliced Wasserstein Barycenter
The Sliced Wasserstein barycenter (SWB) is a widely acknowledged method for efficiently generalizing the averaging operation within probability measure spaces. However, achieving marginal fairness SWB, ensuring approximately equal distances from the barycenter to marginals, remains unexplored. The uniform weighted SWB is not necessarily the optimal choice to obtain the desired marginal fairness barycenter due to the heterogeneous structure of marginals and the non-optimality of the optimization. As the first attempt to tackle the problem, we define the marginal fairness sliced Wasserstein barycenter (MFSWB) as a constrained SWB problem. Due to the computational disadvantages of the formal definition, we propose two hyperparameter-free and computationally tractable surrogate MFSWB problems that implicitly minimize the distances to marginals and encourage marginal fairness at the same time. To further improve the efficiency, we perform slicing distribution selection and obtain the third surrogate definition by introducing a new slicing distribution that focuses more on marginally unfair projecting directions. We discuss the relationship of the three proposed problems and their relationship to sliced multi-marginal Wasserstein distance. Finally, we conduct experiments on finding 3D point-clouds averaging, color harmonization, and training of sliced Wasserstein autoencoder with class-fairness representation to show the favorable performance of the proposed surrogate MFSWB problems.
****
Ziniu Li,Congliang Chen,Tian Xu,Zeyu Qin,Jiancong Xiao,Zhi-Quan Luo,Ruoyu Sun
Preserving Diversity in Supervised Fine-Tuning of Large Language Models
Large Language Models (LLMs) typically rely on Supervised Fine-Tuning (SFT) to specialize in downstream tasks, with the Cross Entropy (CE) loss being the de facto choice. However, CE maximizes the likelihood of observed data without accounting for alternative possibilities. As such, CE usually leads to reduced diversi

ty in the model's outputs, which hinders further development that requires sampl
ing to explore better responses. To address this limitation, this paper introduc
es a new game-theoretic formulation for SFT. In this framework, an auxiliary var
iable is introduced to regulate the learning process. We prove that the proposed
 game-theoretic approach connects to the problem of reverse KL minimization with
 entropy regularization. This regularization prevents over-memorization of train
ing data and promotes output diversity. To implement this framework, we develop
GEM, a new training algorithm that is computationally efficient as CE by leverag
ing some unique properties of LLMs. Empirical studies of pre-trained models from
 3B to 70B parameters show that GEM achieves comparable downstream performance t
o CE while significantly enhancing output diversity. This increased diversity tr
anslates to performance gains in test-time compute scaling for chat and code gen
eration tasks. Moreover, we observe that preserving output diversity has the add
ed benefit of mitigating forgetting, as maintaining diverse outputs encourages m
odels to retain pre-trained knowledge throughout the training process.
****
Patrick Seifner,Kostadin Cvejoski,Antonia Körner,Ramses J Sanchez
Zero-shot Imputation with Foundation Inference Models for Dynamical Systems
Dynamical systems governed by ordinary differential equations (ODEs) serve as mo
dels for a vast number of natural and social phenomena. In this work, we offer a
 fresh perspective on the classical problem of imputing missing time series data
, whose underlying dynamics are assumed to be determined by ODEs. Specifically,
we revisit ideas from amortized inference and neural operators, and propose a no
vel supervised learning framework for *zero-shot time series imputation*, throug
h parametric functions satisfying some (hidden) ODEs. Our proposal consists of t
wo components. First, a broad probability distribution over the space of ODE sol
utions, observation times and noise mechanisms, with which we generate a large,
synthetic dataset of (hidden) ODE solutions, along with their noisy and sparse o
bservations. Second, a neural recognition model that is trained *offline*, to ma
p the generated time series onto the spaces of initial conditions and time deriv
atives of the (hidden) ODE solutions, which we then integrate to impute the miss
ing data. We empirically demonstrate that *one and the same* (pretrained) recogn
ition model can perform zero-shot imputation across 63 distinct time series with
 missing values, each sampled from widely different dynamical systems. Likewise,
 we demonstrate that it can perform zero-shot imputation of missing high-dimensi
onal data in 10 vastly different settings, spanning human motion, air quality, t
raffic and electricity studies, as well as Navier-Stokes simulations — *without
requiring any fine-tuning*. What is more, our proposal often outperforms state-o
f-the-art methods, which are trained on the target datasets.

Our pretrained model, repository and tutorials are available online.
****
Jiquan Wang,Sha Zhao,Zhiling Luo,Yangxuan Zhou,Haiteng Jiang,Shijian Li,Tao Li,G
ang Pan
CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding
Electroencephalography (EEG) is a non-invasive technique to measure and record b
rain electrical activity, widely used in various BCI and healthcare applications
. Early EEG decoding methods rely on supervised learning, limited by specific ta
sks and datasets, hindering model performance and generalizability. With the suc
cess of large language models, there is a growing body of studies focusing on EE
G foundation models. However, these studies still leave challenges: Firstly, mos
t of existing EEG foundation models employ full EEG modeling strategy. It models
 the spatial and temporal dependencies between all EEG patches together, but ign
ores that the spatial and temporal dependencies are heterogeneous due to the uni
que structural characteristics of EEG signals. Secondly, existing EEG foundation
 models have limited generalizability on a wide range of downstream BCI tasks du
e to varying formats of EEG data, making it challenging to adapt to. To address
these challenges, we propose a novel foundation model called CBraMod. Specifical
ly, we devise a criss-cross transformer as the backbone to thoroughly leverage t
he structural characteristics of EEG signals, which can model spatial and tempor

al dependencies separately through two parallel attention mechanisms. And we uti
lize an asymmetric conditional positional encoding scheme which can encode posit
ional information of EEG patches and be easily adapted to the EEG with diverse f
ormats. CBraMod is pre-trained on a very large corpus of EEG through patch-based
 masked EEG reconstruction. We evaluate CBraMod on up to 10 downstream BCI tasks
 (12 public datasets). CBraMod achieves the state-of-the-art performance across
the wide range of tasks, proving its strong capability and generalizability. The
 source code is publicly available at https://github.com/wjq-learning/CBraMod.
****

Florian E. Dorner,Vivian Yvonne Nastl,Moritz Hardt

Limits to scalable evaluation at the frontier: LLM as judge won’t beat twice the
 data

High quality annotations are increasingly a bottleneck in the explosively growin
g machine learning ecosystem. Scalable evaluation methods that avoid costly anno
tation have therefore become an important research ambition. Many hope to use st
rong existing models in lieu of costly labels to provide cheap model evaluations
. Unfortunately, this method of using models as judges introduces biases, such a
s self-preferencing, that can distort model comparisons. An emerging family of d
ebiasing tools promises to fix these issues by using a few high quality labels t
o debias a large number of model judgments. In this paper, we study how far such
 debiasing methods, in principle, can go. Our main result shows that when the ju
dge is no more accurate than the evaluated model, no debiasing method can decrea
se the required amount of ground truth labels by more than half. Our result spea
ks to the severe limitations of the LLM-as-a-judge paradigm at the evaluation fr
ontier where the goal is to assess newly released models that are possibly bette
r than the judge. Through an empirical evaluation, we demonstrate that the sampl
e size savings achievable in practice are even more modest than what our theoret
ical limit suggests. Along the way, our work provides new observations about deb
iasing methods for model evaluation, and points out promising avenues for future
 work.
****

Xinran Wang,Qi Le,Ammar Ahmed,Enmao Diao,Yi Zhou,Nathalie Baracaldo,Jie Ding,Ali
 Anwar

MAP: Multi-Human-Value Alignment Palette

Ensuring that generative AI systems align with human values is essential but cha
llenging, especially when considering multiple human values and their potential
trade-offs. Since human values can be personalized and dynamically change over t
ime, the desirable levels of value alignment vary across different ethnic groups
, industry sectors, and user cohorts. Within existing frameworks, it is hard to
define human values and align AI systems accordingly across different directions
 simultaneously, such as harmlessness, helpfulness, and positiveness. To address
 this, we develop a novel, first-principle approach called Multi-Human-Value Ali
gnment Palette (MAP), which navigates the alignment across multiple human values
 in a structured and reliable way. MAP formulates the alignment problem as an op
timization task with user-defined constraints, which define human value targets.
 It can be efficiently solved via a primal-dual approach, which determines wheth
er a user-defined alignment target is achievable and how to achieve it. We condu
ct a detailed theoretical analysis of MAP by quantifying the trade-offs between
values, the sensitivity to constraints, the fundamental connection between multi
-value alignment and sequential alignment, and proving that linear weighted rewa
rds are sufficient for multi-value alignment. Extensive experiments demonstrate
MAP's ability to align multiple values in a principled manner while delivering s
trong empirical performance across various tasks.
****

Zijian Liu,Zhengyuan Zhou

Nonconvex Stochastic Optimization under Heavy-Tailed Noises: Optimal Convergence
 without Gradient Clipping

Recently, the study of heavy-tailed noises in first-order nonconvex stochastic o
ptimization has gotten a lot of attention since it was recognized as a more real
istic condition as suggested by many empirical observations. Specifically, the s

tochastic noise (the difference between the stochastic and true gradient) is con
sidered to have only a finite $\mathfrak{p}$-th moment where $\mathfrak{p}\in\le
ft(1,2\right]$ instead of assuming it always satisfies the classical finite vari
ance assumption. To deal with this more challenging setting, people have propose
d different algorithms and proved them to converge at an optimal $\mathcal{O}(T^
{\frac{1-\mathfrak{p}}{3\mathfrak{p}-2}})$ rate for smooth objectives after T
iterations. Notably, all these new-designed algorithms are based on the same tec
hnique – gradient clipping. Naturally, one may want to know whether the clipping
 method is a necessary ingredient and the only way to guarantee convergence unde
r heavy-tailed noises. In this work, by revisiting the existing Batched Normaliz
ed Stochastic Gradient Descent with Momentum (Batched NSGDM) algorithm, we provi
de the first convergence result under heavy-tailed noises but without gradient c
lipping. Concretely, we prove that Batched NSGDM can achieve the optimal $\mathc
al{O}(T^{\frac{1-\mathfrak{p}}{3\mathfrak{p}-2}})$ rate even under the relaxed s
mooth condition. More interestingly, we also establish the first $\mathcal{O}(T^
{\frac{1-\mathfrak{p}}{2\mathfrak{p}}})$ convergence rate in the case where the
tail index $\mathfrak{p}$ is unknown in advance, which is arguably the common sc
enario in practice.
****

Chenwei Wu,Zitao Shuai,Zhengxu Tang,Luning Wang,Liyue Shen
Dynamic Modeling of Patients, Modalities and Tasks via Multi-modal Multi-task Mi
xture of Experts
Multi-modal multi-task learning holds significant promise in tackling complex di
agnostic tasks and many significant medical imaging problems. It fulfills the ne
eds in real-world diagnosis protocol to leverage information from different data
 sources and simultaneously perform mutually informative tasks. However, medical
 imaging domains introduce two key challenges: dynamic modality fusion and modal
ity-task dependence. The quality and amount of task-related information from dif
ferent modalities could vary significantly across patient samples, due to biolog
ical and demographic factors. Traditional fusion methods apply fixed combination
 strategies that fail to capture this dynamic relationship, potentially underuti
lizing modalities that carry stronger diagnostic signals for specific patients.
Additionally, different clinical tasks may require dynamic feature selection and
 combination from various modalities, a phenomenon we term “modality-task depend
ence.” To address these issues, we propose M4oE, a novel Multi-modal Multi-task
Mixture of Experts framework for precise Medical diagnosis. M4oE comprises Modal
ity-Specific (MSoE) modules and a Modality-shared Modality-Task MoE (MToE) modul
e. With collaboration from both modules, our model dynamically decomposes and le
arns distinct and shared information from different modalities and achieves dyna
mic fusion. MToE provides a joint probability model of modalities and tasks by u
sing experts as a link and encourages experts to learn modality-task dependence
via conditional mutual information loss. By doing so, M4oE offers sample and pop
ulation-level interpretability of modality contributions. We evaluate M4oE on fo
ur public multi-modal medical benchmark datasets for solving two important medic
al diagnostic problems including breast cancer screening and retinal disease dia
gnosis. Results demonstrate our method's superiority over state-of-the-art metho
ds under different metrics of classification and segmentation tasks like Accurac
y, AUROC, AUPRC, and DICE.
****

Alex Lewandowski,Dale Schuurmans,Marlos C. Machado
Plastic Learning with Deep Fourier Features
Deep neural networks can struggle to learn continually in the face of non-statio
narity, a
  phenomenon known as loss of plasticity.
  In this paper, we identify underlying principles that lead to plastic algorith
ms.
  We provide theoretical results showing that linear function approximation, as
well as a special case of deep linear networks, do not suffer from loss of plast
icity.
  We then propose deep Fourier features, which are the concatenation of a sine a

nd cosine in every layer, and we show that this combination provides a dynamic b
alance between the trainability obtained through linearity and the effectiveness
 obtained through the nonlinearity of neural networks.
  Deep networks composed entirely of deep Fourier features are highly trainable
and sustain their trainability over the course of learning.
  Our empirical results show that continual learning performance can be improved
 by replacing ReLU activations with deep Fourier features combined with regulari
zation.
  These results hold for different continual learning scenarios (e.g., label noi
se, class incremental learning, pixel permutations)
  on all major supervised learning datasets used for continual learning research
, such as CIFAR10, CIFAR100, and tiny-ImageNet.
****
Sohan Patnaik,Milan Aggarwal,Sumit Bhatia,Balaji Krishnamurthy
It Helps to Take a Second Opinion: Teaching Smaller LLMs To Deliberate Mutually
via Selective Rationale Optimisation
Very large language models (LLMs) such as GPT-4 have shown the ability to handle
 complex tasks by generating and self-refining step-by-step rationales. Smaller
language models (SLMs), typically with < 13B parameters, have been improved by u
sing the data generated from very-large LMs through knowledge distillation. Howe
ver, various practical constraints such as API costs, copyright, legal and ethic
al policies restrict using large (often opaque) models to train smaller models f
or commercial use. Limited success has been achieved at improving the ability of
 an SLM to explore the space of possible rationales and evaluate them by itself
through self-deliberation. To address this, we propose COALITION, a trainable fr
amework that facilitates interaction between two variants of the same SLM and tr
ains them to generate and refine rationales optimized for the end-task. The vari
ants exhibit different behaviors to produce a set of diverse candidate rationale
s during the generation and refinement steps. The model is then trained via Sele
ctive Rationale Optimization (SRO) to prefer generating rationale candidates tha
t maximize the likelihood of producing the ground-truth answer. During inference
, COALITION employs a controller to select the suitable variant for generating a
nd refining the rationales. On five different datasets covering mathematical pro
blems, commonsense reasoning, and natural language inference, COALITION outperfo
rms several baselines by up to 5%. Our ablation studies reveal that cross-commun
ication between the two variants performs better than using the single model to
self-refine the rationales. We also demonstrate the applicability of COALITION f
or LMs of varying scales (4B to 14B parameters) and model families (Mistral, Lla
ma, Qwen, Phi). We release the code for this work here.
****
Renpu Liu,Ruida Zhou,Cong Shen,Jing Yang
On the Learn-to-Optimize Capabilities of Transformers in In-Context Sparse Recov
ery
An intriguing property of the Transformer is its ability to perform in-context l
earning (ICL), where the Transformer can solve different inference tasks without
 parameter updating based on the contextual information provided by the correspo
nding input-output demonstration pairs. It has been theoretically proved that IC
L is enabled by the capability of Transformers to perform gradient-descent algor
ithms (Von Oswald et al., 2023a; Bai et al., 2024). This work takes a step furth
er and shows that Transformers can perform learning-to-optimize (L2O) algorithms
. Specifically, for the ICL sparse recovery (formulated as LASSO) tasks, we show
 that a K-layer Transformer can perform an L2O algorithm with a provable converg
ence rate linear in K. This provides a new perspective explaining the superior I
CL capability of Transformers, even with only a few layers, which cannot be achi
eved by the standard gradient-descent algorithms. Moreover, unlike the conventio
nal L2O algorithms that require the measurement matrix involved in training to m
atch that in testing, the trained Transformer is able to solve sparse recovery p
roblems generated with different measurement matrices. Besides, Transformers as
 an L2O algorithm can leverage structural information embedded in the training t
asks to accelerate its convergence during ICL, and generalize across different l

engths of demonstration pairs, where conventional L2O algorithms typically strug
gle or fail. Such theoretical findings are supported by our experimental results
.
****
Jincheng Wang,Penny Karanasou,Pengyuan Wei,Elia Gatti,Diego Martinez Plasencia,D
imitrios Kanoulas
Long-Short Decision Transformer: Bridging Global and Local Dependencies for Gene
ralized Decision-Making
Decision Transformers (DTs) effectively capture long-range dependencies using se
lf-attention but struggle with fine-grained local relationships, especially the
Markovian properties in many offline-RL datasets. Conversely, Decision Convforme
r (DC) utilizes convolutional filters for capturing local patterns but shows lim
itations in tasks demanding long-term dependencies, such as Maze2d. To address t
hese limitations and leverage both strengths, we propose the Long-Short Decision
 Transformer (LSDT), a general-purpose architecture to effectively capture globa
l and local dependencies across two specialized parallel branches (self-attentio
n and convolution). We explore how these branches complement each other by model
ing various ranged dependencies across different environments, and compare it ag
ainst other baselines. Experimental results demonstrate our LSDT achieves state-
of-the-art performance and notable gains over the standard DT in D4RL offline RL
 benchmark. Leveraging the parallel architecture, LSDT performs consistently on
diverse datasets, including Markovian and non-Markovian. We also demonstrate the
 flexibility of LSDT's architecture, where its specialized branches can be repla
ced or integrated into models like DC to improve their performance in capturing
diverse dependencies. Finally, we also highlight the role of goal states in impr
oving decision-making for goal-reaching tasks like Antmaze.
****
Chen Sun,Renat Aksitov,Andrey Zhmoginov,Nolan Andrew Miller,Max Vladymyrov,Ulric
h Rueckert,Been Kim,Mark Sandler
How new data permeates LLM knowledge and how to dilute it
Large language models continually learn through the accumulation of gradient-bas
ed updates, but how individual pieces of new information affect existing knowled
ge, leading to both beneficial generalization and problematic hallucination, rem
ains poorly understood. We demonstrate that when learning new information, LLMs
exhibit a "priming" effect: learning a new fact can cause the model to inappropr
iately apply that knowledge in unrelated contexts.
To systematically study this phenomenon, we introduce "Outlandish," a carefully
curated dataset of 1320 diverse text samples designed to probe how new knowledge
 permeates through an LLM's existing knowledge base. Using this dataset, we show
 that the degree of priming after learning new information can be predicted by m
easuring the token probability of key words before training. This relationship h
olds robustly across different model architectures (PALM-2, Gemma, Llama), sizes
, and training stages.
Finally, we develop two novel techniques to modulate how new knowledge affects e
xisting model behavior: (1) a ``stepping-stone'' text augmentation strategy and
(2) an ``ignore-k'' update pruning method. These approaches reduce undesirable p
riming effects by 50-95% while preserving the model's ability to learn new infor
mation. Our findings provide both empirical insights into how LLMs learn and pra
ctical tools for improving the specificity of knowledge insertion in language mo
dels. Further materials: https://sunchipsster1.github.io/projects/outlandish/
****
Qi CHEN,Jierui Zhu,Florian Shkurti
Generalization in VAE and Diffusion Models: A Unified Information-Theoretic Anal
ysis
Despite the empirical success of Diffusion Models (DMs) and Variational Autoenco
ders (VAEs), their generalization performance remains theoretically underexplore
d, especially lacking a full consideration of the shared encoder-generator struc
ture. Leveraging recent information-theoretic tools, we propose a unified theore
tical framework that provides guarantees for the generalization of both the enco
der and generator by treating them as randomized mappings. This framework furthe

r enables (1) a refined analysis for VAEs, accounting for the generator's genera
lization, which was previously overlooked; (2) illustrating an explicit trade-of
f in generalization terms for DMs that depends on the diffusion time T; and (3
) providing computable bounds for DMs based solely on the training data, allowin
g the selection of the optimal T and the integration of such bounds into the o
ptimization process to improve model performance. Empirical results on both synt
hetic and real datasets illustrate the validity of the proposed theory.
****
Hang Wang,Xin Ye,Feng Tao,Chenbin Pan,Abhirup Mallik,Burhaneddin Yaman,Liu Ren,J
unshan Zhang
AdaWM: Adaptive World Model based Planning for Autonomous Driving
World model based reinforcement learning (RL) has emerged as a promising approac
h for autonomous driving, which learns a latent dynamics model and uses it to tr
ain a planning policy. To speed up the learning process, the pretrain-finetune
paradigm is often used, where online RL is initialized by a pretrained model an
d a policy learned offline. However, naively performing such initialization in R
L may result in dramatic performance degradation during the online interactions
in the new task. To tackle this challenge, we first analyze the performance deg
radation and identify two primary root causes therein: the mismatch of the plann
ing policy and the mismatch of the dynamics model, due to distribution shift. W
e further analyze the effects of these factors on performance degradation durin
g finetuning, and our findings reveal that the choice of finetuning strategies p
lays a pivotal role in mitigating these effects. We then introduce AdaWM, an Ada
ptive World Model based planning method, featuring two key steps: (a) mismatch i
dentification, which quantifies the mismatches and informs the finetuning strate
gy, and (b) alignment-driven finetuning, which selectively updates either the po
licy or the model as needed using efficient low-rank updates. Extensive experim
ents on the challenging CARLA driving tasks demonstrate that AdaWM significantl
y improves the finetuning process, resulting in more robust and efficient perfor
mance in autonomous driving systems.
****
Sua Lee,Kyubum Shin,Jung Ho Park
Weighted Multi-Prompt Learning with Description-free Large Language Model Distil
lation
Recent advances in pre-trained Vision Language Models (VLM) have shown promising
potential for effectively adapting to downstream tasks through _prompt learning
_, without the need for additional annotated paired datasets.
To supplement the text information in VLM trained on correlations with vision da
ta, new approaches leveraging Large Language Models (LLM) in prompts have been p
roposed, enhancing robustness to unseen and diverse data.
Existing methods typically extract text-based responses (i.e., _descriptions_) f
rom LLM to incorporate into prompts; however, this approach suffers from high va
riability and low reliability.
In this work, we propose **De**scription-free **Mul**ti-prompt Learning(**DeMul*
*), a novel method that eliminates the process of extracting descriptions and in
stead directly distills knowledge from LLM into prompts.
By adopting a description-free approach, prompts can encapsulate richer semantic
s while still being represented as continuous vectors for optimization, thereby
eliminating the need for discrete pre-defined templates.
Additionally, in a multi-prompt setting, we empirically demonstrate the potentia
l of prompt weighting in reflecting the importance of different prompts during t
raining.
Experimental results show that our approach achieves superior performance across
11 recognition datasets.
****
Ruochen Zhang,Qinan Yu,Matianyu Zang,Carsten Eickhoff,Ellie Pavlick
The Same but Different: Structural Similarities and Differences in Multilingual
Language Modeling
We employ new tools from mechanistic interpretability to ask whether the interna
l structure of large language models (LLMs) shows correspondence to the linguist

ic structures which underlie the languages on which they are trained. In particu
lar, we ask (1) when two languages employ the same morphosyntactic processes, do
 LLMs handle them using shared internal circuitry? and (2) when two languages re
quire different morphosyntactic processes, do LLMs handle them using different i
nternal circuitry? In a focused case study on English and Chinese multilingual a
nd monolingual models, we analyze the internal circuitry involved in two tasks.
We find evidence that models employ the same circuit to handle the same syntacti
c process independently of the language in which it occurs, and that this is the
 case even for monolingual models trained completely independently. Moreover, we
 show that multilingual models employ language-specific components (attention he
ads and feed-forward networks) when needed to handle linguistic processes (e.g.,
 morphological marking) that only exist in some languages. Together, our results
 are revealing about how LLMs trade off between exploiting common structures and
 preserving linguistic differences when tasked with modeling multiple languages
simultaneously, opening the door for future work in this direction.
****
Yu Wu,Yansong Li,Zeyu Dong,Nitya Sathyavageeswaran,Anand D. Sarwate
Learning to Help in Multi-Class Settings
Deploying complex machine learning models on resource-constrained devices is cha
llenging due to limited computational power, memory, and model retrainability. T
o address these limitations, a hybrid system can be established by augmenting th
e local model with a server-side model, where samples are selectively deferred b
y a *rejector* and then sent to the server for processing. The hybrid system ena
bles efficient use of computational resources while minimizing the overhead asso
ciated with server usage. The recently proposed Learning to Help (L2H) model pro
posed training a server model given a fixed local (client) model. This differs f
rom the Learning to Defer (L2D) framework which trains the client for a fixed (e
xpert) server. In both L2D and L2H, the training includes learning a rejector at
 the client to determine when to query the server. In this work, we extend the L
2H model from binary to multi-class classification problems and demonstrate its
applicability in a number of different scenarios of practical interest in which
access to the server may be limited by cost, availability, or policy. We derive
a stage-switching surrogate loss function that is differentiable, convex, and co
nsistent with the Bayes rule corresponding to the 0-1 loss for the L2H model. Ex
periments show that our proposed methods offer an efficient and practical soluti
on for multi-class classification in resource-constrained environments.
****
T-H. Hubert Chan,Shaofeng H.-C. Jiang,Tianyi Wu,Mengshi Zhao
Online Clustering with Nearly Optimal Consistency
We give online algorithms for k-Means(more generally, (k, z)-Clustering) wit
h nearly optimal consistency (a notion suggested by Lattanzi & Vassilvitskii (20
17)).
Our result turns any α-approximate offline algorithm for clustering into
an $(1+\epsilon)\alpha^2$-competitive online algorithm for clustering with $O(k
\text{poly} \log n)$ consistency.
This consistency bound is optimal up to $\text{poly} \log(n)$ factors.
Plugging in the offline algorithm that returns the exact optimal solution,
we obtain the first
$(1 + \epsilon)$-competitive online algorithm for clustering that achieves a lin
ear in k consistency.
This simultaneously improves several previous results (Lattanzi & Vassilvitskii,
 2017; Fichtenberger et al., 2021).
We validate the performance of our algorithm on real datasets by plugging in the
 practically efficient k-Means++ algorithm.
Our online algorithm makes k-Means++ achieve good consistency with little over
head to the quality of solutions.
****
Taha EL BAKKALI EL KADI,Omar Saadi
On the Almost Sure Convergence of the Stochastic Three Points Algorithm
The stochastic three points (STP) algorithm is a derivative-free optimization te

chnique designed for unconstrained optimization problems in $\mathbb{R}^d$. In t
his paper, we analyze this algorithm for three classes of functions: smooth func
tions that may lack convexity, smooth convex functions, and smooth functions tha
t are strongly convex. Our work provides the first almost sure convergence resul
ts of the STP algorithm, alongside some convergence results in expectation.
For the class of smooth functions, we establish that the best gradient iterate o
f the STP algorithm converges almost surely to zero at a rate of $o(1/{T^{\frac{
1}{2}-\epsilon}})$ for any $\epsilon\in (0,\frac{1}{2})$, where T is the numbe
r of iterations. Furthermore, within the same class of functions, we establish b
oth almost sure convergence and convergence in expectation of the final gradient
 iterate towards zero.
For the class of smooth convex functions, we establish that $f(\theta^T)$ conver
ges to $\inf_{\theta \in \mathbb{R}^d} f(\theta)$ almost surely at a rate of $o(
1/{T^{1-\epsilon}})$ for any $\epsilon\in (0,1)$, and in expectation at a rate o
f $O(\frac{d}{T})$ where d is the dimension of the space.
Finally, for the class of smooth functions that are strongly convex, we establis
h that when step sizes are obtained by approximating the directional derivatives
 of the function, $f(\theta^T)$ converges to $\inf_{\theta \in \mathbb{R}^d} f(\
theta)$ in expectation at a rate of $O((1-\frac{\mu}{dL})^T)$, and almost surely
 at a rate of $o((1-s\frac{\mu}{dL})^T)$ for any $s\in (0,1)$, where μ and
L
are the strong convexity and smoothness parameters of the function.
****
Enze Xie,Junsong Chen,Junyu Chen,Han Cai,Haotian Tang,Yujun Lin,Zhekai Zhang,Muy
ang Li,Ligeng Zhu,Yao Lu,Song Han
SANA: Efficient High-Resolution Text-to-Image Synthesis with Linear Diffusion Tr
ansformers
We introduce Sana, a text-to-image framework that can efficiently generate image
s up to 4096$\times$4096 resolution. Sana can synthesize high-resolution, high-q
uality images with strong text-image alignment at a remarkably fast speed, deplo
yable on laptop GPU. Core designs include: (1) Deep compression autoencoder: unl
ike traditional AEs, which compress images only 8$\times$, we trained an AE that
 can compress images 32$\times$, effectively reducing the number of latent token
s. (2) Linear DiT: we replace all vanilla attention in DiT with linear attention
, which is more efficient at high resolutions without sacrificing quality. (3) D
ecoder-only text encoder: we replaced T5 with modern decoder-only small LLM as t
he text encoder and designed complex human instruction with in-context learning
to enhance the image-text alignment. (4) Efficient training and sampling: we pr
opose Flow-DPM-Solver to reduce sampling steps, with efficient caption labeling
and selection to accelerate convergence. As a result, Sana-0.6B is very competit
ive with modern giant diffusion model (e.g. Flux-12B), being 20 times smaller an
d 100+ times faster in measured throughput. Moreover, Sana-0.6B can be deployed
on a 16GB laptop GPU, taking less than 1 second to generate a 1024$\times$1024 r
esolution image. Sana enables content creation at low cost. Code and model will
be publicly released upon publication.
****
Chenxiao Yang,Zhiyuan Li,David Wipf
Chain-of-Thought Provably Enables Learning the (Otherwise) Unlearnable
Modern language models have demonstrated remarkable reasoning capabilities by us
ing chain-of-thought (CoT). One hypothesis about the inner workings of CoT is th
at it breaks down originally complex tasks into smaller subtasks that are more a
menable to learning. We formalize this notion by showing possibility and impossi
bility results of learning from in-context demonstrations with and without CoT.
In particular, with CoT, we examine a family of learning algorithms that learn a
 task step-by-step, capable of composing simpler functions from individual reaso
ning steps to form an overall complex function. This process reduces the difficu
lty of learning a task to that of the hardest reasoning step in the chain. Moreo
ver, we prove Transformers can express this algorithm and thus they can efficien
tly in-context learn arbitrary tasks as long as these tasks can be decomposed in
to a finite number of subtasks, each of which are efficiently learnable. In cont

rast, without CoT, we demonstrate that there exist tasks that are inherently unl
earnable by the same algorithm. Overall, our results suggest several provably ef
fective ways for decomposing target problems to instantiate CoT. Empirically, we
 demonstrate our proposed CoT construction significantly enhances the reasoning
capabilities of real-world LLMs in solving challenging arithmetic reasoning task
s, including learning polynomials and Boolean formulas.
****

Hengzhuang Li,Teng Zhang

Outlier Synthesis via Hamiltonian Monte Carlo for Out-of-Distribution Detection

Out-of-distribution (OOD) detection is crucial for developing trustworthy and re
liable machine learning systems. Recent advances in training with auxiliary OOD
data demonstrate efficacy in enhancing detection capabilities. Nonetheless, thes
e methods heavily rely on acquiring a large pool of high-quality natural outlier
s. Some prior methods try to alleviate this problem by synthesizing virtual outl
iers but suffer from either poor quality or high cost due to the monotonous samp
ling strategy and the heavy-parameterized generative models. In this paper, we o
vercome all these problems by proposing the Hamiltonian Monte Carlo Outlier Synt
hesis (HamOS) framework, which views the synthesis process as sampling from Mark
ov chains. Based solely on the in-distribution data, the Markov chains can exten
sively traverse the feature space and generate diverse and representative outlie
rs, hence exposing the model to miscellaneous potential OOD scenarios. The Hamil
tonian Monte Carlo with sampling acceptance rate almost close to 1 also makes ou
r framework enjoy great efficiency. By empirically competing with SOTA baselines
 on both standard and large-scale benchmarks, we verify the efficacy and efficie
ncy of our proposed HamOS.
****

Keisuke Kamahori,Tian Tang,Yile Gu,Kan Zhu,Baris Kasikci

Fiddler: CPU-GPU Orchestration for Fast Inference of Mixture-of-Experts Models

Large Language Models (LLMs) with the Mixture-of-Experts (MoE) architectures hav
e shown promising performance on various tasks. However, due to the huge model s
izes, running them in resource-constrained environments where the GPU memory is
not abundant is challenging. Some existing systems propose to use CPU resources
to solve that, but they either suffer from the significant overhead of frequentl
y moving data between CPU and GPU, or fail to consider distinct characteristics
of CPUs and GPUs. This paper proposes Fiddler, a resource-efficient inference sy
stem for MoE models with limited GPU resources. Fiddler strategically utilizes C
PU and GPU resources by determining the optimal execution strategy. Our evaluati
on shows that, unlike state-of-the-art systems that optimize for specific scenar
ios such as single batch inference or long prefill, Fiddler performs better in a
ll scenarios. Compared against different baselines, Fiddler achieves 1.26 times
speed up in single batch inference, 1.30 times in long prefill processing, and 1
1.57 times in beam search inference. The code of Fiddler is publicly available a
t https://github.com/efeslab/fiddler.
****

Yatai Ji,Shilong Zhang,Jie Wu,Peize Sun,Weifeng Chen,Xuefeng Xiao,Sidi Yang,Yuji
u Yang,Ping Luo

IDA-VLM: Towards Movie Understanding via ID-Aware Large Vision-Language Model

The rapid advancement of Large Vision-Language models (LVLMs) has demonstrated a
 spectrum of emergent capabilities. Nevertheless, current models only focus on t
he visual content of a single scenario, while their ability to associate instanc
es across different scenes has not yet been explored, which is essential for und
erstanding complex visual content, such as movies with multiple characters and i
ntricate plots. Towards movie understanding, a critical initial step for LVLMs i
s to unleash the potential of character identities memory and recognition across
 multiple visual scenarios. To achieve the goal, we propose visual instruction t
uning with ID reference and develop an ID-Aware Large Vision-Language Model, IDA
-VLM. Furthermore, our research introduces a novel benchmark MM-ID, to examine L
VLMs on instance IDs memory and recognition across four dimensions: matching, lo
cation, question-answering, and captioning. Our findings highlight the limitatio
ns of existing LVLMs in recognizing and associating instance identities with ID

reference. This paper paves the way for future artificial intelligence systems t
o possess multi-identity visual inputs, thereby facilitating the comprehension o
f complex visual narratives like movies.
****

Chunkai Fu,Brandon G. Nguyen,Jung Hoon Seo,Ryan S. Zesch,Samson Zhou
Learning-Augmented Search Data Structures
We study the integration of machine learning advice to improve upon traditional
data structure designed for efficient search queries. Although there has been re
cent effort in improving the performance of binary search trees using machine le
arning advice, e.g., Lin et. al. (ICML 2022), the resulting constructions never
theless suffer from inherent weaknesses of binary search trees, such as complexi
ty of maintaining balance across multiple updates and the inability to handle pa
rtially-ordered or high-dimensional datasets. For these reasons, we focus on ski
p lists and KD trees in this work. Given access to a possibly erroneous oracle t
hat outputs estimated fractional frequencies for search queries on a set of item
s, we construct skip lists and KD trees that provably provides the optimal expec
ted search time, within nearly a factor of two. In fact, our learning-augmented
skip lists and KD trees are still optimal up to a constant factor, even if the o
racle is only accurate within a constant factor. We also demonstrate robustness
by showing that our data structures achieves an expected search time that is wit
hin a constant factor of an oblivious skip list/KD tree construction even when t
he predictions are arbitrarily incorrect. Finally, we empirically show that our
learning-augmented search data structures outperforms their corresponding tradit
ional analogs on both synthetic and real-world datasets.
****

Ge Li,Dong Tian,Hongyi Zhou,Xinkai Jiang,Rudolf Lioutikov,Gerhard Neumann
TOP-ERL: Transformer-based Off-Policy Episodic Reinforcement Learning
This work introduces Transformer-based Off-Policy Episodic Reinforcement Learnin
g (TOP-ERL), a novel algorithm that enables off-policy updates in the ERL framew
ork. In ERL, policies predict entire action trajectories over multiple time step
s instead of single actions at every time step. These trajectories are typically
 parameterized by trajectory generators such as Movement Primitives (MP), allow
ing for smooth and efficient exploration over long horizons while capturing high
-level temporal correlations. However, ERL methods are often constrained to on-p
olicy frameworks due to the difficulty of evaluating state-action values for ent
ire action sequences, limiting their sample efficiency and preventing the use of
 more efficient off-policy architectures. TOP-ERL addresses this shortcoming by
segmenting long action sequences and estimating the state-action values for each
 segment using a transformer-based critic architecture alongside an n-step retur
n estimation. These contributions result in efficient and stable training that i
s reflected in the empirical results conducted on sophisticated robot learning e
nvironments. TOP-ERL significantly outperforms state-of-the-art RL methods. Thor
ough ablation studies additionally show the impact of key design choices on the
model performance.
****

Yucheng Zhou,Jianbing Shen,Yu Cheng
Weak to Strong Generalization for Large Language Models with Multi-capabilities
As large language models (LLMs) grow in sophistication, some of their capabiliti
es surpass human abilities, making it essential to ensure their alignment with h
uman values and intentions, i.e., Superalignment. This superalignment challenge
is particularly critical for complex tasks, as annotations provided by humans, a
s weak supervisors, may be overly simplistic, incomplete, or incorrect. Previous
 work has demonstrated the potential of training a strong model using the weak d
ataset generated by a weak model as weak supervision. However, these studies hav
e been limited to a single capability. In this work, we conduct extensive experi
ments to investigate weak to strong generalization for LLMs with multi-capabilit
ies. The experiments reveal that different capabilities tend to remain relativel
y independent in this generalization, and the effectiveness of weak supervision
is significantly impacted by the quality and diversity of the weak datasets. Mor
eover, the self-bootstrapping of the strong model leads to performance degradati

on due to its overconfidence and the limited diversity of its generated dataset. To address these issues, we proposed a novel training framework using reward mo dels to select valuable data, thereby providing weak supervision for strong mode l training. In addition, we propose a two-stage training method on both weak and selected datasets to train the strong model. Experimental results demonstrate o ur method significantly improves the weak to strong generalization with multi-ca pabilities.
****
Ashish J Khisti,MohammadReza Ebrahimi,Hassan Dbouk,Arash Behboodi,Roland Memisev ic,Christos Louizos
Multi-Draft Speculative Sampling: Canonical Decomposition and Theoretical Limits
We consider multi-draft speculative sampling, where the proposal sequences are s ampled independently from different draft models. At each step, a token-level draft selection scheme takes a list of valid tokens as input and produces an out put token whose distribution matches that of the target model. Previous works ha ve demonstrated that the optimal scheme (which maximizes the probability of acce pting one of the input tokens) can be cast as a solution to a linear program. In this work we show that the optimal scheme can be decomposed into a two-step sol ution: in the first step an importance sampling (IS) type scheme is used to sele ct one intermediate token; in the second step (single-draft) speculative samplin g is applied to generate the output token. For the case of two identical draft models we further 1) establish a necessary and sufficient condition on the distr ibutions of the target and draft models for the acceptance probability to equal one and 2) provide an explicit expression for the optimal acceptance probability . Our theoretical analysis also motives a new class of token-level selection sc hemes based on weighted importance sampling. Our experimental results demonstrat e consistent improvements in the achievable block efficiency and token rates ove r baseline schemes in a number of scenarios.
****
Shahriar Rifat,Jonathan Ashdown,Michael J. De Lucia,Ananthram Swami,Francesco Re stuccia
On the Adversarial Vulnerability of Label-Free Test-Time Adaptation
Despite the success of Test-time adaptation (TTA), recent work has shown that ad ding relatively small adversarial perturbations to a limited number of samples l eads to significant performance degradation. Therefore, it is crucial to rigorou sly evaluate existing TTA algorithms against relevant threats and implement appr opriate security countermeasures. Importantly, existing threat models assume tes t-time samples will be labeled, which is impractical in real-world scenarios. To address this gap, we propose a new attack algorithm that does not rely on access to labeled test samples, thus providing a concrete way to assess the secu rity vulnerabilities of TTA algorithms. Our attack design is grounded in theoret ical foundations and can generate strong attacks against different state of the art TTA methods. In addition, we show that existing defense mechanisms are almos t ineffective, which emphasizes the need for further research on TTA security. T hrough extensive experiments on CIFAR10-C, CIFAR100-C, and ImageNet-C, we demons trate that our proposed approach closely matches the performance of state-of-the -art attack benchmarks, even without access to labeled samples. In certain cases , our approach generates stronger attacks, e.g., more than 4% higher error rate on CIFAR10-C.
****
Benjamin Feuer,Micah Goldblum,Teresa Datta,Sanjana Nambiar,Raz Besaleli,Samuel D ooley,Max Cembalest,John P Dickerson
Style Outweighs Substance: Failure Modes of LLM Judges in Alignment Benchmarking
The release of ChatGPT in November 2022 sparked an explosion of interest in post -training and an avalanche of new preference optimization (PO) methods. These me thods claim superior alignment by virtue of better correspondence with human pai rwise preferences, often measured by LLM-judges. In this work, we attempt to ans wer the following question -- do LLM-judge preferences translate to progress on other, more concrete metrics for alignment, and if not, why not? We define a con crete metric for alignment, and introduce SOS-Bench (Substance Outweighs Style B

enchmark), the largest standardized, reproducible LLM meta-benchmark to date. We find that (1) LLM-judge preferences do not correlate with concrete measures of safety, world knowledge, and instruction following; (2) LLM-judges have powerful implicit biases, prioritizing style over factuality and safety; and (3) the supervised fine-tuning (SFT) stage of post-training has a large impact on alignment, with data scaling and prompt diversity as the driving factors.
****
Adarsh Kappiyath,Abhra Chaudhuri,AJAY KUMAR JAISWAL,Ziquan Liu,Yunpeng Li,Xiatian Zhu,Lu Yin
SEBRA : Debiasing through Self-Guided Bias Ranking
Ranking samples by fine-grained estimates of spuriosity (the degree to which spurious cues are present) has recently been shown to significantly benefit bias mitigation, over the traditional binary biased-vs-unbiased partitioning of train sets. However, this spuriousity ranking comes with the requirement of human supervision. In this paper, we propose a debiasing framework based on our novel Self-Guided Bias Ranking (Sebra), that mitigates biases via an automatic ranking of data points by spuriosity within their respective classes. Sebra leverages a key local symmetry in Empirical Risk Minimization (ERM) training -- the ease of learning a sample via ERM inversely correlates with its spuriousity; the fewer spurious correlations a sample exhibits, the harder it is to learn, and vice versa. However, globally across iterations, ERM tends to deviate from this symmetry. Sebra dynamically steers ERM to correct this deviation, facilitating the sequential learning of attributes in increasing order of difficulty, ie, decreasing order of spuriosity. As a result, the sequence in which Sebra learns samples naturally provides spuriousity rankings. We use the resulting fine-grained bias characterization in a contrastive learning framework to mitigate biases from multiple sources. Extensive experiments show that Sebra consistently outperforms previous state-of-the-art unsupervised debiasing techniques across multiple standard benchmarks, including UrbanCars, BAR, and CelebA.
****
Renjie Pan,Jihao Dong,Hua Yang
Discovering Clone Negatives via Adaptive Contrastive Learning for Image-Text Matching
In this paper, we identify a common yet challenging issue in image-text matching, i.e., clone negatives: negative image-text pairs that semantically resemble positive pairs, leading to ambiguous and sub-optimal matching outcomes. To tackle this issue, we propose Adaptive Contrastive Learning (AdaCL), which introduces two margin parameters along with a modulating anchor to dynamically strengthen the compactness between positives and mitigate the influence of clone negatives. The modulating anchor is selected based on the distribution of negative samples without the need for explicit training, allowing for progressive tuning and advanced in-batch supervision. Extensive experiments across several tasks demonstrate the effectiveness of AdaCL in image-text matching. Furthermore, we extend AdaCL to weakly-supervised image-text matching by replacing human-annotated descriptions with automatically generated captions, thereby increasing the number of potential clone negatives. AdaCL maintains robustness in this setting, alleviating the reliance on crowd-sourced annotations and laying a foundation for scalable vision-language contrastive learning.
****
Jaden Fried Fiotto-Kaufman,Alexander Russell Loftus,Eric Todd,Jannik Brinkmann,Koyena Pal,Dmitrii Troitskii,Michael Ripa,Adam Belfki,Can Rager,Caden Juang,Aaron Mueller,Samuel Marks,Arnab Sen Sharma,Francesca Lucchetti,Nikhil Prakash,Carla E. Brodley,Arjun Guha,Jonathan Bell,Byron C Wallace,David Bau
NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals
We introduce NNsight and NDIF, technologies that work in tandem to enable scientific study of the representations and computations learned by very large neural networks. NNsight is an open-source system that extends PyTorch to introduce deferred remote execution. The National Deep Inference Fabric (NDIF) is a scalable inference service that executes NNsight requests, allowing users to share GPU resources and pretrained models. These technologies are enabled by the Interventio

n Graph, an architecture developed to decouple experimental design from model ru
ntime. Together, this framework provides transparent and efficient access to the
 internals of deep neural networks such as very large language models (LLMs) wit
hout imposing the cost or complexity of hosting customized models individually.
We conduct a quantitative survey of the machine learning literature that reveals
 a growing gap in the study of the internals of large-scale AI. We demonstrate t
he design and use of our framework to address this gap by enabling a range of re
search methods on huge models. Finally, we conduct benchmarks to compare perform
ance with previous approaches.

Code, documentation, and tutorials are available at https://nnsight.net/.
****
Aiqing Zhu,Yuting Pan,Qianxiao Li
Continuity-Preserving Convolutional Autoencoders for Learning Continuous Latent
 Dynamical Models from Images
Continuous dynamical systems are cornerstones of many scientific and engineering
 disciplines.
While machine learning offers powerful tools to model these systems from traject
ory data, challenges arise when these trajectories are captured as images, resul
ting in pixel-level observations that are discrete in nature.
Consequently, a naive application of a convolutional autoencoder can result in l
atent coordinates that are discontinuous in time.
To resolve this, we propose continuity-preserving convolutional autoencoders (Cp
AEs) to learn continuous latent states and their corresponding continuous latent
 dynamical models from discrete image frames.
We present a mathematical formulation for learning dynamics from image frames, w
hich illustrates issues with previous approaches and motivates our methodology b
ased on promoting the continuity of convolution filters, thereby preserving the
continuity of the latent states.
This approach enables CpAEs to produce latent states that evolve continuously wi
th the underlying dynamics, leading to more accurate latent dynamical models.
Extensive experiments across various scenarios demonstrate the effectiveness of
CpAEs.
****
Hongyu Ke,Jack Morris,Kentaro Oguchi,Xiaofei Cao,Yongkang Liu,Haoxin Wang,Yi Din
g
MamBEV: Enabling State Space Models to Learn Birds-Eye-View Representations
3D visual perception tasks, such as 3D detection from multi-camera images, are e
ssential components of autonomous driving and assistance systems. However, desig
ning computationally efficient methods remains a significant challenge. In this
paper, we propose a Mamba-based framework called MamBEV, which learns unified Bi
rd's Eye View (BEV) representations using linear spatio-temporal SSM-based atten
tion. This approach supports multiple 3D perception tasks with significantly imp
roved computational and memory efficiency. Furthermore, we introduce SSM based c
ross-attention, analogous to standard cross attention, where BEV query represent
ations can interact with relevant image features. Extensive experiments demonstr
ate MamBEV's promising performance across diverse visual perception metrics, hig
hlighting its advantages in input scaling efficiency compared to existing benchm
ark models.
****
Jun-Yan He,Zhi-Qi Cheng,Chenyang Li,Jingdong Sun,Qi He,Wangmeng Xiang,Hanyuan Ch
en,Jin-Peng Lan,Xianhui Lin,kang zhu,Bin Luo,Yifeng Geng,Xuansong Xie,Alexander
G Hauptmann
MetaDesigner: Advancing Artistic Typography through AI-Driven, User-Centric, and
 Multilingual WordArt Synthesis
MetaDesigner introduces a transformative framework for artistic typography synth
esis, powered by Large Language Models (LLMs) and grounded in a user-centric des
ign paradigm. Its foundation is a multi-agent system comprising the Pipeline, Gl
yph, and Texture agents, which collectively orchestrate the creation of customiz
able WordArt, ranging from semantic enhancements to intricate textural elements.

A central feedback mechanism leverages insights from both multimodal models and user evaluations, enabling iterative refinement of design parameters. Through this iterative process, MetaDesigner dynamically adjusts hyperparameters to align with user-defined stylistic and thematic preferences, consistently delivering WordArt that excels in visual quality and contextual resonance. Empirical evaluations underscore the system's versatility and effectiveness across diverse WordArt applications, yielding outputs that are both aesthetically compelling and context-sensitive.

****

REN Lianhai,Qianxiao Li

BP-Modified Local Loss for Efficient Training of Deep Neural Networks

The training of large models is memory-constrained, one direction to relieve this is training using local loss, like GIM, LoCo, and Forward-Forward algorithms. However, the local loss methods often face the issue of slow or non-convergence. In this paper, we propose a novel BP-modified local loss method that uses the true Backward Propagation (BP) gradient to modify the local loss gradient to improve the performance of local loss training. We use the stochastic modified equation to analyze our method and show that modified offset decreases the bias between the BP gradient and local loss gradient, but introduces additional variance, which results in a bias-variance balance. Numerical experiments on full-tuning and LoKr tuning on the ResNet-50 model and LoRA tuning on the ViT-b16 model on CIFAR-100 datasets show 20.5\% test top-1 accuracy improvement for the Forward-Forward algorithm, 18.6\% improvement for LoCo algorithm and achieve only on average 7.7\% of test accuracy loss compared to the BP algorithm, with up to 75\% memory savings.

****

Junyu Zhang,Daochang Liu,Eunbyung Park,Shichao Zhang,Chang Xu

Anti-Exposure Bias in Diffusion Models

Diffusion models (DMs) have achieved record-breaking performance in image generation tasks.

Nevertheless, in practice, the training-sampling discrepancy, caused by score estimation error and discretization error, limits the modeling ability of DMs, a phenomenon known as exposure bias.

To alleviate such exposure bias and further improve the generative performance, we put forward a prompt learning framework built upon a lightweight prompt prediction model.

Concretely, our model learns an anti-bias prompt for the generated sample at each sampling step, aiming to compensate for the exposure bias that arises.

Following this design philosophy, our framework rectifies the sampling trajectory to match the training trajectory, thereby reducing the divergence between the target data distribution and the modeling distribution.

To train the prompt prediction model, we simulate exposure bias by constructing training data and introduce a time-dependent weighting function for optimization.

Empirical results on various DMs demonstrate the superiority of our prompt learning framework across three benchmark datasets.

Importantly, the optimized prompt prediction model effectively improves image quality with only a 5\% increase in sampling overhead, which remains negligible.

****

M Saiful Bari,Yazeed Alnumay,Norah A. Alzahrani,Nouf M. Alotaibi,Hisham Abdullah Alyahya,Sultan AlRashed,Faisal Abdulrahman Mirza,Shaykhah Z. Alsubaie,Hassan A. Alahmed,Ghadah Alabduljabbar,Raghad Alkhathran,Yousef Almushayqih,Raneem Alnajim,Salman Alsubaihi,Maryam Al Mansour,Saad Amin Hassan,Dr. Majed Alrubaian,Ali Alammari,Zaki Alawami,Abdulmohsen Al-Thubaity,Ahmed Abdelali,Jeril Kuriakose,Abdalghani Abujabal,Nora Al-Twairesh,Areeb Alowisheq,Haidar Khan

ALLaM: Large Language Models for Arabic and English

In this work, we present ALLaM: Arabic Large Language Model, a series of large language models to support the ecosystem of Arabic Language Technologies (ALT). ALLaM is carefully trained, considering the values of language alignment and tran

sferability of knowledge at scale. The models are based on an autoregressive dec
oder-only architecture and are pretrained on a mixture of Arabic and English tex
ts. We illustrate how the second-language acquisition via vocabulary expansion c
an help steer a language model towards a new language without any major catastro
phic forgetting in English. Furthermore, we highlight the effectiveness of using
 translation data and the process of knowledge encoding within the language mode
l's latent space. Finally, we show that effective alignment with human preferenc
es can significantly enhance the performance of a large language model (LLM) com
pared to less aligned models of a larger scale. Our methodology enables us to ac
hieve state-of-the-art performance in various Arabic benchmarks, including MMLU
Arabic, ACVA, and Arabic Exams. Our aligned models improve both in Arabic and En
glish from its base aligned models.
****
Etrit Haxholli,Yeti Z. Gurbuz,O■ul Can,Eli Waxman
Efficient Perplexity Bound and Ratio Matching in Discrete Diffusion Language Mod
els
While continuous diffusion models excel in modeling continuous distributions, th
eir application to categorical data has been less effective. Recent work has sho
wn that ratio-matching through *score-entropy* within a continuous-time discrete
 Markov chain (CTMC) framework serves as a competitive alternative to autoregres
sive models in language modeling.
To enhance this framework, we first introduce three new theorems concerning the
KL divergence between the data and learned distribution. Our results serve as th
e discrete counterpart to those established for continuous diffusion models and
allow us to derive an improved upper bound of the perplexity. Second, we empiric
ally show that ratio-matching performed by minimizing the *denoising cross-entro
py* between the clean and corrupted data enables models to outperform those util
izing score-entropy with up to 10\% lower perplexity/generative-perplexity, and
15\% faster training steps.
 To further support our findings, we introduce and evaluate a novel CTMC transit
ion-rate matrix that allows prediction refinement, and derive the analytic expre
ssion for its matrix exponential which facilitates the computation of conditiona
l ratios thus enabling efficient training and generation.
****
Zhilin Wang,Alexander Bukharin,Olivier Delalleau,Daniel Egert,Gerald Shen,Jiaqi
Zeng,Oleksii Kuchaiev,Yi Dong
HelpSteer2-Preference: Complementing Ratings with Preferences
Reward models are critical for aligning models to follow instructions, and are t
ypically trained following one of two popular paradigms: Bradley-Terry style or
Regression style. However, there is a lack of evidence that either approach is b
etter than the other, when adequately matched for data. This is primarily becaus
e these approaches require data collected in different (but incompatible) format
s, meaning that adequately matched data is not available in existing public data
sets. To tackle this problem, we release preference annotations (designed for Br
adley-Terry training) to complement existing ratings (designed for Regression st
yle training) in the HelpSteer2 dataset. To improve data interpretability, prefe
rence annotations are accompanied with human-written justifications. Using this
data, we conduct the first head-to-head comparison of Bradley-Terry and Regressi
on models when adequately matched for data. Based on insights derived from such
a comparison, we propose a novel approach to combine Bradley-Terry and Regressio
n reward modeling. A Llama-3.1-70B-Instruct model tuned with this approach score
s 94.1 on RewardBench, emerging top of more than 140 reward models as of 1 Oct 2
024. This reward model can then be used with REINFORCE algorithm (RLHF) to align
 an Instruct model to reach 85.0 on Arena Hard, which is No. 1 as of 1 Oct 2024.
We open-source this dataset (CC-BY-4.0 license) at https://huggingface.co/datase
ts/nvidia/HelpSteer2#preferences-new---1-oct-2024 and openly release the trained
 Reward and Instruct models at https://huggingface.co/nvidia/Llama-3.1-Nemotron-
70B-Reward and https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct .
****
Renrui Zhang,Xinyu Wei,Dongzhi Jiang,Ziyu Guo,Yichi Zhang,Chengzhuo Tong,Jiaming

Liu,Aojun Zhou,Shanghang Zhang,Peng Gao,Hongsheng Li
MAVIS: Mathematical Visual Instruction Tuning with an Automatic Data Engine
Multi-modal Large Language Models (MLLMs) have recently showcased superior profi ciency in general visual scenarios. However, we identify their mathematical capa bilities remain under-explored with three areas to be improved: visual encoding of math diagrams, diagram-language alignment, and chain-of-thought (CoT) reasoni ng. This draws forth an urgent demand for an effective training paradigm and a l arge-scale, comprehensive dataset with detailed CoT rationales, which is challen ging to collect and costly to annotate manually. To tackle this issue, we propos e MAVIS, a MAthematical VISual instruction tuning pipeline for MLLMs, featuring an automatic data engine to efficiently create mathematical visual datasets.
We design the data generation process to be entirely independent of human interv ention or GPT API usage, while ensuring the diagram-caption correspondence, ques tion-answer correctness, and CoT reasoning quality. With this approach, we curat e two datasets, MAVIS-Caption (558K diagram-caption pairs) and MAVIS-Instruct (8 34K visual math problems with CoT rationales), and propose four progressive stag es for training MLLMs from scratch.
First, we utilize MAVIS-Caption to fine-tune a math-specific vision encoder (CLI P-Math) through contrastive learning, tailored for improved diagram visual encod ing. Second, we also leverage MAVIS-Caption to align the CLIP-Math with a large language model (LLM) by a projection layer, enhancing vision-language alignment in mathematical domains. Third, we adopt MAVIS-Instruct to perform the instructi on tuning for robust problem-solving skills, and term the resulting model as MAV IS-7B. Fourth, we apply Direct Preference Optimization (DPO) to enhance the CoT capabilities of our model, further refining its step-wise reasoning performance.
On various mathematical benchmarks, our MAVIS-7B achieves leading results among open-source MLLMs, e.g., surpassing other 7B models by +9.3% and the second-best LLaVA-NeXT (110B) by +6.9%, demonstrating the effectiveness of our method.
****
Donggoo Jung,Daehyun Kim,Tae Hyun Kim
Continuous Exposure Learning for Low-light Image Enhancement using Neural ODEs
Low-light image enhancement poses a significant challenge due to the limited inf ormation captured by image sensors in low-light environments.
Despite recent improvements in deep learning models, the lack of paired traini ng datasets remains a significant obstacle.
Therefore, unsupervised methods have emerged as a promising solution.
In this work, we focus on the strength of curve-adjustment-based approaches to tackle unsupervised methods.
The majority of existing unsupervised curve-adjustment approaches iteratively estimate higher order curve parameters to enhance the exposure of images while e fficiently preserving the details of the images.
However, the convergence of the enhancement procedure cannot be guaranteed, le ading to sensitivity to the number of iterations and limited performance.
To address this problem, we consider the iterative curve-adjustment update pro cess as a dynamic system and formulate it as a Neural Ordinary Differential Equa tions (NODE) for the first time, and this allows us to learn a continuous dynami cs of the latent image.
The strategy of utilizing NODE to leverage continuous dynamics in iterative me thods enhances unsupervised learning and aids in achieving better convergence co mpared to discrete-space approaches. Consequently, we achieve state-of-the-art p erformance in unsupervised low-light image enhancement across various benchmark datasets.
****
Zhuorui Ye,Stephanie Milani,Geoffrey J. Gordon,Fei Fang
LICORICE: Label-Efficient Concept-Based Interpretable Reinforcement Learning
Recent advances in reinforcement learning (RL) have predominantly leveraged neur al network policies for decision-making, yet these models often lack interpretab ility, posing challenges for stakeholder comprehension and trust. Concept bottle neck models offer an interpretable alternative by integrating human-understandab le concepts into policies. However, prior work assumes that concept annotations

are readily available during training. For RL, this requirement poses a signific
ant limitation: it necessitates continuous real-time concept annotation, which e
ither places an impractical burden on human annotators or incurs substantial cos
ts in API queries and inference time when employing automated labeling methods.
To overcome this limitation, we introduce a novel training scheme that enables R
L agents to efficiently learn a concept-based policy by only querying annotators
 to label a small set of data. Our algorithm, LICORICE, involves three main cont
ributions: interleaving concept learning and RL training, using an ensemble to a
ctively select informative data points for labeling, and decorrelating the conce
pt data. We show how LICORICE reduces human labeling efforts to 500 or fewer con
cept labels in three environments, and 5000 or fewer in two more complex environ
ments, all at no cost to performance. We also explore the use of VLMs as automat
ed concept annotators, finding them effective in some cases but imperfect in oth
ers. Our work significantly reduces the annotation burden for interpretable RL,
making it more practical for real-world applications that necessitate transparen
cy. Our code is released.
****

Wenjie Wei,Malu Zhang,Zijian Zhou,Ammar Belatreche,Yimeng Shan,Yu Liang,Honglin
Cao,Jieyuan Zhang,Yang Yang
QP-SNN: Quantized and Pruned Spiking Neural Networks
Brain-inspired Spiking Neural Networks (SNNs) leverage sparse spikes to encode i
nformation and operate in an asynchronous event-driven manner, offering a highly
 energy-efficient paradigm for machine intelligence. However, the current SNN co
mmunity focuses primarily on performance improvement by developing large-scale m
odels, which limits the applicability of SNNs in resource-limited edge devices.
In this paper, we propose a hardware-friendly and lightweight SNN, aimed at effe
ctively deploying high-performance SNN in resource-limited scenarios. Specifical
ly, we first develop a baseline model that integrates uniform quantization and s
tructured pruning, called QP-SNN baseline. While this baseline significantly red
uces storage demands and computational costs, it suffers from performance declin
e. To address this, we conduct an in-depth analysis of the challenges in quantiz
ation and pruning that lead to performance degradation and propose solutions to
enhance the baseline's performance. For weight quantization, we propose a weight
 rescaling strategy that utilizes bit width more effectively to enhance the mode
l's representation capability. For structured pruning, we propose a novel prunin
g criterion using the singular value of spatiotemporal spike activities to enabl
e more accurate removal of redundant kernels. Extensive experiments demonstrate
that integrating two proposed methods into the baseline allows QP-SNN to achieve
 state-of-the-art performance and efficiency, underscoring its potential for enh
ancing SNN deployment in edge intelligence computing.
****

Xingyu Su,Haiyang Yu,Degui Zhi,Shuiwang Ji
Learning to Discover Regulatory Elements for Gene Expression Prediction
We consider the problem of predicting gene expressions from DNA sequences. A key
 challenge of this task is to find the regulatory elements that control gene exp
ressions. Here, we introduce Seq2Exp, a Sequence to Expression network explicitl
y designed to discover and extract regulatory elements that drive target gene ex
pression, enhancing the accuracy of the gene expression prediction. Our approach
 captures the causal relationship between epigenomic signals, DNA sequences and
their associated regulatory elements. Specifically, we propose to decompose the
epigenomic signals and the DNA sequence conditioned on the causal active regulat
ory elements, and apply an information bottleneck with the Beta distribution to
combine their effects while filtering out non-causal components. Our experiments
 demonstrate that Seq2Exp outperforms existing baselines in gene expression pred
iction tasks and discovers influential regions compared to commonly used statist
ical methods for peak detection such as MACS3. The source code is released as pa
rt of the AIRS library (https://github.com/divelab/AIRS/).
****

Elizaveta Tennant,Stephen Hailes,Mirco Musolesi
Moral Alignment for LLM Agents

Decision-making agents based on pre-trained Large Language Models (LLMs) are inc
reasingly being deployed across various domains of human activity. While their a
pplications are currently rather specialized, several research efforts are under
way to develop more generalist agents. As LLM-based systems become more agentic,
 their influence on human activity will grow and their transparency will decreas
e. Consequently, developing effective methods for aligning them to human values
is vital.

The prevailing practice in alignment often relies on human preference data (e.g.
, in RLHF or DPO), in which values are implicit, opaque and are essentially dedu
ced from relative preferences over different model outputs. In this work, instea
d of relying on human feedback, we introduce the design of reward functions that
 explicitly and transparently encode core human values for Reinforcement Learnin
g-based fine-tuning of foundation agent models. Specifically, we use intrinsic r
ewards for the moral alignment of LLM agents.

We evaluate our approach using the traditional philosophical frameworks of Deont
ological Ethics and Utilitarianism, quantifying moral rewards for agents in term
s of actions and consequences on the Iterated Prisoner's Dilemma (IPD) environme
nt. We also show how moral fine-tuning can be deployed to enable an agent to unl
earn a previously developed selfish strategy. Finally, we find that certain mora
l strategies learned on the IPD game generalize to several other matrix game env
ironments. In summary, we demonstrate that fine-tuning with intrinsic rewards is
 a promising general solution for aligning LLM agents to human values, and it mi
ght represent a more transparent and cost-effective alternative to currently pre
dominant alignment techniques.
****
William Lingxiao Tong,Cengiz Pehlevan
MLPs Learn In-Context on Regression and Classification Tasks
In-context learning (ICL), the remarkable ability to solve a task from only inpu
t exemplars, is often assumed to be a unique hallmark of Transformer models. By
examining commonly employed synthetic ICL tasks, we demonstrate that multi-layer
 perceptrons (MLPs) can also learn in-context. Moreover, MLPs, and the closely r
elated MLP-Mixer models, learn in-context comparably with Transformers under the
 same compute budget in this setting. We further show that MLPs outperform Trans
formers on a series of classical tasks from psychology designed to test relation
al reasoning, which are closely related to in-context classification. These resu
lts underscore a need for studying in-context learning beyond attention-based ar
chitectures, while also challenging prior arguments against MLPs' ability to sol
ve relational tasks. Altogether, our results highlight the unexpected competence
 of MLPs in a synthetic setting, and support the growing interest in all-MLP alt
ernatives to Transformer architectures. It remains unclear how MLPs perform agai
nst Transformers at scale on real-world tasks, and where a performance gap may o
riginate. We encourage further exploration of these architectures in more comple
x settings to better understand the potential comparative advantage of attention
-based schemes.
****
Hengyu Fu,Zehao Dou,Jiawei Guo,Mengdi Wang,Minshuo Chen
Diffusion Transformer Captures Spatial-Temporal Dependencies: A Theory for Gauss
ian Process Data
Diffusion Transformer, the backbone of Sora for video generation, successfully s
cales the capacity of diffusion models, pioneering new avenues for high-fidelity
 sequential data generation. Unlike static data such as images, sequential data
consists of consecutive data frames indexed by time, exhibiting rich spatial and
 temporal dependencies. These dependencies represent the underlying dynamic mode
l and are critical to validate the generated data. In this paper, we make the fi
rst theoretical step towards bridging diffusion transformers for capturing spati
al-temporal dependencies. Specifically, we establish score approximation and dis
tribution estimation guarantees of diffusion transformers for learning Gaussian
process data with covariance functions of various decay patterns. We highlight h

ow the spatial-temporal dependencies are captured and affect learning efficiency . Our study proposes a novel transformer approximation theory, where the transfo rmer acts to unroll an algorithm. We support our theoretical results by numerica l experiments, providing strong evidence that spatial-temporal dependencies are captured within attention layers, aligning with our approximation theory.
****

dewei Zhou,Ji Xie,Zongxin Yang,Yi Yang
3DIS: Depth-Driven Decoupled Image Synthesis for Universal Multi-Instance Genera tion
The increasing demand for controllable outputs in text-to-image generation has s purred advancements in multi-instance generation (MIG), allowing users to define both instance layouts and attributes. However, unlike image-conditional generat ion methods such as ControlNet, MIG techniques have not been widely adopted in s tate-of-the-art models like SD2 and SDXL, primarily due to the challenge of buil ding robust renderers that simultaneously handle instance positioning and attrib ute rendering. In this paper, we introduce Depth-Driven Decoupled Image Synthesi s (3DIS), a novel framework that decouples the MIG process into two stages: (i) generating a coarse scene depth map for accurate instance positioning and scene composition, and (ii) rendering fine-grained attributes using pre-trained Contro lNet on any foundational model, without additional training. Our 3DIS framework integrates a custom adapter into LDM3D for precise depth-based layouts and emplo ys a finetuning-free method for enhanced instance-level attribute rendering. Ext ensive experiments on COCO-Position and COCO-MIG benchmarks demonstrate that 3DI S significantly outperforms existing methods in both layout precision and attrib ute rendering. Notably, 3DIS offers seamless compatibility with diverse foundati onal models, providing a robust, adaptable solution for advanced multi-instance generation. The code is available at: https://github.com/limuloo/3DIS.
****

Chenghan Li,Mingchen Li,Ruisheng Diao
TVNet: A Novel Time Series Analysis Method Based on Dynamic Convolution and 3D-V ariation
With the recent development and advancement of Transformer and MLP architectures , significant strides have been made in time series analysis. Conversely, the pe rformance of Convolutional Neural Networks (CNNs) in time series analysis has fa llen short of expectations, diminishing their potential for future applications. Our research aims to enhance the representational capacity of Convolutional Neu ral Networks (CNNs) in time series analysis by introducing novel perspectives an d design innovations. To be specific, We introduce a novel time series reshaping technique that considers the inter-patch, intra-patch, and cross-variable dimen sions. Consequently, we propose TVNet, a dynamic convolutional network leveragin g a 3D perspective to employ time series analysis. TVNet retains the computation al efficiency of CNNs and achieves state-of-the-art results in five key time ser ies analysis tasks, offering a superior balance of efficiency and performance ov er the state-of-the-art Transformer-based and MLP-based models. Additionally, ou r findings suggest that TVNet exhibits enhanced transferability and robustness. Therefore, it provides a new perspective for applying CNN in advanced time serie s analysis tasks.
****

Sepanta Zeighami,Zac Wellmer,Aditya Parameswaran
NUDGE: Lightweight Non-Parametric Fine-Tuning of Embeddings for Retrieval
k-Nearest Neighbor search on dense vector embeddings (k-NN retrieval) from p re-trained embedding models is the predominant retrieval method for text and ima ges, as well as Retrieval-Augmented Generation (RAG) pipelines. In practice, app lication developers often fine-tune the embeddings to improve their accuracy on the dataset and query workload in hand. Existing approaches either fine-tune the pre-trained model itself or, more efficiently, but at the cost of accuracy, tra in adaptor models to transform the output of the pre-trained model. We present N UDGE, a family of novel *non-parametric* embedding fine-tuning approaches that a re significantly more accurate and efficient than both sets of existing approach es. NUDGE directly modifies the embeddings of data records to maximize the accur

acy of k-NN retrieval. We present a thorough theoretical and experimental study of NUDGE's non-parametric approach. We show that even though the underlying problem is NP-Hard, constrained variations can be solved efficiently. These constraints additionally ensure that the changes to the embeddings are modest, avoiding large distortions to the semantics learned during pre-training. In experiments across five pre-trained models and nine standard text and image retrieval datasets, *NUDGE runs in minutes and often improves NDCG@10 by more than 10\% over existing fine-tuning methods. On average, NUDGE provides 3.3$\times$ and 4.3$\times$ higher increase in accuracy and runs 200$\times$ and 3$\times$ faster, respectively, over fine-tuning the pre-trained model and training adaptors.*
****

Yunyue Wei,Shanning Zhuang,Vincent Zhuang,Yanan Sui
Motion Control of High-Dimensional Musculoskeletal Systems with Hierarchical Model-Based Planning
Controlling high-dimensional nonlinear systems, such as those found in biological and robotic applications, is challenging due to large state and action spaces. While deep reinforcement learning has achieved a number of successes in these domains, it is computationally intensive and time consuming, and therefore not suitable for solving large collections of tasks that require significant manual tuning. In this work, we introduce Model Predictive Control with Morphology-aware Proportional Control (MPC^2), a hierarchical model-based learning algorithm for zero-shot and near-real-time control of high-dimensional complex dynamical systems. MPC^2 uses a sampling-based model predictive controller for target posture planning, and enables robust control for high-dimensional tasks by incorporating a morphology-aware proportional controller for actuator coordination. The algorithm enables motion control of a high-dimensional human musculoskeletal model in a variety of motion tasks, such as standing, walking on different terrains, and imitating sports activities. The reward function of MPC^2 can be tuned via black-box optimization, drastically reducing the need for human-intensive reward engineering.
****

Boning Li,Longbo Huang
Efficient Online Pruning and Abstraction for Imperfect Information Extensive-Form Games
Efficiently computing approximate equilibrium strategies in large Imperfect Information Extensive-Form Games (IIEFGs) poses significant challenges due to the game tree's exponential growth. While pruning and abstraction techniques are essential for complexity reduction, existing methods face two key limitations: (i) Seamless integration of pruning with Counterfactual Regret Minimization (CFR) is nontrivial, and (ii) Pruning and abstraction approaches incur prohibitive computational costs, hindering real-world deployment. We propose Expected-Value Pruning and Abstraction (EVPA), a novel online framework that addresses these challenges through three synergistic components: (i) Expected value estimation using approximate Nash equilibrium strategies to quantify information set utilities, (ii) Minimax pruning before CFR to eliminate a large number of sub-optimal actions permanently, and (iii) Dynamic online information abstraction merging information sets based on their current and future expected values in subgames. Experiments on Heads-up No-Limit Texas Hold'em (HUNL) show EVPA outperforms DeepStack's replication and Slumbot with significant win-rate margins in multiple settings. Remarkably, EVPA requires only 1\%-2\% of the solving time to reach an approximate Nash equilibrium compared to DeepStack's replication.
****

Shivam Gupta,Linda Cai,Sitan Chen
Faster Diffusion Sampling with Randomized Midpoints: Sequential and Parallel
Sampling algorithms play an important role in controlling the quality and runtime of diffusion model inference. In recent years, a number of works (Chen et al., 2023c;b; Benton et al., 2023; Lee et al., 2022) have analyzed algorithms for diffusion sampling with provable guarantees; these works show that for essentially any data distribution, one can approximately sample in polynomial time given a sufficiently accurate estimate of its score functions at different noise levels.

In this work, we propose a new scheme inspired by Shen and Lee's randomized midpoint method for log-concave sampling (Shen & Lee, 2019). We prove that this approach achieves the best known dimension dependence for sampling from arbitrary smooth distributions in total variation distance ($\widetilde O(d^{5/12})$ compared to $\widetilde O(\sqrt{d})$ from prior work). We also show that our algorithm can be parallelized to run in only $\widetilde O(\log^2 d)$ parallel rounds, constituting the first provable guarantees for parallel sampling with diffusion models.

As a byproduct of our methods, for the well-studied problem of log-concave sampling in total variation distance, we give an algorithm and simple analysis achieving dimension dependence $\widetilde O(d^{5/12})$ compared to $\widetilde O(\sqrt{d})$ from prior work.
****
Woosung Koh,Wonbeen Oh,Siyeol Kim,Suhin Shin,Hyeongjin Kim,Jaein Jang,Junghyun Lee,Se-Young Yun
FlickerFusion: Intra-trajectory Domain Generalizing Multi-agent Reinforcement Learning
Multi-agent reinforcement learning has demonstrated significant potential in addressing complex cooperative tasks across various real-world applications. However, existing MARL approaches often rely on the restrictive assumption that the number of entities (e.g., agents, obstacles) remains constant between training and inference. This overlooks scenarios where entities are dynamically removed or $\textit{added}$ $\textit{during}$ the inference trajectory—a common occurrence in real-world environments like search and rescue missions and dynamic combat situations. In this paper, we tackle the challenge of intra-trajectory dynamic entity composition under zero-shot out-of-domain (OOD) generalization, where such dynamic changes cannot be anticipated beforehand. Our empirical studies reveal that existing MARL methods suffer $\textit{significant}$ performance degradation and increased uncertainty in these scenarios. In response, we propose FlickerFusion, a novel OOD generalization method that acts as a $\textit{universally}$ applicable augmentation technique for MARL backbone methods. FlickerFusion stochastically drops out parts of the observation space, emulating being in-domain when inferenced OOD. The results show that FlickerFusion not only achieves superior inference rewards but also $\textit{uniquely}$ reduces uncertainty vis-à-vis the backbone, compared to existing methods. Benchmarks, implementations, and model weights are organized and open-sourced at $\texttt{\href{flickerfusion305.github.io}{\textbf{flickerfusion305.github.io}}}$, accompanied by ample demo video renderings.
****
Yunzhen Feng,Elvis Dohmatob,Pu Yang,Francois Charton,Julia Kempe
Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification
Large Language Models (LLM) are increasingly trained on data generated by other LLMs, either because generated text and images become part of the pre-training corpus, or because synthetized data is used as a replacement for expensive human-annotation. This raises concerns about *model collapse*, a drop in model performance when their training sets include generated data. Considering that it is easier for both humans and machines to tell between good and bad examples than to generate high-quality samples, we investigate the use of verification on synthesized data to prevent model collapse. We provide a theoretical characterization using Gaussian mixtures, linear classifiers, and linear verifiers to derive conditions with measurable proxies to assess whether the verifier can effectively select synthesized data that leads to optimal performance. We experiment with two practical tasks -- computing matrix eigenvalues with transformers and news summarization with LLMs -- which both exhibit model collapse when trained on generated data, and show that verifiers, even imperfect ones, can indeed be harnessed to prevent model collapse and that our proposed proxy measure strongly correlates with performance.

****
Yuming Chen,Jiangyan Feng,Haodong Zhang,Lijun GONG,Feng Zhu,Rui Zhao,Qibin Hou,Ming-Ming Cheng,Yibing Song
Re-Aligning Language to Visual Objects with an Agentic Workflow
Language-based object detection (LOD) aims to align visual objects with language expressions. A large amount of paired data is utilized to improve LOD model generalizations. During the training process, recent studies leverage vision-language models (VLMs) to automatically generate human-like expressions for visual objects, facilitating training data scaling up. In this process, we observe that VLM hallucinations bring inaccurate object descriptions (e.g., object name, color, and shape) to deteriorate VL alignment quality. To reduce VLM hallucinations, we propose an agentic workflow controlled by an LLM to re-align language to visual objects via adaptively adjusting image and text prompts. We name this workflow Real-LOD, which includes planning, tool use, and reflection steps. Given an image with detected objects and VLM raw language expressions, Real-LOD reasons its state automatically and arranges action based on our neural symbolic designs (i.e., planning). The action will adaptively adjust the image and text prompts and send them to VLMs for object re-description (i.e., tool use). Then, we use another LLM to analyze these refined expressions for feedback (i.e., reflection). These steps are conducted in a cyclic form to gradually improve language descriptions for re-aligning to visual objects. We construct a dataset that contains a tiny amount of 0.18M images with re-aligned language expression and train a prevalent LOD model to surpass existing LOD methods by around 50% on the standard benchmarks. Our Real-LOD workflow, with automatic VL refinement, reveals a potential to preserve data quality along with scaling up data quantity, which further improves LOD performance from a data-alignment perspective.
****
Wenting Zhao,Nan Jiang,Celine Lee,Justin T Chiu,Claire Cardie,Matthias Gallé,Alexander M Rush
Commit0: Library Generation from Scratch
With the goal of benchmarking generative systems beyond expert software development ability, we introduce Commit0, a benchmark that challenges AI agents to write libraries from scratch. Agents are provided with a specification document outlining the library’s API as well as a suite of interactive unit tests, with the goal of producing an implementation of this API accordingly. The implementation is validated through running these unit tests. As a benchmark, Commit0 is designed to move beyond static one-shot code generation towards agents that must process long-form natural language specifications, adapt to multi-stage feedback, and generate code with complex dependencies. Commit0 also offers an interactive environment where models receive static analysis and execution feedback on the code they generate. Our experiments demonstrate that while current agents can pass some unit tests, none can yet fully reproduce full libraries. Results also show that interactive feedback is quite useful for models to generate code that passes more unit tests, validating the benchmarks that facilitate its use. We publicly release the benchmark, the interactive environment, and the leaderboard.
****
Mehdi Yazdani-Jahromi,Mangal Prakash,Tommaso Mansi,Artem Moskalev,Rui Liao
HELM: Hierarchical Encoding for mRNA Language Modeling
Messenger RNA (mRNA) plays a crucial role in protein synthesis, with its codon structure directly impacting biological properties. While Language Models (LMs) have shown promise in analyzing biological sequences, existing approaches fail to account for the hierarchical nature of mRNA's codon structure. We introduce Hierarchical Encoding for mRNA Language Modeling (HELM), a novel pre-training strategy that incorporates codon-level hierarchical structure into language model training. HELM modulates the loss function based on codon synonymity, aligning the model's learning process with the biological reality of mRNA sequences. We evaluate HELM on diverse mRNA datasets and tasks, demonstrating that HELM outperforms standard language model pre-training as well as existing foundation model baselines on six diverse downstream property prediction tasks and an antibody region annotation tasks on average by around 8%. Additionally, HELM enhances the genera

tive capabilities of language model, producing diverse mRNA sequences that bette r align with the underlying true data distribution compared to non-hierarchical baselines.
****
XIAOCHEN ZHAO,Hongyi Xu,Guoxian Song,You Xie,Chenxu Zhang,Xiu Li,Linjie Luo,Jinl i Suo,Yebin Liu
X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention
We propose X-NeMo, a novel zero-shot diffusion-based portrait animation pipeline that animates a static portrait using facial movements from a driving video of a different individual. Our work first identifies the root causes of the limitat ions in prior approaches, such as identity leakage and difficulty in capturing s ubtle and extreme expressions. To address these challenges, we introduce a fully end-to-end training framework that distills a 1D identity-agnostic latent motio n descriptor from driving image, effectively controlling motion through cross-at tention during image generation. Our implicit motion descriptor captures express ive facial motion in fine detail, learned end-to-end from a diverse video datase t without reliance on any pre-trained motion detectors. We further disentangle motion latents from identity cues with enhanced expressiveness by supervising th eir learning with a dual GAN decoder, alongside spatial and color augmentations. By embedding the driving motion into a 1D latent vector and controlling motion via cross-attention instead of additive spatial guidance, our design effectively eliminates the transmission of spatial-aligned structural clues from the drivin g condition to the diffusion backbone, substantially mitigating identity leakage . Extensive experiments demonstrate that X-NeMo surpasses state-of-the-art basel ines, producing highly expressive animations with superior identity resemblance. Our code and models will be available for research.
****
Bai LiChen,Shitong Shao,zikai zhou,Zipeng Qi,zhiqiang xu,Haoyi Xiong,Zeke Xie
Zigzag Diffusion Sampling: Diffusion Models Can Self-Improve via Self-Reflection
Diffusion models, the most popular generative paradigm so far, can inject condit ional information into the generation path to guide the latent towards desired d irections. However, existing text-to-image diffusion models often fail to mainta in high image quality and high prompt-image alignment for those challenging prom pts. To mitigate this issue and enhance existing pretrained diffusion models, we mainly made three contributions in this paper. First, we propose **diffusion se lf-reflection** that alternately performs denoising and inversion and demonstrat e that such diffusion self-reflection can leverage the guidance gap between deno ising and inversion to capture prompt-related semantic information with theoreti cal and empirical evidence. Second, motivated by theoretical analysis, we derive Zigzag Diffusion Sampling (Z-Sampling), a novel self-reflection-based diffusion sampling method that leverages the guidance gap between denosing and inversion to accumulate semantic information step by step along the sampling path, leading to improved sampling results. Moreover, as a plug-and-play method, Z-Sampling c an be generally applied to various diffusion models (e.g., accelerated ones and Transformer-based ones) with very limited coding and computational costs. Third, our extensive experiments demonstrate that Z-Sampling can generally and signifi cantly enhance generation quality across various benchmark datasets, diffusion m odels, and performance evaluation metrics. For example, DreamShaper with Z-Sampl ing can self-improve with the HPSv2 winning rate up to **94%** over the original results. Moreover, Z-Sampling can further enhance existing diffusion models com bined with other orthogonal methods, including Diffusion-DPO. The code is public ly available at
[github.com/xie-lab-ml/Zigzag-Diffusion-Sampling](https://github.com/xie-lab-ml/ Zigzag-Diffusion-Sampling).
****
Shuhao Cao,Francesco Brarda,Ruipeng Li,Yuanzhe Xi
Spectral-Refiner: Accurate Fine-Tuning of Spatiotemporal Fourier Neural Operator for Turbulent Flows
Recent advancements in operator-type neural networks have shown promising result s in approximating the solutions of spatiotemporal Partial Differential Equation

s (PDEs). However, these neural networks often entail considerable training expe
nses, and may not always achieve the desired accuracy required in many scientifi
c and engineering disciplines.
Recent advancements in operator-type neural networks have shown promising result
s in approximating the solutions of spatiotemporal Partial Differential Equation
s (PDEs). However, these neural networks often entail considerable training expe
nses, and may not always achieve the desired accuracy required in many scientifi
c and engineering disciplines.
In this paper, we propose a new learning framework to address these issues. A
new spatiotemporal adaptation is proposed to generalize any Fourier Neural Opera
tor (FNO) variant to learn maps between Bochner spaces, which can perform an arb
itrary-length temporal super-resolution for the first time.
To better exploit this capacity, a new paradigm is proposed to refine the comm
only adopted end-to-end neural operator training and evaluations with the help f
rom the wisdom from traditional numerical PDE theory and techniques.
Specifically, in the learning problems for the turbulent flow modeled by the N
avier-Stokes Equations (NSE), the proposed paradigm trains an FNO only for a few
 epochs. Then, only the newly proposed spatiotemporal spectral convolution layer
 is fine-tuned without the frequency truncation. The spectral fine-tuning loss f
unction uses a negative Sobolev norm for the first time in operator learning, de
fined through a reliable functional-type a posteriori error estimator whose eval
uation is exact thanks to the Parseval identity. Moreover, unlike the difficult
nonconvex optimization problems in the end-to-end training, this fine-tuning los
s is convex.
Numerical experiments on commonly used NSE benchmarks demonstrate significant
improvements in both computational efficiency and accuracy, compared to end-to-e
nd evaluation and traditional numerical PDE solvers under certain conditions. Th
e source code is publicly available at https://github.com/scaomath/torch-cfd.
****
Bill Yuchen Lin,Yuntian Deng,Khyathi Chandu,Abhilasha Ravichander,Valentina Pyat
kin,Nouha Dziri,Ronan Le Bras,Yejin Choi
WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild
We introduce WildBench, an automated evaluation framework designed to benchmark
large language models (LLMs) using challenging, real-world user queries. WildBen
ch consists of 1,024 tasks carefully selected from over one million human-chatbo
t conversation logs. For automated evaluation with WildBench, we have developed
two metrics, WB-Reward and WB-Score, which are computable using advanced LLMs su
ch as GPT-4-turbo. WildBench evaluation uses task-specific checklists to evaluat
e model outputs systematically and provides structured explanations that justify
 the scores and comparisons, resulting in more reliable and interpretable automa
tic judgments. WB-Reward employs fine-grained pairwise comparisons between model
 responses, generating five potential outcomes: much better, slightly better, sl
ightly worse, much worse, or a tie. Unlike previous evaluations that employed a
single baseline model, we selected three baseline models at varying performance
levels to ensure a comprehensive pairwise evaluation. Additionally, we propose a
 simple method to mitigate length bias, by converting outcomes of “slightly bett
er/worse” to “tie” if the winner response exceeds the loser one by more than K c
haracters. WB-Score evaluates the quality of model outputs individually, making
it a fast and cost-efficient evaluation metric. WildBench results demonstrate a
strong correlation with the human-voted Elo ratings from Chatbot Arena on hard t
asks. Specifically, WB-Reward achieves a Pearson correlation of 0.98 with top-ra
nking models. Additionally, WB-Score reaches 0.95, surpassing both ArenaHard’s 0
.91 and AlpacaEval2.0’s 0.89 for length-controlled win rates, as well as the 0.8
7 for regular win rates.
****
Junjie Wu,Qilong Wang,Jiangtao Xie,Pengfei Zhu,Qinghua Hu
Asymmetric Factorized Bilinear Operation for Vision Transformer
As a core component of Transformer-like deep architectures, a feed-forward netwo
rk (FFN) for channel mixing is responsible for learning features of each token.
Recent works show channel mixing can be enhanced by increasing computational bur

den or can be slimmed at the sacrifice of performance. Although some efforts hav
e been made, existing works are still struggling to solve the paradox of perform
ance and complexity trade-offs. In this paper, we propose an Asymmetric Factoriz
ed Bilinear Operation (AFBO) to replace FFN of vision transformer (ViT), which a
ttempts to efficiently explore rich statistics of token features for achieving b
etter performance and complexity trade-off. Specifically, our AFBO computes seco
nd-order statistics via a spatial-channel factorized bilinear operation for feat
ure learning, which replaces a simple linear projection in FFN and enhances the
feature learning ability of ViT by modeling second-order correlation among token
 features. Furthermore, our AFBO presents two structured-sparsity channel mappin
g strategies, namely Grouped Cross Channel Mapping (GCCM) and Overlapped Cycle C
hannel Mapping (OCCM). They decompose bilinear operation into grouped channel fe
atures by considering information interaction between groups, significantly redu
cing computational complexity while guaranteeing model performance. Finally, our
 AFBO is built with GCCM and OCCM in an asymmetric way, aiming to achieve a bett
er trade-off. Note that our AFBO is model-agnostic, which can be flexibly integr
ated with existing ViTs. Experiments are conducted with twenty ViTs on various t
asks, and the results show our AFBO is superior to its counterparts while improv
ing existing ViTs in terms of generalization and robustness.
****

Sulin Liu,Juno Nam,Andrew Campbell,Hannes Stark,Yilun Xu,Tommi Jaakkola,Rafael G
omez-Bombarelli
Think while You Generate: Discrete Diffusion with Planned Denoising
Discrete diffusion has achieved state-of-the-art performance, outperforming or a
pproaching autoregressive models on standard benchmarks. In this work, we introd
uce *Discrete Diffusion with Planned Denoising* (DDPD), a novel framework that s
eparates the generation process into two models: a planner and a denoiser. At in
ference time, the planner selects which positions to denoise next by identifying
 the most corrupted positions in need of denoising, including both initially cor
rupted and those requiring additional refinement. This plan-and-denoise approach
 enables more efficient reconstruction during generation by iteratively identify
ing and denoising corruptions in the optimal order. DDPD outperforms traditional
 denoiser-only mask diffusion methods, achieving superior results on language mo
deling benchmarks such as *text8*, *OpenWebText*, and token-based generation on
ImageNet 256 × 256. Notably, in language modeling, DDPD significantly reduces
the performance gap between diffusion-based and autoregressive methods in terms
of generative perplexity. Code is available at [github.com/liusulin/DDPD](https
://github.com/liusulin/DDPD).
****

Mansi Sakarvadia,Aswathy Ajith,Arham Mushtaq Khan,Nathaniel C Hudson,Caleb Genie
sse,Kyle Chard,Yaoqing Yang,Ian Foster,Michael W. Mahoney
Mitigating Memorization in Language Models
Language models (LMs) can “memorize” information, i.e., encode training data in
their weights in such a way that inference-time queries can lead to verbatim reg
urgitation of that data. This ability to extract training data can be problemati
c, for example, when data are private or sensitive. In this work, we investigate
 methods to mitigate memorization: three regularizer-based, three fine-tuning-ba
sed, and eleven machine unlearning-based methods, with five of the latter being
new methods that we introduce. We also introduce TinyMem, a suite of small, comp
utationally-efficient LMs for the rapid development and evaluation of memorizati
on-mitigation methods. We demonstrate that the mitigation methods that we develo
p using TinyMem can successfully be applied to production-grade LMs, and we dete
rmine via experiment that: regularizer-based mitigation methods are slow and ine
ffective at curbing memorization; fine-tuning-based methods
are effective at curbing memorization, but overly expensive, especially for reta
ining higher accuracies; and unlearning-based methods are faster and more effect
ive, allowing for the precise localization and removal of memorized information
from LM weights prior to inference. We show, in particular, that our proposed un
learning method BalancedSubnet outperforms other mitigation methods at removing
memorized information while preserving performance on target tasks.

****
Yansong Peng,Hebei Li,Peixi Wu,Yueyi Zhang,Xiaoyan Sun,Feng Wu
D-FINE: Redefine Regression Task of DETRs as Fine-grained Distribution Refinemen
t
We introduce D-FINE, a powerful real-time object detector that achieves outstand
ing localization precision by redefining the bounding box regression task in DET
R models. D-FINE comprises two key components: Fine-grained Distribution Refinem
ent (FDR) and Global Optimal Localization Self-Distillation (GO-LSD). FDR transf
orms the regression process from predicting fixed coordinates to iteratively ref
ining probability distributions, providing a fine-grained intermediate represent
ation that significantly enhances localization accuracy. GO-LSD is a bidirection
al optimization strategy that transfers localization knowledge from refined dist
ributions to shallower layers through self-distillation, while also simplifying
the residual prediction tasks for deeper layers. Additionally, D-FINE incorporat
es lightweight optimizations in computationally intensive modules and operations
, achieving a better balance between speed and accuracy. Specifically, D-FINE-L
/ X achieves 54.0% / 55.8% AP on the COCO dataset at 124 / 78 FPS on an NVIDIA T
4 GPU. When pretrained on Objects365, D-FINE-L / X attains 57.1% / 59.3% AP, sur
passing all existing real-time detectors. Furthermore, our method significantly
enhances the performance of a wide range of DETR models by up to 5.3% AP with ne
gligible extra parameters and training costs. Our code and models: https://githu
b.com/Peterande/D-FINE.
****
Marc Anton Finzi,Sanyam Kapoor,Diego Granziol,Anming Gu,Christopher De Sa,J Zico
Kolter,Andrew Gordon Wilson
Compute-Optimal LLMs Provably Generalize Better with Scale
Why do larger language models generalize better? To explore this question, we de
velop generalization bounds on the pretraining objective of large language model
s (LLMs) in the compute-optimal regime, as described by the Chinchilla scaling l
aws. We introduce a novel, fully empirical Freedman-type martingale concentratio
n inequality that tightens existing bounds by accounting for the variance of the
loss function. The generalization bound can be broken into three contributions:
the number of parameters per token, the loss variance, and the quantization err
or at a fixed bitrate. As language models are scaled up, the number of parameter
s per data point stays constant; however, both the loss variance and the quantiz
ation error decrease, implying that larger models should have \emph{smaller} gen
eralization gaps. We examine why larger models tend to be more quantizable from
an information theoretic perspective, showing that the rate at which they can in
tegrate new information grows slower than their capacity on the compute optimal
frontier. From these findings we produce a scaling law for the generalization ga
p, showing that our bounds decrease in a predictable way.
****
Nikita Balagansky,Ian Maksimov,Daniil Gavrilov
Mechanistic Permutability: Match Features Across Layers
Understanding how features evolve across layers in deep neural networks is a fun
damental challenge in mechanistic interpretability, particularly due to polysema
nticity and feature superposition. While Sparse Autoencoders (SAEs) have been us
ed to extract interpretable features from individual layers, aligning these feat
ures across layers has remained an open problem. In this paper, we introduce SAE
Match, a novel, data-free method for aligning SAE features across different lay
ers of a neural network. Our approach involves matching features by minimizing t
he mean squared error between the folded parameters of SAEs, a technique that in
corporates activation thresholds into the encoder and decoder weights to account
for differences in feature scales. Through extensive experiments on the Gemma 2
language model, we demonstrate that our method effectively captures feature evo
lution across layers, improving feature matching quality. We also show that feat
ures persist over several layers and that our approach can approximate hidden st
ates across layers. Our work advances the understanding of feature dynamics in n
eural networks and provides a new tool for mechanistic interpretability studies.
****

Jonathan Light,Yue Wu,Yiyou Sun,Wenchao Yu,Yanchi Liu,Xujiang Zhao,Ziniu Hu,Haif
eng Chen,Wei Cheng
SFS: Smarter Code Space Search improves LLM Inference Scaling
We frame code generation as a black-box optimization problem within the code
space and demonstrate how optimization-inspired techniques can enhance inference
scaling over text. Based on this perspective, we propose **SCATTERED FOREST
SEARCH (SFS)**, a novel approach that improves solution diversity during evoluti
onary search,
thereby avoiding local optima. Our theoretical analysis illustrates how these
methods improve exploration and enhance efficiency. Extensive experiments
on *HumanEval, MBPP, APPS, CodeContests,* and *Leetcode* reveal significant
performance gains. For instance, our method achieves a **pass@1 rate of 67.1% on
HumanEval+** and **87.2% on HumanEval with GPT-3.5**, marking improvements of
8.6% and **4.3%** over the state-of-the-art, while also halving the iteratio
ns needed
to find the correct solution. Furthermore, our approach scales more efficiently
than existing search techniques, including **tree search, line search,** and **r
epeated
sampling (Best of N)**.
****
Zeyu Yun,Juexiao Zhang,Yann LeCun,Yubei Chen
URLOST: Unsupervised Representation Learning without Stationarity or Topology
Unsupervised representation learning has seen tremendous progress. However, it i
s constrained by its reliance on domain specific stationarity and topology, a li
mitation not found in biological intelligence systems. For instance, unlike comp
uter vision, human vision can process visual signals sampled from highly irregul
ar and non-stationary sensors. We introduce a novel framework that learns from h
igh-dimensional data without prior knowledge of stationarity and topology. Our m
odel, abbreviated as URLOST, combines a learnable self-organizing layer, spectra
l clustering, and a masked autoencoder (MAE). We evaluate its effectiveness on t
hree diverse data modalities including simulated biological vision data, neural
recordings from the primary visual cortex, and gene expressions. Compared to sta
te-of-the-art unsupervised learning methods like SimCLR and MAE, our model excel
s at learning meaningful representations across diverse modalities without knowi
ng their stationarity or topology. It also outperforms other methods that are no
t dependent on these factors, setting a new benchmark in the field. We position
this work as a step toward unsupervised learning methods capable of generalizing
 across diverse high-dimensional data modalities.
****
Weixian Lei,Difei Gao,Mike Zheng Shou
Grounding Multimodal Large Language Model in GUI World
Recent advancements in Multimodal Large Language Models (MLLMs) have accelerated
 the development of Graphical User Interface (GUI) agents capable of automating
complex tasks across digital platforms. However, precise GUI element grounding r
emains a key challenge for accurate interaction and generalization. In this work
, we present an effective GUI grounding framework, which includes an automated d
ata collection engine that gathers extensive GUI screenshots and annotations to
ensure broad generalization. We also propose a lightweight and flexible GUI grou
nding module designed to efficiently localize UI elements by pre-training on the
 collected data, and introduce a novel method to integrate this module with MLLM
s for the effective execution of GUI tasks. Our approach demonstrates superior p
erformance in task accuracy and adaptability, as validated by benchmarks such as
 ScreenSpot, MiniWob, AITW, and Mind2Web.
****
Futa Kai Waseda,Ching-Chun Chang,Isao Echizen
Rethinking Invariance Regularization in Adversarial Training to Improve Robustne
ss-Accuracy Trade-off
Adversarial training often suffers from a robustness-accuracy trade-off, where a
chieving high robustness comes at the cost of accuracy.
One approach to mitigate this trade-off is leveraging invariance regularization,

which encourages model invariance under adversarial perturbations; however, it still leads to accuracy loss.
In this work, we closely analyze the challenges of using invariance regularization in adversarial training and understand how to address them.
Our analysis identifies two key issues: (1) a "gradient conflict" between invariance and classification objectives, leading to suboptimal convergence, and (2) the mixture distribution problem arising from diverged distributions between clean and adversarial inputs.
To address these issues, we propose Asymmetric Representation-regularized Adversarial Training (ARAT), which incorporates asymmetric invariance loss with stop-gradient operation and a predictor to avoid gradient conflict, and a split-BatchNorm (BN) structure to resolve the mixture distribution problem.
Our detailed analysis demonstrates that each component effectively addresses the identified issues, offering novel insights into adversarial defense.
ARAT shows superiority over existing methods across various settings. Finally, we discuss the implications of our findings to knowledge distillation-based defenses, providing a new perspective on their relative successes.
****

Seohong Park,Kevin Frans,Benjamin Eysenbach,Sergey Levine
OGBench: Benchmarking Offline Goal-Conditioned RL
Offline goal-conditioned reinforcement learning (GCRL) is a major problem in reinforcement learning (RL) because it provides a simple, unsupervised, and domain-agnostic way to acquire diverse behaviors and representations from unlabeled data without rewards. Despite the importance of this setting, we lack a standard benchmark that can systematically evaluate the capabilities of offline GCRL algorithms. In this work, we propose OGBench, a new, high-quality benchmark for algorithms research in offline goal-conditioned RL. OGBench consists of 8 types of environments, 85 datasets, and reference implementations of 6 representative offline GCRL algorithms. We have designed these challenging and realistic environments and datasets to directly probe different capabilities of algorithms, such as stitching, long-horizon reasoning, and the ability to handle high-dimensional inputs and stochasticity. While representative algorithms may rank similarly on prior benchmarks, our experiments reveal stark strengths and weaknesses in these different capabilities, providing a strong foundation for building new algorithms.
Project page: https://seohong.me/projects/ogbench
****

Boshen Xu,Ziheng Wang,Yang Du,Zhinan Song,Sipeng Zheng,Qin Jin
Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions?
Egocentric video-language pretraining is a crucial step in advancing the understanding of hand-object interactions in first-person scenarios. Despite successes on existing testbeds, we find that current EgoVLMs can be easily misled by simple modifications, such as changing the verbs or nouns in interaction descriptions, with models struggling to distinguish between these changes. This raises the question: "Do EgoVLMs truly understand hand-object interactions?'' To address this question, we introduce a benchmark called $\textbf{EgoHOIBench}$, revealing the performance limitation of current egocentric models when confronted with such challenges. We attribute this performance gap to insufficient fine-grained supervision and the greater difficulty EgoVLMs experience in recognizing verbs compared to nouns. To tackle these issues, we propose a novel asymmetric contrastive objective named $\textbf{EgoNCE++}$. For the video-to-text objective, we enhance text supervision by generating negative captions using large language models or leveraging pretrained vocabulary for HOI-related word substitutions. For the text-to-video objective, we focus on preserving an object-centric feature space that clusters video representations based on shared nouns. Extensive experiments demonstrate that EgoNCE++ significantly enhances EgoHOI understanding, leading to improved performance across various EgoVLMs in tasks such as multi-instance retrieval, action recognition, and temporal understanding. Our code is available at https://github.com/xuboshen/EgoNCEpp.
****

Luca Scofano,Alessio Sampieri,Tommaso Campari,Valentino Sacco,Indro Spinelli,Lam

berto Ballan,Fabio Galasso
Following the Human Thread in Social Navigation
The success of collaboration between humans and robots in shared environments relies on the robot's real-time adaptation to human motion. Specifically, in Social Navigation, the agent should be close enough to assist but ready to back up to let the human move freely, avoiding collisions. Human trajectories emerge as crucial cues in Social Navigation, but they are partially observable from the robot's egocentric view and computationally complex to process.

We present the first Social Dynamics Adaptation model (SDA) based on the robot's state-action history to infer the social dynamics. We propose a two-stage Reinforcement Learning framework: the first learns to encode the human trajectories into social dynamics and learns a motion policy conditioned on this encoded information, the current status, and the previous action. Here, the trajectories are fully visible, i.e., assumed as privileged information. In the second stage, the trained policy operates without direct access to trajectories. Instead, the model infers the social dynamics solely from the history of previous actions and statuses in real-time.
Tested on the novel Habitat 3.0 platform, SDA sets a novel state-of-the-art (SotA) performance in finding and following humans.

The code can be found at https://github.com/L-Scofano/SDA.
****
Hengwei Bian,Lingdong Kong,Haozhe Xie,Liang Pan,Yu Qiao,Ziwei Liu
DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes
Urban scene generation has been developing rapidly recently. However, existing methods primarily focus on generating static and single-frame scenes, overlooking the inherently dynamic nature of real-world driving environments. In this work, we introduce DynamicCity, a novel 4D occupancy generation framework capable of generating large-scale, high-quality dynamic 4D scenes with semantics. DynamicCity mainly consists of two key models. **1)** A VAE model for learning HexPlane as the compact 4D representation. Instead of using naive averaging operations, DynamicCity employs a novel **Projection Module** to effectively compress 4D features into six 2D feature maps for HexPlane construction, which significantly enhances HexPlane fitting quality (up to **12.56** mIoU gain). Furthermore, we utilize an **Expansion & Squeeze Strategy** to reconstruct 3D feature volumes in parallel, which improves both network training efficiency and reconstruction accuracy than naively querying each 3D point (up to **7.05** mIoU gain, **2.06x** training speedup, and **70.84\%** memory reduction). **2)** A DiT-based diffusion model for HexPlane generation. To make HexPlane feasible for DiT generation, a **Padded Rollout Operation** is proposed to reorganize all six feature planes of the HexPlane as a squared 2D feature map. In particular, various conditions could be introduced in the diffusion or sampling process, supporting **versatile 4D generation applications**, such as trajectory- and command-driven generation, inpainting, and layout-conditioned generation. Extensive experiments on the CarlaSC and Waymo datasets demonstrate that DynamicCity significantly outperforms existing state-of-the-art 4D occupancy generation methods across multiple metrics. The code and models have been released to facilitate future research.
****
Xue Li,Yiyou Sun,Wei Cheng,Yinglun Zhu,Haifeng Chen
Chain-of-region: Visual Language Models Need Details for Diagram Analysis
Visual Language Models (VLMs) like GPT-4V have broadened the scope of LLM applications, yet they face significant challenges in accurately processing visual details, particularly in scientific diagrams.
This paper explores the necessity of meticulous visual detail collection and region decomposition for enhancing the performance of VLMs in scientific diagram analysis. We propose a novel approach that combines traditional computer vision techniques with VLMs to systematically decompose diagrams into discernible visual elements and aggregate essential metadata. Our method employs techniques in OpenCV library to identify and label regions, followed by a refinement process using

shape detection and region merging algorithms, which are particularly suited to the structured nature of scientific diagrams. This strategy not only improves the granularity and accuracy of visual information processing but also extends the capabilities of VLMs beyond their current limitations. We validate our approach through a series of experiments that demonstrate enhanced performance in diagram analysis tasks, setting a new standard for integrating visual and language processing in a multimodal context.
****
Yuning Cui,Syed Waqas Zamir,Salman Khan,Alois Knoll,Mubarak Shah,Fahad Shahbaz Khan
AdaIR: Adaptive All-in-One Image Restoration via Frequency Mining and Modulation
In the image acquisition process, various forms of degradation, including noise, blur, haze, and rain, are frequently introduced. These degradations typically arise from the inherent limitations of cameras or unfavorable ambient conditions. To recover clean images from their degraded versions, numerous specialized restoration methods have been developed, each targeting a specific type of degradation. Recently, all-in-one algorithms have garnered significant attention by addressing different types of degradations within a single model without requiring the prior information of the input degradation type. However, most methods purely operate in the spatial domain and do not delve into the distinct frequency variations inherent to different degradation types. To address this gap, we propose an adaptive all-in-one image restoration network based on frequency mining and modulation. Our approach is motivated by the observation that different degradation types impact the image content on different frequency subbands, thereby requiring different treatments for each restoration task. Specifically, we first mine low- and high-frequency information from the input features, guided by the adaptively decoupled spectra of the degraded image. The extracted features are then modulated by a bidirectional operator to facilitate interactions between different frequency components. Finally, the modulated features are merged into the original input for a progressively guided restoration. With this approach, the model achieves adaptive reconstruction by accentuating the informative frequency subbands according to different input degradations. Extensive experiments demonstrate that the proposed method, AdaIR, achieves state-of-the-art performance on different image restoration tasks, including image denoising, dehazing, deraining, motion deblurring, and low-light image enhancement. The code is available at https://github.com/c-yn/AdaIR.
****
Junjielong Xu,Qinan Zhang,Zhiqing Zhong,Shilin He,Chaoyun Zhang,Qingwei Lin,Dan Pei,Pinjia He,Dongmei Zhang,Qi Zhang
OpenRCA: Can Large Language Models Locate the Root Cause of Software Failures?
Large language models (LLMs) are driving substantial advancements in software engineering, with successful applications like Copilot and Cursor transforming real-world development practices. However, current research predominantly focuses on the early stages of development, such as code generation, while overlooking the post-development phases that are crucial to user experience. To explore the potential of LLMs in this direction, we propose OpenRCA, a benchmark dataset and evaluation framework for assessing LLMs' ability to identify the root cause of software failures. OpenRCA includes 335 failures from three enterprise software systems, along with over 68 GB of telemetry data (logs, metrics, and traces). Given a failure case and its associated telemetry, the LLM is tasked to identify the root cause that triggered the failure, requiring comprehension of software dependencies and reasoning over heterogeneous, long-context telemetry data. Our results show substantial room for improvement, as current models can only handle the simplest cases. Even with the specially designed RCA-agent, the best-performing model, Claude 3.5, solved only 11.34% failure cases. Our work paves the way for future research in this direction.
****
Xiaosong Jia,Junqi You,Zhiyuan Zhang,Junchi Yan
DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving
End-to-end autonomous driving (E2E-AD) has emerged as a trend in the field of au

tonomous driving, promising a data-driven, scalable approach to system design. H
owever, existing E2E-AD methods usually adopt the sequential paradigm of percept
ion-prediction-planning, which leads to cumulative errors and training instabili
ty. The manual ordering of tasks also limits the system’s ability to leverage sy
nergies between tasks (for example, planning-aware perception and game-theoretic
 interactive prediction and planning). Moreover, the dense BEV representation ad
opted by existing methods brings computational challenges for long-range percept
ion and long-term temporal fusion. To address these challenges, we present Driv
eTransformer, a simplified E2E-AD framework for the ease of scaling up, characte
rized by three key features: Task Parallelism (All agent, map, and planning que
ries direct interact with each other at each block), Sparse Representation (Task
 queries direct interact with raw sensor features), and Streaming Processing (Ta
sk queries are stored and passed as history information). As a result, the new f
ramework is composed of three unified operations: task self-attention, sensor cr
oss-attention, temporal cross-attention, which significantly reduces the complex
ity of system and leads to better training stability. DriveTransformer achieves
state-of-the-art performance in both simulated closed-loop benchmark Bench2Drive
 and real world open-loop benchmark nuScenes with high FPS.
****

Xinyan Liang,Pinhan Fu,Yuhua Qian,Qian Guo,Guoqing Liu
Trusted Multi-View Classification via Evolutionary Multi-View Fusion
Multi-view classification based on the Dempster-Shafer theory is widely recogniz
ed for its reliability in safety-critical domains with multi-view data. However,
 the adoption of a late fusion strategy constrains information interaction among
 views, thereby leading to suboptimal utilization of multi-view data. A recent a
dvancement addressing this limitation involves generating a pseudo view by conca
tenating individual views. Yet, the efficacy of this pseudo view may diminish wh
en incorporating underperforming views like noisy views. Additionally, the integ
ration of a pseudo view exacerbates the issue of imbalanced multi-view learning,
 as it contains a disproportionate amount of information compared to individual
views. To address these issues, we propose the enhancing Trusted multi-view cla
ssification via Evolutionary multi-view Fusion (TEF) approach. TEF employs an ev
olutionary multi-view architecture search method to create a high-quality fusion
 architecture serving as the pseudo view, facilitating adaptive view and fusion
operator selection. Furthermore, TEF enhances each view within the fusion archit
ecture by concatenating the fusion architecture's decision output with its respe
ctive view. Our experimental results demonstrate the effectiveness of this strai
ghtforward yet powerful strategy in mitigating imbalanced multi-view learning is
sues, particularly on complex many-view datasets exceeding three views. Extensiv
e evaluations across 13 multi-view datasets validate the superior performance of
 our proposed method compared to other trusted multi-view learning approaches. T
he code is available at https://github.com/fupinhan123/TEF.
****

Nikolai Kalischek,Michael Oechsle,Fabian Manhardt,Philipp Henzler,Konrad Schindl
er,Federico Tombari
CubeDiff: Repurposing Diffusion-Based Image Models for Panorama Generation
We introduce a novel method for generating 360° panoramas from text prompts or i
mages. Our approach leverages recent advances in 3D generation by employing mult
i-view diffusion models to jointly synthesize the six faces of a cubemap. Unlike
 previous methods that rely on processing equirectangular projections or autoreg
ressive generation, our method treats each face as a standard perspective image,
 simplifying the generation process and enabling the use of existing multi-view
diffusion models. We demonstrate that these models can be adapted to produce hig
h-quality cubemaps without requiring correspondence-aware attention layers. Our
model allows for fine-grained text control, generates high resolution panorama i
mages and generalizes well beyond its training set, whilst achieving state-of-th
e-art results, both qualitatively and quantitatively.
****

Jiawei Yang,Jiahui Huang,Boris Ivanovic,Yuxiao Chen,Yan Wang,Boyi Li,Yurong You,
Apoorva Sharma,Maximilian Igl,Peter Karkus,Danfei Xu,Yue Wang,Marco Pavone

STORM: Spatio-TempOral Reconstruction Model For Large-Scale Outdoor Scenes
We present STORM, a spatio-temporal reconstruction model designed for reconstructing dynamic outdoor scenes from sparse observations. Existing dynamic reconstruction methods often rely on per-scene optimization, dense observations across space and time, and strong motion supervision, resulting in lengthy optimization times, limited generalization to novel views or scenes, and degenerated quality caused by noisy pseudo-labels for dynamics. To address these challenges, STORM leverages a data-driven Transformer architecture that directly infers dynamic 3D scene representations—parameterized by 3D Gaussians and their velocities—in a single forward pass. Our key design is to aggregate 3D Gaussians from all frames using self-supervised scene flows, transforming them to the target timestep to enable complete (i.e., "amodal") reconstructions from arbitrary viewpoints at any moment in time. As an emergent property, STORM automatically captures dynamic instances and generates high-quality masks using only reconstruction losses. Extensive experiments on public datasets show that STORM achieves precise dynamic scene reconstruction, surpassing state-of-the-art per-scene optimization methods (+4.3 to 6.6 PSNR) and existing feed-forward approaches (+2.1 to 4.7 PSNR) in dynamic regions. STORM reconstructs large-scale outdoor scenes in 200ms, supports real-time rendering, and outperforms competitors in scene flow estimation, improving 3D EPE by 0.422m and Acc5 by 28.02%. Beyond reconstruction, we showcase four additional applications of our model, illustrating the potential of self-supervised learning for broader dynamic scene understanding. For more details, please visit our project at https://jiawei-yang.github.io/STORM/.
****
Kangjie Zheng,Siyue Liang,Junwei Yang,Bin Feng,Zequn Liu,Wei Ju,Zhiping Xiao,Ming Zhang
SMI-Editor: Edit-based SMILES Language Model with Fragment-level Supervision
SMILES, a crucial textual representation of molecular structures, has garnered significant attention as a foundation for pre-trained language models (LMs). However, most existing pre-trained SMILES LMs focus solely on the single-token level supervision during pre-training, failing to fully leverage the substructural information of molecules. This limitation makes the pre-training task overly simplistic, preventing the models from capturing richer molecular semantic information. Moreover, during pre-training, these SMILES LMs only process corrupted SMILES inputs, never encountering any valid SMILES, which leads to a train-inference mismatch. To address these challenges, we propose SMI-Editor, a novel edit-based pre-trained SMILES LM. SMI-Editor disrupts substructures within a molecule at random and feeds the resulting SMILES back into the model, which then attempts to restore the original SMILES through an editing process. This approach not only introduces fragment-level training signals, but also enables the use of valid SMILES as inputs, allowing the model to learn how to reconstruct complete molecules from these incomplete structures. As a result, the model demonstrates improved scalability and an enhanced ability to capture fragment-level molecular information. Experimental results show that SMI-Editor achieves state-of-the-art performance across multiple downstream molecular tasks, and even outperforming several 3D molecular representation models.
****
Chenglei Si,Diyi Yang,Tatsunori Hashimoto
Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers
Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on

current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel (p < 0.05) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation.
****

Hanlin Wu,Yuxuan Song,Jingjing Gong,Ziyao Cao,Yawen Ouyang,Jianbing Zhang,Hao Zhou,Wei-Ying Ma,Jingjing Liu
A Periodic Bayesian Flow for Material Generation
Generative modeling of crystal data distribution is an important yet challenging task due to the unique periodic physical symmetry of crystals. Diffusion-based methods have shown early promise in modeling crystal distribution. More recently, Bayesian Flow Networks were introduced to aggregate noisy latent variables, resulting in a variance-reduced parameter space that has been shown to be advantageous for modeling Euclidean data distributions with structural constraints (Song, et al.,2023). Inspired by this, we seek to unlock its potential for modeling variables located in non-Euclidean manifolds e.g. those within crystal structures, by overcoming challenging theoretical issues. We introduce CrysBFN, a novel crystal generation method by proposing a periodic Bayesian flow, which essentially differs from the original Gaussian-based BFN by exhibiting non-monotonic entropy dynamics. To successfully realize the concept of periodic Bayesian flow, CrysBFN integrates a new entropy conditioning mechanism and empirically demonstrates its significance compared to time-conditioning. Extensive experiments over both crystal ab initio generation and crystal structure prediction tasks demonstrate the superiority of CrysBFN, which consistently achieves new state-of-the-art on all benchmarks. Surprisingly, we found that CrysBFN enjoys a significant improvement in sampling efficiency, e.g., 200x speedup (10 v.s. 2000 steps network forwards) compared with previous Diffusion-based methods on MP-20 dataset.
****

Cheng Lu,Yang Song
Simplifying, Stabilizing and Scaling Continuous-time Consistency Models
Consistency models (CMs) are a powerful class of diffusion-based generative models optimized for fast sampling. Most existing CMs are trained using discretized timesteps, which introduce additional hyperparameters and are prone to discretization errors. While continuous-time formulations can mitigate these issues, their success has been limited by training instability. To address this, we propose a simplified theoretical framework that unifies previous parameterizations of diffusion models and CMs, identifying the root causes of instability. Based on this analysis, we introduce key improvements in diffusion process parameterization, network architecture, and training objectives. These changes enable us to train continuous-time CMs at an unprecedented scale, reaching 1.5B parameters on ImageNet 512×512. Our proposed training algorithm, using only two sampling steps, achieves FID scores of 2.06 on CIFAR-10, 1.48 on ImageNet 64×64, and 1.88 on ImageNet 512×512, narrowing the gap in FID scores with the best existing diffusion models to within 10\%.
****

Dmitrii Zhemchuzhnikov,Sergei Grudinin
On the Fourier analysis in the SO(3) space : the EquiLoPO Network
Analyzing volumetric data with rotational invariance or equivariance is currently an active research topic. Existing deep-learning approaches utilize either group convolutional networks limited to discrete rotations or steerable convolutional networks with constrained filter structures. This work proposes a novel equivariant neural network architecture that achieves analytical Equivariance to Local Pattern Orientation on the continuous SO(3) group while allowing unconstrained trainable filters - EquiLoPO Network. Our key innovations are a group convolutional operation leveraging irreducible representations as the Fourier basis and a local activation function in the SO(3) space that provides a well-defined mapping from input to output functions, preserving equivariance. By integrating these operations into a ResNet-style architecture, we propose a model that overcomes the limitations of prior methods. A comprehensive evaluation on diverse 3D medic

al imaging datasets from MedMNIST3D demonstrates the effectiveness of our approach, which consistently outperforms state of the art. This work suggests the benefits of true rotational equivariance on SO(3) and flexible unconstrained filters enabled by the local activation function, providing a flexible framework for equivariant deep learning on volumetric data with potential applications across domains. Our code is publicly available at https://gricad-gitlab.univ-grenoble-alpes.fr/GruLab/ILPO/-/tree/main/EquiLoPO.
****
Arnav Kumar Jain,Harley Wiltzer,Jesse Farebrother,Irina Rish,Glen Berseth,Sanjiban Choudhury
Non-Adversarial Inverse Reinforcement Learning via Successor Feature Matching
In inverse reinforcement learning (IRL), an agent seeks to replicate expert demonstrations through interactions with the environment.
Traditionally, IRL is treated as an adversarial game, where an adversary searches over reward models, and a learner optimizes the reward through repeated RL procedures.
This game-solving approach is both computationally expensive and difficult to stabilize.
In this work, we propose a novel approach to IRL by _direct policy search_:
by exploiting a linear factorization of the return as the inner product of successor features and a reward vector, we design an IRL algorithm by policy gradient descent on the gap between the learner and expert features.
Our non-adversarial method does not require learning an explicit reward function and can be solved seamlessly with existing RL algorithms.
Remarkably, our approach works in state-only settings without expert action labels, a setting which behavior cloning (BC) cannot solve.
Empirical results demonstrate that our method learns from as few as a single expert demonstration and achieves improved performance on various control tasks.
****
Yehonathan Refael,Jonathan Svirsky,Boris Shustin,Wasim Huleihel,Ofir Lindenbaum
AdaRankGrad: Adaptive Gradient Rank and Moments for Memory-Efficient LLMs Training and Fine-Tuning
Training and fine-tuning large language models (LLMs) come with challenges related to memory and computational requirements due to the increasing size of the model weights and the optimizer states. To tackle these challenges, various techniques have been developed, such as low-rank adaptation (LoRA), which involves introducing a parallel trainable low-rank matrix to the fixed pre-trained weights at each layer. However, these methods often fall short compared to the full-rank weight training approach, as they restrict the parameter search to a low-rank subspace. This limitation can disrupt training dynamics and may require a full-rank warm start to mitigate the impact.
In this paper, we introduce a new method inspired by a phenomenon we formally prove: as training progresses, the rank of the estimated layer gradients gradually decreases and asymptotically approaches rank one. Leveraging this, our approach involves adaptively reducing the rank of the gradients during Adam optimization steps, using an efficient online-updating low-rank projections rule. We further present a randomized-svd scheme for efficiently finding the projection matrix.
Our technique enables full-parameter fine-tuning with adaptive low-rank gradient updates, significantly reducing overall memory requirements during training compared to state-of-the-art methods while improving model performance in both pretraining and fine-tuning. Finally, we provide a convergence analysis of our method and demonstrate its merits for training and fine-tuning language and biological foundation models.
****
Belinda Z. Li,Alex Tamkin,Noah Goodman,Jacob Andreas
Eliciting Human Preferences with Language Models
Language models (LMs) can be directed to perform user- and context-dependent tasks by using labeled examples or natural language prompts.
But selecting examples or writing prompts can be challenging---especially in tasks that require users to precisely articulate nebulous preferences or reason abo

ut complex edge cases. For such tasks, we introduce **Generative Active Task Elicitation (GATE)**, a method for using *LMs themselves* to guide the task specification process. GATE is a learning framework in which models elicit and infer human preferences through free-form, language-based interaction with users.
We identify prototypical challenges that users face when specifying preferences, and design three preference modeling tasks to study these challenges: content recommendation, moral reasoning, and email validation.
In preregistered experiments, we show that LMs that learn to perform these tasks using GATE (by interactively querying users with open-ended questions) obtain preference specifications that are more informative than user-written prompts or examples. GATE matches existing task specification methods in the moral reasoning task, and significantly outperforms them in the content recommendation and email validation tasks. Users additionally report that interactive task elicitation requires less effort than prompting or example labeling and surfaces considerations that they did not anticipate on their own. Our findings suggest that LM-driven elicitation can be a powerful tool for aligning models to complex human preferences and values.
****

Yuxuan Wu,Ziyu Wang,Bhiksha Raj,Gus Xia
Unsupervised Disentanglement of Content and Style via Variance-Invariance Constraints
We contribute an unsupervised method that effectively learns disentangled content and style representations from sequences of observations. Unlike most disentanglement algorithms that rely on domain-specific labels or knowledge, our method is based on the insight of domain-general statistical differences between content and style --- content varies more among different fragments within a sample but maintains an invariant vocabulary across data samples, whereas style remains relatively invariant within a sample but exhibits more significant variation across different samples. We integrate such inductive bias into an encoder-decoder architecture and name our method after V3 (variance-versus-invariance). Experimental results show that V3 generalizes across multiple domains and modalities, successfully learning disentangled content and style representations, such as pitch and timbre from music audio, digit and color from images of hand-written digits, and action and character appearance from simple animations. V3 demonstrates strong disentanglement performance compared to existing unsupervised methods, along with superior out-of-distribution generalization and few-shot learning capabilities compared to supervised counterparts. Lastly, symbolic-level interpretability emerges in the learned content codebook, forging a near one-to-one alignment between machine representation and human knowledge.
****

Yuanfei Wang,Xiaojie Zhang,Ruihai Wu,Yu Li,Yan Shen,Mingdong Wu,Zhaofeng He,Yizhou Wang,Hao Dong
AdaManip: Adaptive Articulated Object Manipulation Environments and Policy Learning
Articulated object manipulation is a critical capability for robots to perform various tasks in real-world scenarios.
Composed of multiple parts connected by joints, articulated objects are endowed with diverse functional mechanisms through complex relative motions. For example, a safe consists of a door, a handle, and a lock, where the door can only be opened when the latch is unlocked. The internal structure, such as the state of a lock or joint angle constraints, cannot be directly observed from visual observation. Consequently, successful manipulation of these objects requires adaptive adjustment based on trial and error rather than a one-time visual inference. However, previous datasets and simulation environments for articulated objects have primarily focused on simple manipulation mechanisms where the complete manipulation process can be inferred from the object's appearance. To enhance the diversity and complexity of adaptive manipulation mechanisms, we build a novel articulated object manipulation environment and equip it with 9 categories of objects. Based on the environment and objects, we further propose an adaptive demonstration collection and 3D visual diffusion-based imitation learning pipeline that lear

ns the adaptive manipulation policy. The effectiveness of our designs and propos
ed method is validated through both simulation and real-world experiments.
****

Zhuo Huang,Gang Niu,Bo Han,Masashi Sugiyama,Tongliang Liu
Towards Out-of-Modal Generalization without Instance-level Modal Correspondence
The world is understood from various modalities, such as appearance, sound, lang
uage, etc. Since each modality only partially represents objects in a certain ph
ysical meaning, leveraging additional ones is beneficial in both theory and prac
tice. However, exploiting novel modalities normally requires cross-modal pairs c
orresponding to the same instance, which is extremely resource-consuming and som
etimes even impossible, making knowledge exploration of novel modalities largely
 restricted. To seek practical multi-modal learning, here we study Out-of-Modal
(OOM) Generalization as an initial attempt to generalize to an unknown modality
without given instance-level modal correspondence. Specifically, we consider Sem
i-Supervised and Unsupervised scenarios of OOM Generalization, where the first h
as scarce correspondences and the second has none, and propose connect & explore
 (COX) to solve these problems. COX first connects OOM data and known In-Modal (
IM) data through a variational information bottleneck framework to extract share
d information. Then, COX leverages the shared knowledge to create emergent corre
spondences, which is theoretically justified from an information-theoretic persp
ective. As a result, the label information on OOM data emerges along with the co
rrespondences, which help explore the OOM data with unknown knowledge, thus bene
fiting generalization results. We carefully evaluate the proposed COX method und
er various OOM generalization scenarios, verifying its effectiveness and extensi
bility.
****

Wesley A. Suttle,Aamodh Suresh,Carlos Nieto-Granda
Behavioral Entropy-Guided Dataset Generation for Offline Reinforcement Learning
Entropy-based objectives are widely used to perform state space exploration in r
einforcement learning (RL) and dataset generation for offline RL. Behavioral ent
ropy (BE), a rigorous generalization of classical entropies that incorporates co
gnitive and perceptual biases of agents, was recently proposed for discrete sett
ings and shown to be a promising metric for robotic exploration problems. In thi
s work, we propose using BE as a principled exploration objective for systematic
ally generating datasets that provide diverse state space coverage in complex, c
ontinuous, potentially high-dimensional domains. To achieve this, we extend the
notion of BE to continuous settings, derive tractable k-nearest neighbor estim
ators, provide theoretical guarantees for these estimators, and develop practica
l reward functions that can be used with standard RL methods to learn BE-maximiz
ing policies. Using standard MuJoCo environments, we experimentally compare the
performance of offline RL algorithms for a variety of downstream tasks on datase
ts generated using BE, R\'{e}nyi, and Shannon entropy-maximizing policies, as we
ll as the SMM and RND algorithms. We find that offline RL algorithms trained on
datasets collected using BE outperform those trained on datasets collected using
 Shannon entropy, SMM, and RND on all tasks considered, and on 80\% of the tasks
 compared to datasets collected using Renyi entropy.
****

Isabella Liu,Hao Su,Xiaolong Wang
Dynamic Gaussians Mesh: Consistent Mesh Reconstruction from Dynamic Scenes
Modern 3D engines and graphics pipelines require mesh as a memory-efficient repr
esentation, which allows efficient rendering, geometry processing, texture editi
ng, and many other downstream operations. However, it is still highly difficult
to obtain high-quality mesh in terms of detailed structure and time consistency
from dynamic observations. To this end, we introduce Dynamic Gaussians Mesh (DG-
Mesh), a framework to reconstruct a high-fidelity and time-consistent mesh from
dynamic input. Our work leverages the recent advancement in 3D Gaussian Splattin
g to construct the mesh sequence with temporal consistency from dynamic observat
ions. Building on top of this representation, DG-Mesh recovers high-quality mesh
es from the Gaussian points and can track the mesh vertices over time, which ena
bles applications such as texture editing on dynamic objects. We introduce the G

aussian-Mesh Anchoring, which encourages evenly distributed Gaussians, resulting better mesh reconstruction through mesh-guided densification and pruning on the deformed Gaussians. By applying cycle-consistent deformation between the canonical and the deformed space, we can project the anchored Gaussian back to the canonical space and optimize Gaussians across all time frames. During the evaluation on different datasets, DG-Mesh provides significantly better mesh reconstruction and rendering than baselines.
****
Hainan Xu,Travis M. Bartley,Vladimir Bataev,Boris Ginsburg
Three-in-One: Fast and Accurate Transducer for Hybrid-Autoregressive ASR
We present Hybrid-Autoregressive INference TrANsducers (HAINAN), a novel architecture for speech recognition that extends the Token-and-Duration Transducer (TDT) model. Trained with randomly masked predictor network outputs, HAINAN supports both autoregressive inference with all network components and non-autoregressive inference without the predictor. Additionally, we propose a novel semi-autoregressive inference method that first generates an initial hypothesis using non-autoregressive inference, followed by refinement steps where each token prediction is regenerated using parallelized autoregression on the initial hypothesis. Experiments on multiple datasets across different languages demonstrate that HAINAN achieves efficiency parity with CTC in non-autoregressive mode and with TDT in autoregressive mode. In terms of accuracy, autoregressive HAINAN achieves parity with TDT and RNN-T, while non-autoregressive HAINAN significantly outperforms CTC. Semi-autoregressive inference further enhances the model's accuracy with minimal computational overhead, and even outperforms TDT results in some cases. These results highlight HAINAN's flexibility in balancing accuracy and speed, positioning it as a strong candidate for real-world speech recognition applications.
****
Vaibhav Saxena,Matthew Bronars,Nadun Ranawaka Arachchige,Kuancheng Wang,Woo Chul Shin,Soroush Nasiriany,Ajay Mandlekar,Danfei Xu
What Matters in Learning from Large-Scale Datasets for Robot Manipulation
Imitation learning from large multi-task demonstration datasets has emerged as a promising path for building generally-capable robots. As a result, 1000s of hours have been spent on building such large-scale datasets around the globe. Despite the continuous growth of such efforts, we still lack a systematic understanding of what data should be collected to improve the utility of a robotics dataset and facilitate downstream policy learning. In this work, we conduct a large-scale dataset composition study to answer this question. We develop a data generation framework to procedurally emulate common sources of diversity in existing datasets (such as sensor placements and object types and arrangements), and use it to generate large-scale robot datasets with controlled compositions, enabling a suite of dataset composition studies that would be prohibitively expensive in the real world. We focus on two practical settings: (1) what types of diversity should be emphasized when future researchers collect large-scale datasets for robotics, and (2) how should current practitioners retrieve relevant demonstrations from existing datasets to maximize downstream policy performance on tasks of interest. Our study yields several critical insights -- for example, we find that camera poses and spatial arrangements are crucial dimensions for both diversity in collection and alignment in retrieval. In real-world robot learning settings, we find that not only do our insights from simulation carry over, but our retrieval strategies on existing datasets such as DROID allow us to consistently outperform existing training strategies by up to 70\%.
****
Tao Lin,Yiling Chen
Generalized Principal-Agent Problem with a Learning Agent
Generalized principal-agent problems, including Stackelberg games, contract design, and Bayesian persuasion, are a class of economic problems where an agent best responds to a principal's committed strategy.
We study repeated generalized principal-agent problems under the assumption that the principal does not have commitment power and the agent uses algorithms to learn to respond to the principal. We reduce this problem to a one-shot generaliz

ed principal-agent problem where the agent approximately best responds. Using this reduction, we show that: (1) if the agent uses contextual no-regret learning algorithms with regret $\mathrm{Reg}(T)$, then the principal can guarantee utility at least $U^* - \Theta\big(\sqrt{\tfrac{\mathrm{Reg}(T)}{T}}\big)$, where U^* is the principal's optimal utility in the classic model with a best-responding agent.
(2) If the agent uses contextual no-swap-regret learning algorithms with swap-regret $\mathrm{SReg}(T)$, then the principal cannot obtain utility more than $U^* + O(\frac{\mathrm{SReg(T)}}{T})$.
But (3) if the agent uses mean-based learning algorithms (which can be no-regret but not no-swap-regret), then the principal can sometimes do significantly better than U^*.
These results not only refine previous results in Stackelberg games and contract design, but also lead to new results for Bayesian persuasion with a learning agent and all generalized principal-agent problems where the agent does not have private information.
****
Wen Wang,Qiuyu Wang,Kecheng Zheng,Hao OUYANG,Zhekai Chen,Biao Gong,Hao Chen,Yujun Shen,Chunhua Shen
Framer: Interactive Frame Interpolation
We propose Framer for interactive frame interpolation, which targets producing smoothly transitioning frames between two images as per user creativity. Concretely, besides taking the start and end frames as inputs, our approach supports customizing the transition process by tailoring the trajectory of some selected key points. Such a design enjoys two clear benefits. First, incorporating human interaction mitigates the issue arising from numerous possibilities of transforming one image to another, and in turn enables finer control of local motions. Second, as the most basic form of interaction, keypoints help establish the correspondence across frames, enhancing the model to handle challenging cases (e.g., objects on the start and end frames are of different shapes and styles). It is noteworthy that our system also offers an "autopilot" mode, where we introduce a module to estimate the keypoints and refine the trajectory automatically, to simplify the usage in practice. Extensive experimental results demonstrate the appealing performance of Framer on various applications, such as image morphing, time-lapse video generation, cartoon interpolation, etc. The code, model, and interface are publicly accessible at https://github.com/aim-uofa/Framer.
****
Xiaoyuan Cheng,Yi He,Yiming Yang,Xiao Xue,Sibo Cheng,Daniel Giles,Xiaohang Tang,Yukun Hu
Learning Chaos In A Linear Way
Learning long-term behaviors in chaotic dynamical systems, such as turbulent flows and climate modelling, is challenging due to their inherent instability and unpredictability. These systems exhibit positive Lyapunov exponents, which significantly hinder accurate long-term forecasting. As a result, understanding long-term statistical behavior is far more valuable than focusing on short-term accuracy. While autoregressive deep sequence models have been applied to capture long-term behavior, they often lead to exponentially increasing errors in learned dynamics. To address this, we shift the focus from simple prediction errors to preserving an invariant measure in dissipative chaotic systems. These systems have attractors, where trajectories settle, and the invariant measure is the probability distribution on attractors that remains unchanged under dynamics. Existing methods generate long trajectories of dissipative chaotic systems by aligning invariant measures, but it is not always possible to obtain invariant measures for arbitrary datasets. We propose the Poincaré Flow Neural Network (PFNN), a novel operator learning framework designed to capture behaviors of chaotic systems without any explicit knowledge of the invariant measure. PFNN employs an auto-encoder to map the chaotic system to a finite-dimensional feature space, effectively linearizing the chaotic evolution. It then learns the linear evolution operators to match the physical dynamics by addressing two critical properties in dissipative chaotic systems: (1) contraction, the system’s convergence toward its attra

ctors, and (2) measure invariance, trajectories on the attractors following a pr
obability distribution invariant to the dynamics.
Our experiments on a variety of chaotic systems, including Lorenz systems, Kuram
oto-Sivashinsky equation and Navier–Stokes equation, demonstrate that PFNN has m
ore accurate predictions and physical statistics compared to competitive baselin
es including the Fourier Neural Operator and the Markov Neural Operator.
****
Guibin Zhang,Yanwei Yue,Zhixun Li,Sukwon Yun,Guancheng Wan,Kun Wang,Dawei Cheng,
Jeffrey Xu Yu,Tianlong Chen
Cut the Crap: An Economical Communication Pipeline for LLM-based Multi-Agent Sys
tems
Recent advancements in large language model (LLM)-powered agents have shown that
 collective intelligence can significantly outperform individual capabilities, l
argely attributed to the meticulously designed inter-agent communication topolog
ies. Though impressive in performance, existing multi-agent pipelines inherently
 introduce substantial token overhead, as well as increased economic costs, whic
h pose challenges for their large-scale deployments. In response to this challen
ge, we propose an economical, simple, and robust multi-agent communication frame
work, termed $\texttt{AgentPrune}$, which can seamlessly integrate into mainstre
am multi-agent systems and prunes redundant or even malicious communication mess
ages. Technically, $\texttt{AgentPrune}$ is the first to identify and formally d
efine the $\textit{Communication Redundancy}$ issue present in current LLM-based
 multi-agent pipelines, and efficiently performs one-shot pruning on the spatial
-temporal message-passing graph, yielding a token-economic and high-performing c
ommunication topology.
Extensive experiments across six benchmarks demonstrate that $\texttt{AgentPrune
}$ $\textbf{(I)}$ achieves comparable results as state-of-the-art topologies at
merely $\\$5.6$ cost compared to their $\\$43.7$, $\textbf{(II)}$ integrates sea
mlessly into existing multi-agent frameworks with $28.1\\%\sim72.8\\%\downarrow$
 token reduction, and $\textbf{(III)}$ successfully defend against two types of
agent-based adversarial attacks with $3.5\\%\sim10.8\\%\uparrow$ performance boo
st. The source code is available at \url{https://github.com/yanweiyue/AgentPrune
}.
****
Hubert Baniecki,Giuseppe Casalicchio,Bernd Bischl,Przemyslaw Biecek
Efficient and Accurate Explanation Estimation with Distribution Compression
We discover a theoretical connection between explanation estimation and distribu
tion compression that significantly improves the approximation of feature attrib
utions, importance, and effects. While the exact computation of various machine
learning explanations requires numerous model inferences and becomes impractical
, the computational cost of approximation increases with an ever-increasing size
 of data and model parameters. We show that the standard i.i.d. sampling used in
 a broad spectrum of algorithms for post-hoc explanation leads to an approximati
on error worthy of improvement. To this end, we introduce Compress Then Explain
(CTE), a new paradigm of sample-efficient explainability. It relies on distribut
ion compression through kernel thinning to obtain a data sample that best approx
imates its marginal distribution. CTE significantly improves the accuracy and st
ability of explanation estimation with negligible computational overhead. It oft
en achieves an on-par explanation approximation error 2-3x faster by using fewer
 samples, i.e. requiring 2-3x fewer model evaluations. CTE is a simple, yet powe
rful, plug-in for any explanation method that now relies on i.i.d. sampling.
****
Seyedarmin Azizi,Souvik Kundu,Mohammad Erfan Sadeghi,Massoud Pedram
MambaExtend: A Training-Free Approach to Improve Long Context Extension of Mamba
The inherent quadratic complexity of the attention mechanism in transformer mode
ls has driven the research community to explore alternative architectures with s
ub-quadratic complexity, such as state-space models. Mamba has established itsel
f as a leading model within this emerging paradigm, achieving state-of-the-art r
esults in various language modeling benchmarks. However, despite its impressive
performance, Mamba's effectiveness is limited by its pre-training context length

, resulting in a pronounced degradation when the model is tasked with handling l
onger contexts. Our investigation reveals that Mamba's inability to generalize e
ffectively to long contexts is primarily due to the out-of-distribution (OOD) di
scretization steps. To address this critical limitation, we introduce _**MambaEx
tend**_, a novel framework designed to significantly enhance the context extensi
on capabilities of Mamba. Specifically, MambaExtend leverages a _**training-free
**_ approach to calibrate _only_ the scaling factors of discretization modules f
or different layers. We demonstrate both gradient-based and gradient-free zeroth
-order optimization to learn the optimal scaling factors for each Mamba layer, r
equiring orders of magnitude fewer updates as opposed to the parameter fine-tuni
ng-based alternatives.
Using this approach, we achieve a training-free context extension of up to 32x,
expanding the context from 2k to 64k tokens with minimal increases in perplexity
. In contrast to existing fine-tuning methods, MambaExtend selectively calibrate
s the scaling factors, requiring up to $\mathbf{5.42 * 10^6} \times$ fewer param
eter updates and incurring up to $\mathbf{3.87} \times$ lower peak memory usage,
 while delivering comparable or superior long-context performance across multipl
e tasks. Codes and checkpoints are available here^1.
****
Xinwei Zhang,Zhiqi Bu,Borja Balle,Mingyi Hong,Meisam Razaviyayn,Vahab Mirrokni
DiSK: Differentially Private Optimizer with Simplified Kalman Filter for Noise R
eduction
Differential privacy (DP) offers a robust framework for safeguarding individual
data privacy. To utilize DP in training modern machine learning models, differe
ntially private optimizers have been widely used in recent years. A popular appr
oach to privatize an optimizer is to clip the individual gradients and add suffi
ciently large noise to the clipped gradient. This approach led to the developmen
t of DP optimizers that have comparable performance with their non-private count
erparts in fine-tuning tasks or in tasks with a small number of training paramet
ers. However, a significant performance drop is observed when these optimizers a
re applied to large-scale training. This degradation stems from the substantial
noise injection required to maintain DP, which disrupts the optimizer's dynamics
.
This paper introduces DiSK, a novel framework designed to significantly enhance
the performance of DP optimizers. DiSK employs Kalman filtering, a technique dra
wn from control and signal processing, to effectively denoise privatized gradien
ts and generate progressively refined gradient estimations. To ensure practicali
ty for large-scale training, we simplify the Kalman filtering process, minimizin
g its memory and computational demands.
We establish theoretical privacy-utility trade-off guarantees for DiSK, and demo
nstrate provable improvements over standard DP optimizers like DPSGD in terms of
 iteration complexity upper-bound.
Extensive experiments across diverse tasks, including vision tasks such as CIFAR
-100 and ImageNet-1k and language fine-tuning tasks such as GLUE, E2E, and DART,
 validate the effectiveness of DiSK. The results showcase its ability to signif
icantly improve the performance of DP optimizers, surpassing state-of-the-art re
sults under the same privacy constraints on several benchmarks.
****
Weicai Yan,Wang Lin,Zirun Guo,Ye Wang,Fangming Feng,Xiaoda Yang,Zehan Wang,Tao J
in
Diff-Prompt: Diffusion-driven Prompt Generator with Mask Supervision
Prompt learning has demonstrated promising results in fine-tuning pre-trained mu
ltimodal models. However, the performance improvement is limited when applied to
 more complex and fine-grained tasks. The reason is that most existing methods d
irectly optimize the parameters involved in the prompt generation process throug
h loss backpropagation, which constrains the richness and specificity of the pro
mpt representations. In this paper, we propose Diffusion-Driven Prompt Generator
 (Diff-Prompt), aiming to use the diffusion model to generate rich and fine-grai
ned prompt information for complex downstream tasks. Specifically, our approach
consists of three stages. In the first stage, we train a Mask-VAE to compress th

e masks into latent space. In the second stage, we leverage an improved Diffusio
n Transformer (DiT) to train a prompt generator in the latent space, using the m
asks for supervision. In the third stage, we align the denoising process of the
prompt generator with the pre-trained model in the semantic space, and use the g
enerated prompts to fine-tune the model. We conduct experiments on a complex pix
el-level downstream task, referring expression comprehension, and compare our me
thod with various parameter-efficient fine-tuning approaches. Diff-Prompt achiev
es a maximum improvement of 8.87 in R@1 and 14.05 in R@5 compared to the foundat
ion model and also outperforms other state-of-the-art methods across multiple me
trics. The experimental results validate the effectiveness of our approach and h
ighlight the potential of using generative models for prompt generation. Code is
 available at https://github.com/Kelvin-ywc/diff-prompt.
****
Zhenghai Xue,Bo An,Shuicheng YAN
Policy Optimization under Imperfect Human Interactions with Agent-Gated Shared A
utonomy
We introduce AGSA, an Agent-Gated Shared Autonomy framework that learns from hig
h-level human feedback to tackle the challenges of reward-free training, safe ex
ploration, and imperfect low-level human control. Recent human-in-the loop learn
ing methods enable human participants to intervene a learning agent’s control an
d provide online demonstrations. Nonetheless, these methods rely heavily on perf
ect human interactions, including accurate human-monitored intervention decision
s and near-optimal human demonstrations. AGSA employs a dedicated gating agent t
o determine when to switch control, thereby reducing the need of constant human
monitoring. To obtain a precise and foreseeable gating agent, AGSA trains a long
-term gating value function from human evaluative feedback on the gating agent’s
 intervention requests and preference feedback on pairs of human intervention tr
ajectories. Instead of relying on potentially suboptimal human demonstrations, t
he learning agent is trained using control-switching signals from the gating age
nt. We provide theoretical insights on performance bounds that respectively desc
ribe the ability of the two agents. Experiments are conducted with both simulate
d and real human participants at different skill levels in challenging continuou
s control environments. Comparative results highlight that AGSA achieves signifi
cant improvements over previous human-in-the-loop learning methods in terms of t
raining safety, policy performance, and user-friendliness.
****
Jerry Yao-Chieh Hu,Maojiang Su,En-jui kuo,Zhao Song,Han Liu
Computational Limits of Low-Rank Adaptation (LoRA) Fine-Tuning for Transformer M
odels
We study the computational limits of Low-Rank Adaptation (LoRA) for finetuning t
ransformer-based models using fine-grained complexity theory.
Our key observation is that the existence of low-rank decompositions within the
gradient computation of LoRA adaptation leads to possible algorithmic speedup.
This allows us to (i) identify a phase transition behavior of efficiency \blue{a
ssuming the Strong Exponential Time Hypothesis (SETH)}, and (ii) prove the exist
ence of almost linear algorithms by controlling the LoRA update computation term
 by term.
For the former, we identify a sharp transition in the efficiency of all possible
 rank-r LoRA update algorithms for transformers, based on specific norms resul
ting from the multiplications of the input sequence X, pretrained weights ${W^
\star}$, and adapter matrices $\alpha B A/r$.
Specifically, we derive a shared upper bound threshold for such norms and show t
hat efficient (sub-quadratic) approximation algorithms of LoRA exist only below
this threshold.
For the latter, we prove the existence of almost linear approximation algorithms
 for LoRA adaptation by utilizing the hierarchical low-rank structures of LoRA g
radients and approximating the gradients with a series of chained low-rank appro
ximations.
To showcase our theory, we consider two practical scenarios: partial (e.g., only
 W_V and W_Q) and full adaptations (e.g., W_Q, W_V, and W_K) of weight

s in attention heads.
****
Yiting Chen,Qitian Wu,Junchi Yan
Regularizing Energy among Training Samples for Out-of-Distribution Generalizatio
n
The energy-based model provides a unified framework for various learning models where an energy value is assigned to each configuration of random variables base d on probability. Recently, different methods have been proposed to derive an en ergy value out of the logits of a classifier for out-of-distribution (OOD) detec tion or OOD generalization. However, these methods mainly focus on the energy di fference between in-distribution and OOD data samples, neglecting the energy dif ference among in-distribution data samples. In this paper, we show that the ener gy among in-distribution data also requires attention. We propose to investigate the energy difference between in-distribution data samples. Both empirically an d theoretically, we show that previous methods for subpopulation shift (\emph{e. g.}, long-tail classification) such as data re-weighting and margin control appl y implicit energy regularization and we provide a unified framework from the ene rgy perspective. With the influence function, we further extend the energy regul arization framework to OOD generalization scenarios where the distribution shift is more implicit compared to the long-tail recognition scenario. We conduct exp eriments on long-tail datasets, subpopulation shift benchmarks, and OOD generali zation benchmarks to show the effectiveness of the proposed energy regularizatio n.
****
Matthew Dowling,Cristina Savin
Nonlinear multiregion neural dynamics with parametric impulse response communica tion channels
Cognition arises from the coordinated interaction of brain regions with distinct computational roles. Despite improvements in our ability to extract the dynamic s underlying circuit computation from population activity recorded in individual areas, understanding how multiple areas jointly support distributed computation remains a challenge. As part of this effort, we propose a multi-region neural d ynamics model composed of two building blocks: _i)_ within-region (potentially driven) nonlinear dynamics and _ii)_ communication channels between regions, par ameterized through their impulse response. Together, these choices make it possi ble to learn nonlinear neural population dynamics and understand the flow of inf ormation between regions by drawing from the rich literature of linear systems t heory. We develop a state noise inversion free variational filtering and learni ng algorithm for our model and show, through neuroscientifically inspired numeri cal experiments, how the proposed model can reveal interpretable characterizatio ns of the local computations within and the flow of information between neural p opulations. We further validate the efficacy of our approach using simultaneous population recordings from areas V1 and V2.
****
Haiyang Liu,Xingchao Yang,Tomoya Akiyama,Yuantian Huang,Qiaoge Li,Shigeru Kuriya ma,Takafumi Taketomi
TANGO: Co-Speech Gesture Video Reenactment with Hierarchical Audio Motion Embedd ing and Diffusion Interpolation
We present TANGO, a framework for generating co-speech body-gesture videos. Give n a few-minute, single-speaker reference video and target speech audio, TANGO pr oduces high-fidelity videos with synchronized body gestures. TANGO builds on Ges ture Video Reenactment (GVR), which splits and retrieves video clips using a dir ected graph structure - representing video frames as nodes and valid transitions as edges. We address two key limitations of GVR: audio-motion misalignment and visual artifacts in GAN-generated transition frames. In particular, i) we propos e retrieving gestures using latent feature distance to improve cross-modal align ment. To ensure the latent features could effectively model the relationship bet ween speech audio and gesture motion, we implement a hierarchical joint embeddin g space (AuMoClip); ii) we introduce the diffusion-based model to generate high- quality transition frames. Our diffusion model, Appearance Consistent Interpolat

ion (ACInterp), is built upon AnimateAnyone and includes a reference motion modu
le and homography background flow to preserve appearance consistency between gen
erated and reference videos. By integrating these components into the graph-base
d retrieval framework, TANGO reliably produces realistic, audio-synchronized vid
eos and outperforms all existing generative and retrieval methods. Our code, pre
trained models, and datasets are publicly available at https://github.com/CyberA
gentAILab/TANGO.
****
Shikun Feng,Yuyan Ni,Lu yan,Zhi-Ming Ma,Wei-Ying Ma,Yanyan Lan
UniGEM: A Unified Approach to Generation and Property Prediction for Molecules
Molecular generation and molecular property prediction are both crucial for drug
 discovery, but they are often developed independently. Inspired by recent studi
es, which demonstrate that diffusion model, a prominent generative approach, can
 learn meaningful data representations that enhance predictive tasks, we explore
 the potential for developing a unified generative model in the molecular domain
 that effectively addresses both molecular generation and property prediction ta
sks. However, the integration of these tasks is challenging due to inherent inco
nsistencies, making simple multi-task learning ineffective. To address this, we
propose UniGEM, the first unified model to successfully integrate molecular gene
ration and property prediction, delivering superior performance in both tasks. O
ur key innovation lies in a novel two-phase generative process, where predictive
 tasks are activated in the later stages, after the molecular scaffold is formed
. We further enhance task balance through innovative training strategies. Rigoro
us theoretical analysis and comprehensive experiments demonstrate our significan
t improvements in both tasks. The principles behind UniGEM hold promise for broa
der applications, including natural language processing and computer vision.
****
Yao Teng,Han Shi,Xian Liu,Xuefei Ning,Guohao Dai,Yu Wang,Zhenguo Li,Xihui Liu
Accelerating Auto-regressive Text-to-Image Generation with Training-free Specula
tive Jacobi Decoding
The current large auto-regressive models can generate high-quality, high-resolut
ion images, but these models require hundreds or even thousands of steps of next
-token prediction during inference, resulting in substantial time consumption. I
n existing studies, Jacobi decoding, an iterative parallel decoding algorithm, h
as been used to accelerate the auto-regressive generation and can be executed wi
thout training. However, the Jacobi decoding relies on a deterministic criterion
 to determine the convergence of iterations. Thus, it works for greedy decoding
but is incompatible with sampling-based decoding which is crucial for visual qua
lity and diversity in the current auto-regressive text-to-image generation. In t
his paper, we propose a training-free probabilistic parallel decoding algorithm,
 Speculative Jacobi Decoding (SJD), to accelerate auto-regressive text-to-image
generation. By introducing a probabilistic convergence criterion, our SJD accele
rates the inference of auto-regressive text-to-image generation while maintainin
g the randomness in sampling-based token decoding and allowing the model to gene
rate diverse images. Specifically, SJD facilitates the model to predict multiple
 tokens at each step and accepts tokens based on the probabilistic criterion, en
abling the model to generate images with fewer steps than the conventional next-
token-prediction paradigm. We also investigate the token initialization strategi
es that leverage the spatial locality of visual data to further improve the acce
leration ratio under specific scenarios. We conduct experiments for our proposed
 SJD on multiple auto-regressive text-to-image generation models, showing the ef
fectiveness of model acceleration without sacrificing the visual quality. The co
de of our work is available here: https://github.com/tyshiwo1/Accelerating-T2I-A
R-with-SJD/.
****
Zhe Li,Weihao Yuan,Yisheng HE,Lingteng Qiu,Shenhao Zhu,Xiaodong Gu,Weichao Shen,
Yuan Dong,Zilong Dong,Laurence Tianruo Yang
LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioni
ng
Language plays a vital role in the realm of human motion. Existing methods have

largely depended on CLIP text embeddings for motion generation, yet they fall sh
ort in effectively aligning language and motion due to CLIP’s pretraining on sta
tic image-text pairs. This work introduces LaMP, a novel Language-Motion Pretrai
ning model, which transitions from a language-vision to a more suitable language
-motion latent space. It addresses key limitations by generating motion-informat
ive text embeddings, significantly enhancing the relevance and semantics of gene
rated motion sequences. With LaMP, we advance three key tasks: text-to-motion ge
neration, motion-text retrieval, and motion captioning through aligned language-
motion representation learning. For generation, LaMP instead of CLIP provides th
e text condition, and an autoregressive masked prediction is designed to achieve
 mask modeling without rank collapse in transformers. For retrieval, motion feat
ures from LaMP’s motion transformer interact with query tokens to retrieve text
features from the text transformer, and vice versa. For captioning, we finetune
a large language model with the language-informative motion features to develop
a strong motion captioning model. In addition, we introduce the LaMP-BertScore m
etric to assess the alignment of generated motions with textual descriptions. Ex
tensive experimental results on multiple datasets demonstrate substantial improv
ements over previous methods across all three tasks. Project page: https://aigc3
d.github.io/LaMP
****
Jia Deng,Tianyi Tang,Yanbin Yin,Wenhao yang,Xin Zhao,Ji-Rong Wen
Neuron based Personality Trait Induction in Large Language Models
Large language models (LLMs) have become increasingly proficient at simulating v
arious personality traits, an important capability for supporting related applic
ations (e.g., role-playing). To further improve this capacity, in this paper, we
 present a neuron based approach for personality trait induction in LLMs, with t
hree major technical contributions. First, we construct PERSONALITYBENCH, a larg
e-scale dataset for identifying and evaluating personality traits in LLMs. This
dataset is grounded in the Big Five personality traits from psychology and desig
ned to assess the generative capabilities of LLMs towards specific personality t
raits. Second, by leveraging PERSONALITYBENCH, we propose an efficient method fo
r identifying personality-related neurons within LLMs by examining the opposite
aspects of a given trait. Third, we develop a simple yet effective induction met
hod that manipulates the values of these identified personality-related neurons,
 which enables fine-grained control over the traits exhibited by LLMs without tr
aining and modifying model parameters. Extensive experiments validates the effic
acy of our neuron identification and trait induction methods. Notably, our appro
ach achieves comparable performance as fine-tuned models, offering a more effici
ent and flexible solution for personality trait induction in LLMs.
****
Yuta Saito,Jihan Yao,Thorsten Joachims
POTEC: Off-Policy Contextual Bandits for Large Action Spaces via Policy Decompos
ition
We study off-policy learning (OPL) of contextual bandit policies in large discre
te action spaces where existing methods -- most of which rely crucially on rewar
d-regression models or importance-weighted policy gradients -- fail due to exces
sive bias or variance. To overcome these issues in OPL, we propose a novel two-s
tage algorithm, called Policy Optimization via Two-Stage Policy Decomposition (P
OTEC). It leverages clustering in the action space and learns two different poli
cies via policy- and regression-based approaches, respectively. In particular, w
e derive a novel low-variance gradient estimator that enables to learn a first-s
tage policy for cluster selection efficiently via a policy-based approach. To se
lect a specific action within the cluster sampled by the first-stage policy, POT
EC uses a second-stage policy derived from a regression-based approach within ea
ch cluster. We show that a local correctness condition, which only requires that
 the regression model preserves the relative expected reward differences of the
actions within each cluster, ensures that our policy-gradient estimator is unbia
sed and the second-stage policy is optimal. We also show that POTEC provides a s
trict generalization of policy- and regression-based approaches and their associ
ated assumptions. Comprehensive experiments demonstrate that POTEC provides subs

tantial improvements in OPL effectiveness particularly in large and structured a ction spaces.
****
Yang Cai,Gabriele Farina,Julien Grand-Clément,Christian Kroer,Chung-Wei Lee,Haip eng Luo,Weiqiang Zheng
Last-Iterate Convergence Properties of Regret-Matching Algorithms in Games
We study last-iterate convergence properties of algorithms for solving two-playe r zero-sum games based on Regret Matching$^+$ (RM$^+$). Despite their widespread use for solving real games, virtually nothing is known about their last-iterate convergence. A major obstacle to analyzing RM-type dynamics is that their regre t operators lack Lipschitzness and (pseudo)monotonicity.
We start by showing numerically that several variants used in practice, such as RM$^+$, predictive RM$^+$ and alternating RM$^+$, all lack last-iterate converge nce guarantees even on a simple 3×3 matrix game.
We then prove that recent variants of these algorithms based on a smoothing tech nique, extragradient RM^{+} and smooth Predictive RM$^+$, enjoy asymptotic la st-iterate convergence (without a rate), $1/\sqrt{t}$ best-iterate convergence, and when combined with restarting, linear-rate last-iterate convergence. Our ana lysis builds on a new characterization of the geometric structure of the limit p oints of our algorithms, marking a significant departure from most of the litera ture on last-iterate convergence. We believe that our analysis may be of indepen dent interest and offers a fresh perspective for studying last-iterate convergen ce in algorithms based on non-monotone operators.
****
Chi-Chih Chang,Wei-Cheng Lin,Chien-Yu Lin,Chong-Yan Chen,Yu-Fang Hu,Pei-Shuo Wan g,Ning-Chi Huang,Luis Ceze,Mohamed S. Abdelfattah,Kai-Chiang Wu
Palu: KV-Cache Compression with Low-Rank Projection
Post-training KV-Cache compression methods typically either sample a subset of e ffectual tokens or quantize the data into lower numerical bit width. However, th ese methods cannot exploit redundancy in the hidden dimension of the KV tenors. This paper presents a hidden dimension compression approach called Palu, a KV-Ca che compression framework that utilizes low-rank projection to reduce inference-time LLM memory usage. Palu decomposes the linear layers into low-rank matrices, caches compressed intermediate states, and reconstructs the full keys and value s on the fly. To improve accuracy, compression rate, and efficiency, Palu furthe r encompasses (1) a medium-grained low-rank decomposition scheme, (2) an efficie nt rank search algorithm, (3) low-rank-aware quantization compatibility enhancem ents, and (4) an optimized GPU kernel with matrix fusion. Extensive experiments with popular LLMs show that Palu compresses KV-Cache by 50% while maintaining st rong accuracy and delivering up to 1.89× speedup on the RoPE-based attention mod ule. When combined with quantization, Palu’s
inherent quantization-friendly design yields small to negligible extra accuracy degradation while saving additional memory than quantization-only methods and ac hieving up to 2.91× speedup for the RoPE-based attention. Moreover, it maintains comparable or even better accuracy (up to 1.19 lower perplexity) compared to qu antization-only methods. These results demonstrate Palu’s superior capability to effectively address the efficiency and memory challenges of LLM inference posed by KV-Cache. Our code is publicly available at: https://github.com/shadowpa0327 /Palu.
****
Yanbiao Ma,Wei Dai,Jiayi Chen
Pursuing Better Decision Boundaries for Long-Tailed Object Detection via Categor y Information Amount
In object detection, the number of instances is commonly used to determine wheth er a dataset follows a long-tailed distribution, implicitly assuming that the mo del will perform poorly on categories with fewer instances. This assumption has led to extensive research on category bias in datasets with imbalanced instance distributions. However, even in datasets with relatively balanced instance count s, models still exhibit bias toward certain categories, indicating that instance count alone cannot explain this phenomenon. In this work, we first introduce th

e concept and measurement of category informativeness. We observe a significant negative correlation between a category's informativeness and its accuracy, suggesting that informativeness more accurately reflects the learning difficulty of a category. Based on this observation, we propose the Informativeness-Guided Angular Margin Loss (IGAM Loss), which dynamically adjusts the decision space of categories according to their informativeness, thereby mitigating category bias in long-tailed datasets. IGAM Loss not only achieves superior performance on long-tailed benchmark datasets such as LVIS v1.0 and COCO-LT but also demonstrates significant improvements for underrepresented categories in non-long-tailed datasets like Pascal VOC. Extensive experiments confirm the potential of category informativeness as a tool and the generalizability of our proposed method.
****

Antoine Schnepf,Karim Kassab,Jean-Yves Franceschi,Laurent Caraffa,Flavian Vasile,Jeremie Mary,Andrew I. Comport,Valerie Gouet-Brunet
Bringing NeRFs to the Latent Space: Inverse Graphics Autoencoder
While pre-trained image autoencoders are increasingly utilized in computer vision, the application of inverse graphics in 2D latent spaces has been under-explored. Yet, besides reducing the training and rendering complexity, applying inverse graphics in the latent space enables a valuable interoperability with other latent-based 2D methods. The major challenge is that inverse graphics cannot be directly applied to such image latent spaces because they lack an underlying 3D geometry. In this paper, we propose an Inverse Graphics Autoencoder (IG-AE) that specifically addresses this issue. To this end, we regularize an image autoencoder with 3D-geometry by aligning its latent space with jointly trained latent 3D scenes. We utilize the trained IG-AE to bring NeRFs to the latent space with a latent NeRF training pipeline, which we implement in an open-source extension of the Nerfstudio framework, thereby unlocking latent scene learning for its supported methods. We experimentally confirm that Latent NeRFs trained with IG-AE present an improved quality compared to a standard autoencoder, all while exhibiting training and rendering accelerations with respect to NeRFs trained in the image space. Our project page can be found at https://ig-ae.github.io .
****

Ziqi Lu,Heng Yang,Danfei Xu,Boyi Li,Boris Ivanovic,Marco Pavone,Yue Wang
LoRA3D: Low-Rank Self-Calibration of 3D Geometric Foundation models
Emerging 3D geometric foundation models, such as DUSt3R, offer a promising approach for in-the-wild 3D vision tasks.
However, due to the high-dimensional nature of the problem space and scarcity of high-quality 3D data,
these pre-trained models still struggle to generalize to many challenging circumstances,
such as limited view overlap or low lighting.
To address this, we propose LoRA3D, an efficient self-calibration pipeline to *specialize* the pre-trained models to target scenes using their own multi-view predictions.
Taking sparse RGB images as input, we leverage robust optimization techniques to refine multi-view predictions and align them into a global coordinate frame.
In particular, we incorporate prediction confidence into the geometric optimization process,
automatically re-weighting the confidence to better reflect point estimation accuracy.
We use the calibrated confidence to generate high-quality pseudo labels for the calibrating views and fine-tune the models using low-rank adaptation (LoRA) on the pseudo-labeled data.
Our method does not require any external priors or manual labels. It completes the self-calibration process on a **single standard GPU within just 5 minutes**.
Each low-rank adapter requires only **18MB** of storage.
We evaluated our method on **more than 160 scenes** from the Replica, TUM and Waymo Open datasets,
achieving up to **88\% performance improvement** on 3D reconstruction, multi-view pose estimation and novel-view rendering.

For more details, please visit our project page at https://520xyxyzq.github.io/lora3d/.
****
Jiafei Lyu,Mengbei Yan,Zhongjian Qiao,Runze Liu,Xiaoteng Ma,Deheng Ye,Jing-Wen Yang,Zongqing Lu,Xiu Li
Cross-Domain Offline Policy Adaptation with Optimal Transport and Dataset Constraint
We explore cross-domain offline reinforcement learning (RL) where offline datasets from another domain can be accessed to facilitate policy learning. However, the underlying environments of the two datasets may have dynamics mismatches, incurring inferior performance when simply merging the data of two domains. Existing methods mitigate this issue by training domain classifiers, using contrastive learning methods, etc. Nevertheless, they still rely on a large amount of target domain data to function well. Instead, we address this problem by establishing a concrete performance bound of a policy given datasets from two domains. Motivated by the theoretical insights, we propose to align transitions in the two datasets using optimal transport and selectively share source domain samples, without training any neural networks. This enables reliable data filtering even given a few target domain data. Additionally, we introduce a dataset regularization term that ensures the learned policy remains within the scope of the target domain dataset, preventing it from being biased towards the source domain data. Consequently, we propose the Optimal Transport Data Filtering (dubbed OTDF) method and examine its effectiveness by conducting extensive experiments across various dynamics shift conditions (e.g., gravity shift), given limited target domain data. It turns out that OTDF exhibits superior performance on many tasks and dataset qualities, often surpassing prior strong baselines by a large margin.
****
Zhuoyi Yang,Jiayan Teng,Wendi Zheng,Ming Ding,Shiyu Huang,Jiazheng Xu,Yuanming Yang,Wenyi Hong,Xiaohan Zhang,Guanyu Feng,Da Yin,Yuxuan.Zhang,Weihan Wang,Yean Cheng,Bin Xu,Xiaotao Gu,Yuxiao Dong,Jie Tang
CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer
We present CogVideoX, a large-scale text-to-video generation model based on diffusion transformer, which can generate 10-second continuous videos that align seamlessly with text prompts, with a frame rate of 16 fps and resolution of 768 x 1360 pixels.
Previous video generation models often struggled with limited motion and short durations.
It is especially difficult to generate videos with coherent narratives based on text.
We propose several designs to address these issues.
First, we introduce a 3D Variational Autoencoder (VAE) to compress videos across spatial and temporal dimensions, enhancing both the compression rate and video fidelity.
Second, to improve text-video alignment, we propose an expert transformer with expert adaptive LayerNorm to facilitate the deep fusion between the two modalities.
Third, by employing progressive training and multi-resolution frame packing, CogVideoX excels at generating coherent, long-duration videos with diverse shapes and dynamic movements.
In addition, we develop an effective pipeline that includes various pre-processing strategies for text and video data.
Our innovative video captioning model significantly improves generation quality and semantic alignment.
Results show that CogVideoX achieves state-of-the-art performance in both automated benchmarks and human evaluation.
We publish the code and model checkpoints of CogVideoX along with our VAE model and video captioning model at https://github.com/THUDM/CogVideo.
****
Yuhang Li,Zhuying Li,Yuheng Jia
Complementary Label Learning with Positive Label Guessing and Negative Label Enh

ancement
Complementary label learning (CLL) is a weakly supervised learning paradigm that constructs a multi-class classifier only with complementary labels, specifying classes that the instance does not belong to. We reformulate CLL as an inverse problem that infers the full label information from the output space information. To be specific, we propose to split the inverse problem into two subtasks: positive label guessing (PLG) and negative label enhancement (NLE), collectively called PLNL. Specifically, we use well-designed criteria for evaluating the confidence of the model output, accordingly divide the training instances into three categories: highly-confident, moderately-confident and under-confident. For highly-confident instances, we perform PLG to assign them pseudo labels for supervised training. For moderately-confident and under-confident instances, we perform NLE by enhancing their complementary label set at different levels and train them with the augmented complementary labels iteratively. In addition, we unify PLG and NLE into a consistent framework, in which we can view all the pseudo-labeling-based methods from the perspective of negative label recovery. We prove that the error rates of both PLG and NLE are upper bounded, and based on that we can construct a classifier consistent with that learned by clean full labels. Extensive experiments demonstrate the superiority of PLNL over the state-of-the-art CLL methods, e.g., on STL-10, we increase the classification accuracy from 34.96\% to 55.25\%. The source code is available at https://github.com/yhli-ml/PLNL.
****
Taiyi Wang,Zhihao Wu,Jianheng Liu,Jianye HAO,Jun Wang,Kun Shao
DistRL: An Asynchronous Distributed Reinforcement Learning Framework for On-Device Control Agent
On-device control agents, especially on mobile devices, are responsible for operating mobile devices to fulfill users' requests, enabling seamless and intuitive interactions. Integrating Multimodal Large Language Models (MLLMs) into these agents enhances their ability to understand and execute complex commands, thereby improving user experience. However, fine-tuning MLLMs for on-device control presents significant challenges due to limited data availability and inefficient online training processes. This paper introduces DistRL, a novel framework designed to enhance the efficiency of online RL fine-tuning for mobile device control agents. DistRL employs centralized training and decentralized data acquisition to ensure efficient fine-tuning in the context of dynamic online interactions. Additionally, the framework is backed by our tailor-made RL algorithm, which effectively balances exploration with the prioritized utilization of collected data to ensure stable and robust training. Our experiments show that, on average, DistRL delivers a 3$\times$ improvement in training efficiency and enables training data collection 2.4$\times$ faster than the leading synchronous multi-machine methods. Notably, after training, DistRL achieves a 20\% relative improvement in success rate compared to state-of-the-art methods on general Android tasks from an open benchmark, significantly outperforming existing approaches while maintaining the same training time. These results validate DistRL as a scalable and efficient solution, offering substantial improvements in both training efficiency and agent performance for real-world, in-the-wild device control tasks.
****
Sravanti Addepalli,Yerram Varun,Arun Suggala,Karthikeyan Shanmugam,Prateek Jain
Does Safety Training of LLMs Generalize to Semantically Related Natural Prompts?
Large Language Models (LLMs) are known to be susceptible to crafted adversarial attacks or jailbreaks that lead to the generation of objectionable content despite being aligned to human preferences using safety fine-tuning methods. While the large dimensionality of input token space makes it inevitable to find *adversarial* prompts that can jailbreak these models, we aim to evaluate whether safety fine-tuned LLMs are safe against *natural* prompts which are semantically related to toxic seed prompts that elicit safe responses after alignment. We surprisingly find that popular aligned LLMs such as GPT-4 can be compromised using naive prompts that are NOT even crafted with an objective of jailbreaking the model. Furthermore, we empirically show that given a seed prompt that elicits a toxic response from an unaligned model, one can systematically generate several semanti

cally related *natural* prompts that can jailbreak aligned LLMs. Towards this, w
e propose a method of *Response Guided Question Augmentation (ReG-QA)* to evalua
te the generalization of safety aligned LLMs to natural prompts, that first gene
rates several toxic answers given a seed question using an unaligned LLM (Q to A
), and further leverages an LLM to generate questions that are likely to produce
 these answers (A to Q). We interestingly find that safety fine-tuned LLMs such
as GPT-4o are vulnerable to producing natural jailbreak *questions* from unsafe
content (without denial) and can thus be used for the latter (A to Q) step. We o
btain attack success rates that are comparable to/ better than leading adversari
al attack methods on the JailbreakBench leaderboard, while being significantly m
ore stable against defenses such as Smooth-LLM and Synonym Substitution, which a
re effective against existing all attacks on the leaderboard.
****
Ziyan Wang,Xiaoming Huo,Hao Wang
Towards Domain Adaptive Neural Contextual Bandits
Contextual bandit algorithms are essential for solving real-world decision makin
g problems. In practice, collecting a contextual bandit's feedback from differen
t domains may involve different costs. For example, measuring drug reaction from
 mice (as a source domain) and humans (as a target domain). Unfortunately, adapt
ing a contextual bandit algorithm from a source domain to a target domain with d
istribution shift still remains a major challenge and largely unexplored. In thi
s paper, we introduce the first general domain adaptation method for contextual
bandits. Our approach learns a bandit model for the target domain by collecting
feedback from the source domain. Our theoretical analysis shows that our algorit
hm maintains a sub-linear regret bound even adapting across domains. Empirical r
esults show that our approach outperforms the state-of-the-art contextual bandit
 algorithms on real-world datasets. Code will soon be available at https://githu
b.com/Wang-ML-Lab/DABand.
****
Xin Wang,Yu Zheng,Zhongwei Wan,Mi Zhang
SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model
Compression
The advancements in Large Language Models (LLMs) have been hindered by
their substantial sizes, which necessitates LLM compression methods for practica
l
deployment. Singular Value Decomposition (SVD) offers a promising solution for
LLM compression. However, state-of-the-art SVD-based LLM compression meth-
ods have two key limitations: truncating smaller singular values may lead to hig
her
compression loss, and the lack of update on the compressed weights after SVD
truncation. In this work, we propose SVD-LLM, a SVD-based post-training LLM
compression method that addresses the limitations of existing methods. SVD-LLM
incorporates a truncation-aware data whitening technique to ensure a direct map-
ping between singular values and compression loss. Moreover, SVD-LLM adopts
a parameter update with sequential low-rank approximation to compensate for
the accuracy degradation after SVD compression. We evaluate SVD-LLM on 10
datasets and seven models from three different LLM families at three different
scales. Our results demonstrate the superiority of SVD-LLM over state-of-the-art
s,
especially at high model compression ratios.
****
Sicheng Yu,CHENGKAI JIN,Huanyu Wang,Zhenghao Chen,Sheng Jin,ZHONGRONG ZUO,XU XIA
OLEI,Zhenbang Sun,Bingni Zhang,Jiawei Wu,Hao Zhang,Qianru Sun
Frame-Voyager: Learning to Query Frames for Video Large Language Models
Video Large Language Models (Video-LLMs) have made remarkable progress in video
understanding tasks. However, they are constrained by the maximum length of inpu
t tokens, making it impractical to input entire videos. Existing frame selection
 approaches, such as uniform frame sampling and text-frame retrieval, fail to ac
count for the information density variations in the videos or the complex instru
ctions in the tasks, leading to sub-optimal performance. In this paper, we propo

se Frame-Voyager that learns to query informative frame combinations, based on t
he given textual queries in the task. To train Frame-Voyager, we introduce a new
 data collection and labeling pipeline, by ranking frame combinations using a pr
e-trained Video-LLM. Given a video of M frames, we traverse its T-frame combinat
ions, feed them into a Video-LLM, and rank them based on Video-LLM's prediction
losses. Using this ranking as supervision, we train Frame-Voyager to query the f
rame combinations with lower losses. In experiments, we evaluate Frame-Voyager o
n four Video Question Answering benchmarks by plugging it into two different Vid
eo-LLMs. The experimental results demonstrate that Frame-Voyager achieves impres
sive results in all settings, highlighting its potential as a plug-and-play solu
tion for Video-LLMs.
****
Christina Sartzetaki,Gemma Roig,Cees G. M. Snoek,Iris Groen
One Hundred Neural Networks and Brains Watching Videos: Lessons from Alignment
What can we learn from comparing video models to human brains, arguably the most
 efficient and effective video processing systems in existence? Our work takes a
 step towards answering this question by performing the first large-scale benchm
arking of deep video models on representational alignment to the human brain, us
ing publicly available models and a recently released video brain imaging (fMRI)
 dataset. We disentangle four factors of variation in the models (temporal model
ing, classification task, architecture, and training dataset) that affect alignm
ent to the brain, which we measure by conducting Representational Similarity Ana
lysis across multiple brain regions and model layers. We show that temporal mode
ling is key for alignment to brain regions involved in early visual processing,
while a relevant classification task is key for alignment to higher-level region
s. Moreover, we identify clear differences between the brain scoring patterns ac
ross layers of CNNs and Transformers, and reveal how training dataset biases tra
nsfer to alignment with functionally selective brain areas. Additionally, we unc
over a negative correlation of computational complexity to brain alignment. Meas
uring a total of 99 neural networks and 10 human brains watching videos, we aim
to forge a path that widens our understanding of temporal and semantic video rep
resentations in brains and machines, ideally leading towards more efficient vide
o models and more mechanistic explanations of processing in the human brain.
****
M. Eren Akbiyik,Nedko Savov,Danda Pani Paudel,Nikola Popovic,Christian Vater,Otm
ar Hilliges,Luc Van Gool,Xi Wang
Leveraging Driver Field-of-View for Multimodal Ego-Trajectory Prediction
Understanding drivers’ decision-making is crucial for road safety. Although pred
icting the ego-vehicle’s path is valuable for driver-assistance systems, existin
g methods mainly focus on external factors like other vehicles’ motions, often n
eglecting the driver’s attention and intent. To address this gap, we infer the e
go-trajectory by integrating the driver’s gaze and the surrounding scene. We int
roduce RouteFormer, a novel multimodal ego-trajectory prediction network combini
ng GPS data, environmental context, and the driver's field-of-view—comprising fi
rst-person video and gaze fixations. We also present the Path Complexity Index (
PCI), a new metric for trajectory complexity that enables a more nuanced evaluat
ion of challenging scenarios. To tackle data scarcity and enhance diversity, we
introduce GEM, a comprehensive dataset of urban driving scenarios enriched with
synchronized driver field-of-view and gaze data. Extensive evaluations on GEM an
d DR(eye)VE demonstrate that RouteFormer significantly outperforms state-of-the-
art methods, achieving notable improvements in prediction accuracy across divers
e conditions. Ablation studies reveal that incorporating driver field-of-view da
ta yields significantly better average displacement error, especially in challen
ging scenarios with high PCI scores, underscoring the importance of modeling dri
ver attention. All data and code are available at meakbiyik.github.io/routeforme
r.
****
Joshua Kazdan,Hao Sun,Jiaqi Han,Felix Petersen,Frederick Vu,Stefano Ermon
CPSample: Classifier Protected Sampling for Guarding Training Data During Diffus
ion

Diffusion models have a tendency to exactly replicate their training data, espec
ially when trained on small datasets. Most prior work has sought to mitigate th
is problem by imposing differential privacy constraints or masking parts of the
training data, resulting in a notable substantial decrease in image quality. We
present CPSample, a method that modifies the sampling process to prevent trainin
g data replication while preserving image quality. CPSample utilizes a classifie
r that is trained to overfit on random binary labels attached to the training da
ta. CPSample then uses classifier guidance to steer the generation process away
from the set of points that can be classified with high certainty, a set that in
cludes the training data. CPSample achieves FID scores of 4.97 and 2.97 on CIFAR
-10 and CelebA-64, respectively, without producing exact replicates of the train
ing data. Unlike prior methods intended to guard the training images, CPSample
only requires training a classifier rather than retraining a diffusion model, wh
ich is computationally cheaper. Moreover, our technique provides diffusion model
s with greater robustness against membership inference attacks, wherein an adver
sary attempts to discern which images were in the model's training dataset. We s
how that CPSample behaves like a built-in rejection sampler, and we demonstrate
its capabilities to prevent mode collapse in Stable Diffusion.
****
Zihao Zhou,Rose Yu
Can LLMs Understand Time Series Anomalies?
Large Language Models (LLMs) have gained popularity in time series forecasting,
but their potential for anomaly detection remains largely unexplored. Our study
investigates whether LLMs can understand and detect anomalies in time series dat
a, focusing on zero-shot and few-shot scenarios. Inspired by conjectures about L
LMs' behavior from time series forecasting research, we formulate key hypotheses
 about LLMs' capabilities in time series anomaly detection. We design and conduc
t principled experiments to test each of these hypotheses. Our investigation rev
eals several surprising findings about LLMs for time series: (1) LLMs understand
 time series better as *images* rather than as text, (2) LLMs do not demonstrate
 enhanced performance when prompted to engage in *explicit reasoning* about time
 series analysis. (3) Contrary to common beliefs, LLMs' understanding of time se
ries *do not* stem from their repetition biases or arithmetic abilities. (4) LLM
s' behaviors and performance in time series analysis *vary significantly* across
 different models. This study provides the first comprehensive analysis of conte
mporary LLM capabilities in time series anomaly detection. Our results suggest t
hat while LLMs can understand trivial time series anomalies (we have no evidence
 that they can understand more subtle real-world anomalies), many common conject
ures based on their reasoning capabilities do not hold. All synthetic dataset ge
nerators, final prompts, and evaluation scripts have been made available in http
s://github.com/rose-stl-lab/anomllm.
****
Yu Pan,Chaozheng Wang,Zekai Wu,Qifan Wang,Min Zhang,Zenglin Xu
IDInit: A Universal and Stable Initialization Method for Neural Network Training
Deep neural networks have achieved remarkable accomplishments in practice. The s
uccess of these networks hinges on effective initialization methods, which are v
ital for ensuring stable and rapid convergence during training. Recently, initia
lization methods that maintain identity transition within layers have shown good
 efficiency in network training. These techniques (e.g., Fixup) set specific wei
ghts to zero to achieve identity control. However, settings of remaining weight
(e.g., Fixup uses random values to initialize non-zero weights) will affect the
inductive bias that is achieved only by a zero weight, which may be harmful to t
raining. Addressing this concern, we introduce fully identical initialization (I
DInit), a novel method that preserves identity in both the main and sub-stem lay
ers of residual networks. IDInit employs a padded identity-like matrix to overco
me rank constraints in non-square weight matrices. Furthermore, we show the conv
ergence problem of an identity matrix can be solved by stochastic gradient desce
nt. Additionally, we enhance the universality of IDInit by processing higher-ord
er weights and addressing dead neuron problems. IDInit is a straightforward yet
effective initialization method, with improved convergence, stability, and perfo

rmance across various settings, including large-scale datasets and deep models.
****
Han Lin,Tushar Nagarajan,Nicolas Ballas,Mido Assran,Mojtaba Komeili,Mohit Bansal
,Koustuv Sinha
VEDIT: Latent Prediction Architecture For Procedural Video Representation Learni
ng
Procedural video representation learning is an active research area where the ob
jective is to learn an agent which can anticipate and forecast the future given
the present video input, typically in conjunction with textual annotations. Prio
r works often rely on large-scale pretraining of visual encoders and prediction
models with language supervision. However, the necessity and effectiveness of ex
tending compute intensive pretraining to learn video clip sequences with noisy t
ext supervision have not yet been fully validated by previous works. In this wor
k, we show that a strong off-the-shelf frozen pretrained visual encoder, along w
ith a well designed prediction model, can achieve state-of-the-art (SoTA) perfor
mance in forecasting and procedural planning without the need for pretraining th
e prediction model, nor requiring additional supervision from language or ASR. I
nstead of learning representations from pixel space, our method utilizes the lat
ent embedding space of publicly available vision encoders. By conditioning on fr
ozen clip-level embeddings from observed steps to predict the actions of unseen
steps, our prediction model is able to learn robust representations for forecast
ing through iterative denoising —leveraging the recent advances in diffusion tra
nsformers (Peebles & Xie, 2023). Empirical studies over a total of five procedur
al learning tasks across four datasets (NIV, CrossTask, COIN and Ego4D-v2) show
that our model advances the strong baselines in long-horizon action anticipation
 (+2.6% in Verb ED@20, +3.1% in Noun ED@20), and significantly improves the SoTA
 in step forecasting (+5.0%), task classification (+3.8%), and procedure plannin
g tasks (up to +2.28% in success rate, +3.39% in mAcc, and +0.90% in mIoU).
****
Muthu Chidambaram,Rong Ge
For Better or For Worse? Learning Minimum Variance Features With Label Augmentat
ion
Data augmentation has been pivotal in successfully training deep learning models
 on classification tasks over the past decade. An important subclass of data aug
mentation techniques - which includes both label smoothing and Mixup - involves
modifying not only the input data but also the input label during model training
. In this work, we analyze the role played by the label augmentation aspect of s
uch methods. We first prove that linear models on binary classification data tra
ined with label augmentation learn only the minimum variance features in the dat
a, while standard training (which includes weight decay) can learn higher varian
ce features. We then use our techniques to show that even for nonlinear models a
nd general data distributions, the label smoothing and Mixup losses are lower bo
unded by a function of the model output variance. Lastly, we demonstrate empiric
ally that this aspect of label smoothing and Mixup can be a positive and a negat
ive. On the one hand, we show that the strong performance of label smoothing and
 Mixup on image classification benchmarks is correlated with learning low varian
ce hidden representations. On the other hand, we show that Mixup and label smoot
hing can be more susceptible to low variance spurious correlations in the traini
ng data.
****
Yitian Zhang,Xu Ma,Yue Bai,Huan Wang,Yun Fu
Accessing Vision Foundation Models via ImageNet-1K
Vision foundation models are renowned for the generalization ability due to mass
ive training data. Nevertheless, they demand tremendous training resources, and
the training data is often inaccessible, e.g., CLIP, DINOv2, posing great challe
nges to developing derivatives that could facilitate the research. In this work,
 we offer a very simple and general solution, named Proteus, to distill foundati
on models into smaller equivalents on ImageNet-1K without access to the original
 training data. Specifically, we remove the designs from conventional knowledge
distillation settings that result in dataset bias and present three levels of tr

aining objectives, i.e., token, patch, and feature, to maximize the efficacy of knowledge transfer. In this manner, Proteus is trained at ImageNet-level costs with surprising ability, facilitating the accessibility of training foundation models for the broader research community. When leveraging DINOv2-g/14 as the teacher, Proteus-L/14 matches the performance of the Oracle method DINOv2-L/14 (142M training data) across 19 benchmarks and outperforms other vision foundation models including CLIP-L/14 (400M), OpenCLIP-L/14 (400M/2B) and SynCLR-L/14 (600M) with a significantly smaller training set of 1.2M images.
****
Sheng Liu,Haotian Ye,James Zou
Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering
Hallucination poses a challenge to the deployment of large vision-language models (LVLMs) in applications. Unlike in large language models (LLMs), hallucination in LVLMs often arises from misalignments between visual inputs and textual outputs. This paper investigates the underlying mechanisms of hallucination, focusing on the unique structure of LVLMs that distinguishes them from LLMs. We identify that hallucinations often arise from the sensitivity of text decoders to vision inputs, a natural phenomenon when image encoders and text decoders are pre-trained separately. Inspired by this, we introduce Visual and Textual Intervention (VTI), a novel technique designed to reduce hallucinations by steering latent space representations during inference to enhance the stability of vision features. As a task-agnostic test-time intervention, VTI can be easily applied to any problem without additional training costs. Extensive experiments demonstrate that it can effectively reduce hallucinations and outperform baseline methods across multiple metrics, highlighting the critical role of vision feature stability in LVLMs.
****
Cheng Zhang,Jeffrey T. H. Wong,Can Xiao,George Anthony Constantinides,Yiren Zhao
QERA: an Analytical Framework for Quantization Error Reconstruction
The growing number of parameters and computational demands of large language models (LLMs) present significant challenges for their efficient deployment.
Recently, there is an increasing interest in quantizing weights to extremely low precision while offsetting the resulting error with low-rank, high-precision error reconstruction terms.
The combination of quantization and low-rank approximation is now popular in both adapter-based, parameter-efficient fine-tuning methods such as LoftQ and low-precision inference techniques including ZeroQuant-V2.
Usually, the low-rank terms are calculated via the singular value decomposition (SVD) of the weight quantization error,
minimizing the Frobenius and spectral norms of the weight approximation error.
Recent methods like LQ-LoRA and LQER introduced hand-crafted heuristics to minimize errors in layer outputs (activations) rather than weights, resulting improved quantization results.
However, these heuristic methods lack an analytical solution to guide the design of quantization error reconstruction terms.
In this paper, we revisit this problem and formulate an analytical framework, named Quantization Error Reconstruction Analysis (QERA),
and offer a closed-form solution to the problem.
We show QERA benefits both existing low-precision fine-tuning and inference methods --
QERA achieves a fine-tuned accuracy gain of Δ_{acc} = 6.05\% of 2-bit RoBERTa-base on GLUE compared to LoftQ;
and obtains Δ_{acc} = 2.97\% higher post-training quantization accuracy of 4-bit Llama-3.1-70B on average than ZeroQuant-V2 and Δ_{ppl} = $-$ 0.28 lower perplexity on WikiText2 than LQER.
****
Tianjin Huang,Ziquan Zhu,Gaojie Jin,Lu Liu,Zhangyang Wang,Shiwei Liu
SPAM: Spike-Aware Adam with Momentum Reset for Stable LLM Training
Large Language Models (LLMs) have demonstrated exceptional performance across di

verse tasks, yet their training remains highly resource intensive and susceptibl
e to critical challenges such as training instability. A predominant source of t
his instability stems from gradient and loss spikes, which disrupt the learning
process, often leading to costly interventions like checkpoint recovery and expe
riment restarts, further amplifying inefficiencies. This paper presents a compre
hensive investigation into gradient spikes observed during LLM training, reveali
ng their prevalence across multiple architectures and datasets. Our analysis sho
ws that these spikes can be up to 1000× larger than typical gradients, substanti
ally deteriorating model performance. To address this issue, we propose Spike-Aw
are Adam with Momentum Reset (SPAM), a novel optimizer designed to counteract gr
adient spikes through momentum reset and spike-aware gradient clipping. Extensiv
e experiments, including both pre-training and fine-tuning, demonstrate that SPA
M consistently surpasses Adam and its variants across a range of model scales. A
dditionally, SPAM facilitates memory-efficient training by enabling sparse momen
tum, where only a subset of momentum terms are maintained and updated. When oper
ating under memory constraints, SPAM outperforms state-of-the-art memory-efficie
nt optimizers such as GaLore and Adam-Mini. Our work underscores the importance
of mitigating gradient spikes in LLM training and introduces an effective optimi
zation strategy that enhances both training stability and resource efficiency at
 scale. Code is submitted.
****
Changdae Oh,Yixuan Li,Kyungwoo Song,Sangdoo Yun,Dongyoon Han
DaWin: Training-free Dynamic Weight Interpolation for Robust Adaptation
Adapting a pre-trained foundation model on downstream tasks should ensure robust
ness against distribution shifts without the need to retrain the whole model. Al
though existing weight interpolation methods are simple yet effective, we argue
their static nature limits downstream performance while achieving efficiency. In
 this work, we propose DaWin, a training-free dynamic weight interpolation metho
d that leverages the entropy of individual models over each unlabeled test sampl
e to assess model expertise, and compute per-sample interpolation coefficients d
ynamically. Unlike previous works that typically rely on additional training to
learn such coefficients, our approach requires no training. Then, we propose a m
ixture modeling approach that greatly reduces inference overhead raised by dynam
ic interpolation. We validate DaWin on the large-scale visual recognition benchm
arks, spanning 14 tasks across robust fine-tuning -- ImageNet and derived five d
istribution shift benchmarks -- and multi-task learning with eight classificatio
n tasks. Results demonstrate that DaWin achieves significant performance gain in
 considered settings, with minimal computational overhead. We further discuss Da
Win's analytic behavior to explain its empirical success.
****
Sepehr Dehdashtian,Gautam Sreekumar,Vishnu Boddeti
OASIS Uncovers: High-Quality T2I Models, Same Old Stereotypes
Images generated by text-to-image (T2I) models often exhibit visual biases and s
tereotypes of concepts such as culture and profession. Existing quantitative mea
sures of stereotypes are based on statistical parity that does not align with th
e sociological definition of stereotypes and, therefore, incorrectly categorizes
 biases as stereotypes. Instead of oversimplifying stereotypes as biases, we pro
pose a quantitative measure of stereotypes that aligns with its sociological def
inition. We then propose OASIS to measure the stereotypes in a generated dataset
 and understand their origins within the T2I model. OASIS includes two scores to
 measure stereotypes from a generated image dataset: **(M1)** Stereotype Score t
o measure the distributional violation of stereotypical attributes, and **(M2)**
 WALS to measure spectral variance in the images along a stereotypical attribute
. OASIS
also includes two methods to understand the origins of stereotypes in T2I models
: **(U1)** StOP to discover attributes that the T2I model internally associates
with a given concept, and **(U2)** SPI to quantify the emergence of stereotypica
l attributes in the latent space of the T2I model during image generation. Despi
te the considerable progress in image fidelity, using OASIS, we conclude that ne
wer T2I models such as FLUX.1 and SDv3 contain strong stereotypical predispositi

ons about concepts and still generate images with widespread stereotypical attri
butes. Additionally, the quantity of stereotypes worsens for nationalities with
lower Internet footprints.
****

Haobin Jiang,Jiangxing Wang,Zongqing Lu
Discrete Latent Plans via Semantic Skill Abstractions
Skill learning from language instructions is a critical challenge in developing
intelligent agents that can generalize across diverse tasks and follow complex h
uman instructions. Hierarchical methods address this by decomposing the learning
 problem into multiple levels, where the high-level and low-level policies are m
ediated through a latent plan space. Effective modeling and learning of this lat
ent plan space are key to enabling robust and interpretable skill learning. In t
his paper, we introduce LADS, a hierarchical approach that learns language-condi
tioned discrete latent plans through semantic skill abstractions. Our method dec
ouples the learning of the latent plan space from the language-conditioned high-
level policy to improve training stability. First, we incorporate a trajectory e
ncoder to learn a discrete latent space with the low-level policy, regularized b
y language instructions. Next, we model the high-level policy as a categorical d
istribution over these discrete latent plans to capture the multi-modality of th
e dataset. Through experiments in simulated control environments, we demonstrate
 that LADS outperforms state-of-the-art methods in both skill learning and compo
sitional generalization.
****

Tanqiu Jiang,Zian Wang,Jiacheng Liang,Changjiang Li,Yuhui Wang,Ting Wang
RobustKV: Defending Large Language Models against Jailbreak Attacks via KV Evict
ion
Jailbreak attacks circumvent LLMs' built-in safeguards by concealing harmful que
ries within adversarial prompts. While most existing defenses attempt to mitigat
e the effects of adversarial prompts, they often prove inadequate as adversarial
 prompts can take arbitrary, adaptive forms. This paper introduces RobustKV, a n
ovel jailbreak defense that takes a fundamentally different approach by selectiv
ely removing critical tokens of harmful queries from key-value (KV) caches. Intu
itively, for an adversarial prompt to be effective, its tokens must achieve suff
icient `importance' (measured by attention scores), which consequently lowers th
e importance of tokens in the concealed harmful query. Therefore, by carefully e
victing the KVs of low-ranked tokens, RobustKV minimizes the harmful query's pre
sence in the KV cache, thus preventing the LLM from generating informative respo
nses. Extensive evaluation using benchmark datasets and models demonstrates that
 RobustKV effectively counters state-of-the-art jailbreak attacks while maintain
ing the LLM's performance on benign queries. Notably, RobustKV creates an intere
sting effectiveness-evasiveness dilemma for the adversary, leading to its robust
ness against adaptive attacks.{(Warning: This paper contains potentially harmful
 content generated by LLMs.)}
****

Petr Kouba,Joan Planas-Iglesias,Jiri Damborsky,Jiri Sedlar,Stanislav Mazurenko,J
osef Sivic
Learning to engineer protein flexibility
Generative machine learning models are increasingly being used to design novel p
roteins. However, their major limitation is the inability to account for protein
 flexibility, a property crucial for protein function. Learning to engineer flex
ibility is difficult because the relevant data is scarce, heterogeneous, and cos
tly to obtain using computational and experimental methods. Our contributions ar
e three-fold. First, we perform a comprehensive comparison of methods for evalua
ting protein flexibility and identify relevant data for learning. Second, we ove
rcome the data scarcity issue by leveraging a pre-trained protein language model
. We design and train flexibility predictors utilizing either only sequential or
 both sequential and structural information on the input. Third, we introduce a
method for fine-tuning a protein inverse folding model to make it steerable towa
rd desired flexibility at specified regions. We demonstrate that our method Flex
pert enables guidance of inverse folding models toward increased flexibility. Th

is opens up a transformative possibility of engineering protein flexibility.
****
Ivan Rubachev,Nikolay Kartashev,Yury Gorishniy,Artem Babenko
TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchma rks
Advances in machine learning research drive progress in real-world applications.

To ensure this progress, it is important to understand the potential pitfalls on the way from a novel method's success on academic benchmarks to its practical d eployment. In this work, we analyze existing tabular deep learning benchmarks an d find two common characteristics of tabular data in typical industrial applicat ions that are underrepresented in the datasets usually used for evaluation in th e literature.
First, in real-world deployment scenarios, distribution of data often changes ov er time. To account for this distribution drift, time-based train/test splits sh ould be used in evaluation. However, existing academic tabular datasets often la ck timestamp metadata to enable such evaluation.
Second, a considerable portion of datasets in production settings stem from exte nsive data acquisition and feature engineering pipelines. This can have an impac t on the absolute and relative number of predictive, uninformative, and correlat ed features compared to academic datasets.
In this work, we aim to understand how recent research advances in tabular deep learning transfer to these underrepresented conditions.
To this end, we introduce TabReD -- a collection of eight industry-grade tabular datasets.
We reassess a large number of tabular ML models and techniques on TabReD. We dem onstrate that evaluation on both time-based data splits and richer feature sets leads to different methods ranking, compared to evaluation on random splits and smaller number of features, which are common in academic benchmarks. Furthermore , simple MLP-like architectures and GBDT show the best results on the TabReD dat asets, while other methods are less effective in the new setting.
****
Zhuo Ouyang,Kaiwen Hu,Qi Zhang,Yifei Wang,Yisen Wang
Projection Head is Secretly an Information Bottleneck
Recently, contrastive learning has risen to be a promising paradigm for extracti ng meaningful data representations. Among various special designs, adding a proj ection head on top of the encoder during training and removing it for downstream tasks has proven to significantly enhance the performance of contrastive learni ng. However, despite its empirical success, the underlying mechanism of the proj ection head remains under-explored. In this paper, we develop an in-depth theore tical understanding of the projection head from the information-theoretic perspe ctive. By establishing the theoretical guarantees on the downstream performance of the features before the projector, we reveal that an effective projector shou ld act as an information bottleneck, filtering out the information irrelevant to the contrastive objective. Based on theoretical insights, we introduce modifica tions to projectors with training and structural regularizations. Empirically, o ur methods exhibit consistent improvement in the downstream performance across v arious real-world datasets, including CIFAR-10, CIFAR-100, and ImageNet-100. We believe our theoretical understanding on the role of the projection head will in spire more principled and advanced designs in this field. Code is available at \url{https://github.com/PKU-ML/Projector_Theory}.
****
Cory Daniel Efird,Alex Murphy,Joel Zylberberg,Alona Fyshe
Finding Shared Decodable Concepts and their Negations in the Brain
Prior work has offered evidence for functional localization in the brain; differ ent anatomical regions preferentially activate for certain types of visual input . For example, the fusiform face area preferentially activates for visual stimul i that include a face. However, the spectrum of visual semantics is extensive, a nd only a few semantically-tuned patches of cortex have so far been identified i n the human brain. Using a multimodal (natural language and image) neural networ

k architecture (CLIP, \cite{CLIP}, we train a highly accurate contrastive model that maps brain responses during naturalistic image viewing to CLIP embeddings. We then use a novel adaptation of the DBSCAN clustering algorithm to cluster the parameters of these participant-specific contrastive models. This reveals what we call Shared Decodable Concepts (SDCs): clusters in CLIP space that are decodable from common sets of voxels across multiple participants.

Examining the images most and least associated with each SDC cluster gives us additional insight into the semantic properties of each SDC. We note SDCs for previously reported visual features (e.g. orientation tuning in early visual cortex) as well as visual semantic concepts such as faces, places and bodies. In cases where our method finds multiple clusters for a visuo-semantic concept, the least associated images allow us to dissociate between confounding factors. For example, we discovered two clusters of food images, one driven by color, the other by shape. We also uncover previously unreported areas with visuo-semantic sensitivity such as regions of extrastriate body area (EBA) tuned for legs/hands and sensitivity to numerosity in right intraparietal sulcus, sensitivity associated with visual perspective (close/far) and more. Thus, our contrastive-learning methodology better characterizes new and existing visuo-semantic representations in the brain by leveraging multimodal neural network representations and a novel adaptation of clustering algorithms.
****
Nazanin Mohammadi Sepahvand,Eleni Triantafillou,Hugo Larochelle,Doina Precup,James J. Clark,Daniel M. Roy,Gintare Karolina Dziugaite
Selective Unlearning via Representation Erasure Using Domain Adversarial Training
When deploying machine learning models in the real world, we often face the challenge of “unlearning” specific data points or subsets after training. Inspired by Domain-Adversarial Training of Neural Networks (DANN), we propose a novel algorithm,SURE, for targeted unlearning.SURE treats the process as a domain adaptation problem, where the “forget set” (data to be removed) and a validation set from the same distribution form two distinct domains. We train a domain classifier to discriminate between representations from the forget and validation sets.Using a gradient reversal strategy similar to DANN, we perform gradient updates to the representations to “fool” the domain classifier and thus obfuscate representations belonging to the forget set. Simultaneously, gradient descent is applied to the retain set (original training data minus the forget set) to preserve its classification performance. Unlike other unlearning approaches whose training objectives are built based on model outputs, SURE directly manipulates the representations.This is key to ensure robustness against a set of more powerful attacks than currently considered in the literature, that aim to detect which examples were unlearned through access to learned embeddings. Our thorough experiments reveal that SURE has a better unlearning quality to utility trade-off compared to other standard unlearning techniques for deep neural networks.
****
HyoJung Han,Akiko Eriguchi,Haoran Xu,Hieu Hoang,Marine Carpuat,Huda Khayrallah
Adapters for Altering LLM Vocabularies: What Languages Benefit the Most?
Vocabulary adaptation, which integrates new vocabulary into pre-trained language models, enables expansion to new languages and mitigates token over-fragmentation. However, existing approaches are limited by their reliance on heuristics or external embeddings. We propose VocADT, a novel method for vocabulary adaptation using adapter modules that are trained to learn the optimal linear combination of existing embeddings while keeping the model’s weights fixed. VocADT offers a flexible and scalable solution without depending on external resources or language constraints. Across 11 languages—with diverse scripts, resource availability, and fragmentation—we demonstrate that VocADT outperforms the original Mistral model (Jiang et al., 2023) and other baselines across various multilingual tasks including natural language understanding and machine translation. We find that Latin-script languages and highly fragmented languages
benefit the most from vocabulary adaptation. We further fine-tune the adapted mo

del on the generative task of machine translation and find that vocabulary adapt
ation is still beneficial after fine-tuning and that VocADT is the most effectiv
e.
****

Daniel Cederberg,Xuyang Wu,Stephen P. Boyd,Mikael Johansson
An Asynchronous Bundle Method for Distributed Learning Problems
We propose a novel asynchronous bundle method to solve distributed learning prob
lems. Compared to existing asynchronous methods, our algorithm computes the next
 iterate based on a more accurate approximation of the objective function and do
es not require any prior information about the maximal information delay in the
 system. This makes the proposed method fast and easy to tune. We prove that the
 algorithm converges in both deterministic and stochastic (mini-batch) settings,
 and quantify how the convergence times depend on the level of asynchrony. The p
ractical advantages of our method are illustrated through numerical experiments
on classification problems of varying complexities and scales.
****

So Kuroki,Taishi Nakamura,Takuya Akiba,Yujin Tang
Agent Skill Acquisition for Large Language Models via CycleQD
Training large language models to acquire specific skills remains a challenging
endeavor. Conventional training approaches often struggle with data distribution
 imbalances and inadequacies in objective functions that do not align well with
task-specific performance. To address these challenges, we introduce CycleQD, a
novel approach that leverages the Quality Diversity framework through a cyclic a
daptation of the algorithm, along with a model merging based crossover and an SV
D-based mutation. In CycleQD, each task’s performance metric is alternated as th
e quality measure while the others serve as the behavioral characteristics. This
 cyclic focus on individual tasks allows for concentrated effort on one task at
a time, eliminating the need for data ratio tuning and simplifying the design of
 the objective function. Empirical results from AgentBench indicate that applyin
g CycleQD to LLAMA3-8B-INSTRUCT based models not only enables them to surpass tr
aditional fine-tuning methods in coding, operating systems, and database tasks,
but also achieves performance on par with GPT-3.5-TURBO, which potentially conta
ins much more parameters, across these domains. Crucially, this enhanced perform
ance is achieved while retaining robust language capabilities, as evidenced by i
ts performance on widely adopted language benchmark tasks. We highlight the key
design choices in CycleQD, detailing how these contribute to its effectiveness.
Furthermore, our method is general and can be applied to image segmentation mode
ls, highlighting its applicability across different domains.
****

Itsumi Saito,Haruto Yoshida,Keisuke Sakaguchi
Sketch2Diagram: Generating Vector Diagrams from Hand-Drawn Sketches
We address the challenge of automatically generating high-quality vector diagram
s from hand-drawn sketches.
Vector diagrams are essential for communicating complex ideas across various fie
lds, offering flexibility and scalability.
While recent research has progressed in generating diagrams from text descriptio
ns, converting hand-drawn sketches into vector diagrams remains largely unexplor
ed due to the lack of suitable datasets.
To address this gap, we introduce SketikZ, a dataset comprising 3,231 pairs of h
and-drawn sketches and thier corresponding TikZ codes as well as reference diagr
ams.
Our evaluations reveal the limitations of state-of-the-art vision and language m
odels (VLMs), positioning SketikZ as a key benchmark for future research in sket
ch-to-diagram conversion.
Along with SketikZ, we present ImgTikZ, an image-to-TikZ model that integrates a
 6.7B parameter code-specialized open-source large language model (LLM) with a p
re-trained vision encoder.
Despite its relatively compact size, ImgTikZ performs comparably to GPT-4o.
This success is driven by using our two data augmentation techniques and a multi
-candidate inference strategy.

Our findings open promising directions for future research in sketch-to-diagram conversion and broader image-to-code generation tasks. SketikZ is publicly available.
****
Koichi Saito,Dongjun Kim,Takashi Shibuya,Chieh-Hsin Lai,Zhi Zhong,Yuhta Takida,Yuki Mitsufuji
SoundCTM: Unifying Score-based and Consistency Models for Full-band Text-to-Sound Generation
Sound content creation, essential for multimedia works such as video games and films, often involves extensive trial-and-error, enabling creators to semantically reflect their artistic ideas and inspirations, which evolve throughout the creation process, into the sound.
Recent high-quality diffusion-based Text-to-Sound (T2S) generative models provide valuable tools for creators. However, these models often suffer from slow inference speeds, imposing an undesirable burden that hinders the trial-and-error process.
While existing T2S distillation models address this limitation through 1-step generation, the sample quality of 1-step generation remains insufficient for production use.
Additionally, while multi-step sampling in those distillation models improves sample quality itself, the semantic content changes due to their lack of deterministic sampling capabilities.
Thus, developing a T2S generative model that allows creators to efficiently conduct trial-and-error while producing high-quality sound remains a key challenge.
To address these issues, we introduce Sound Consistency Trajectory Models (SoundCTM), which allow flexible transitions between high-quality 1-step sound generation and superior sound quality through multi-step deterministic sampling.
This allows creators to efficiently conduct trial-and-error with 1-step generation to semantically align samples with their intention, and subsequently refine sample quality with preserving semantic content through deterministic multi-step sampling.
To develop SoundCTM, we reframe the CTM training framework, originally proposed in computer vision, and introduce a novel feature distance using the teacher network for a distillation loss.
Additionally, while distilling classifier-free guided trajectories, we introduce a ν-sampling, a new algorithm that offers another source of quality improvement. For the ν-sampling, we simultaneously train both conditional and unconditional student models.
For production-level generation, we scale up our model to 1B trainable parameters, making SoundCTM-DiT-1B the first large-scale distillation model in the sound community to achieve both promising high-quality 1-step and multi-step full-band (44.1kHz) generation.
Audio samples are available at \url{https://anonymus-soundctm.github.io/soundctm_iclr/}.
****
Amit Chakrabarti,Jeffrey Jiang,David Woodruff,Taisuke Yasuda
Streaming Algorithms For ℓ_p Flows and ℓ_p Regression
We initiate the study of one-pass streaming algorithms for underdetermined ℓ_p linear regression problems of the form
  $$
      \min_{\mathbf A\mathbf x = \mathbf b} \lVert\mathbf x\rVert_p \,, \qquad
      \text{where } \mathbf A \in \mathbb R^{n \times d} \text{ with } n \ll d \,,
  $$
  which generalizes basis pursuit ($p = 1$) and least squares solutions to
  underdetermined linear systems ($p = 2$). We study the column-arrival
  streaming model, in which the columns of $\mathbf A$ are presented one by one in a
  stream. When $\mathbf A$ is the incidence matrix of a graph, this corresponds to an

edge insertion graph stream, and the regression problem captures ℓ_p flows which includes transshipment ($p = 1$), electrical flows ($p = 2$), and max flow ($p = \infty$) on undirected graphs as special cases. Our goal is to design algorithms which use space much less than the entire stream, which has a length of d.

For the task of estimating the cost of the ℓ_p regression problem for $p\in[2,\infty]$, we show a streaming algorithm which constructs a sparse instance supported on $\tilde O(\varepsilon^{-2}n)$ columns of $\mathbf A$ which approximates the cost up to a $(1\pm\varepsilon)$ factor, which corresponds to $\tilde O(\varepsilon^{-2}n^2)$ bits of space in general and an $\tilde O(\varepsilon^{-2}n)$ space semi-streaming algorithm for constructing ℓ_p flow sparsifiers on graphs. This extends to $p\in(1, 2)$ with $\tilde O(\varepsilon^{2}n^{q/2})$ columns, where q is the H\"older conjugate exponent of p. For $p = 2$, we show that $\Omega(n^2)$ bits of space are required in general even for outputting a constant factor solution. For $p = 1$, we show that the cost cannot be estimated even to an $o(\sqrt n)$ factor in $\mathrm{poly}(n)$ space.

On the other hand, if we are interested in outputting a solution $\mathbf x$, then we show that $(1+\varepsilon)$-approximations require $\Omega(d)$ space for $p > 1$, and in general, κ-approximations require $\tilde\Omega(d/\kappa^{2q})$ space for $p > 1$. We complement these lower bounds with the first sublinear space upper bounds for this problem, showing that we can output a κ-approximation using space only $\mathrm{poly}(n) \cdot \tilde O(d/\kappa^q)$ for $p > 1$, as well as a $\sqrt n$-approximation using $\mathrm{poly}(n, \log d)$ space for $p = 1$.
****
Kairong Luo,Haodong Wen,Shengding Hu,Zhenbo Sun,Maosong Sun,Zhiyuan Liu,Kaifeng Lyu,Wenguang Chen
A Multi-Power Law for Loss Curve Prediction Across Learning Rate Schedules
Training large models is both resource-intensive and time-consuming, making it crucial to understand the quantitative relationship between model performance and hyperparameters. In this paper, we derive an empirical law that predicts pretraining loss for large language models for every intermediate training step across various learning rate schedules, including constant, cosine, and step decay schedules. Our proposed law takes a multi-power form, combining a power law based on the sum of learning rates and additional power laws to account for a loss reduction effect as learning rate decays. We validate this law extensively on Llama-2 models of varying sizes and demonstrate that, after fitting on a few learning rate schedules, it accurately predicts the loss curves for unseen schedules of different shapes and horizons. Moreover, by minimizing the predicted final pretraining loss across learning rate schedules, we are able to find a schedule that outperforms the widely-used cosine learning rate schedule. Interestingly, this automatically discovered schedule bears some resemblance to the recently proposed Warmup-Stable-Decay (WSD) schedule (Hu et al, 2024) but achieves a slightly lower final loss. We believe these results could offer valuable insights for understanding the dynamics of pretraining and for designing learning rate schedules to improve efficiency.
****
Jin Zhang,Flood Sung,Zhilin Yang,Yang Gao,Chongjie Zhang
Learning to Plan Before Answering: Self-Teaching LLMs to Learn Abstract Plans for Problem Solving
In the field of large language model (LLM) post-training, the effectiveness of utilizing synthetic data generated by the LLM itself has been well-presented. However, a key question remains unaddressed: what essential information should such self-generated data encapsulate? Existing approaches only produce step-by-step problem solutions, and fail to capture the abstract meta-knowledge necessary for generalization across similar problems. Drawing insights from cognitive science, where humans employ high-level abstraction to simplify complex problems before

delving into specifics, we introduce a novel self-training algorithm: LEarning to Plan before Answering (LEPA). LEPA trains the LLM to formulate anticipatory plans, which serve as abstract meta-knowledge for problem-solving, before engaging with the intricacies of problems. This approach not only outlines the solution generation path but also shields the LLM from the distraction of irrelevant details. During data generation, LEPA first crafts an anticipatory plan based on the problem, and then generates a solution that aligns with both the plan and the problem. LEPA refines the plan through self-reflection, aiming to acquire plans that are instrumental in yielding correct solutions. During model optimization, the LLM is trained to predict both the refined plans and the corresponding solutions. By efficiently extracting and utilizing the anticipatory plans, LEPA demonstrates remarkable superiority over conventional algorithms on various challenging natural language reasoning benchmarks.
****
Hanxun Huang,Sarah Monazam Erfani,Yige Li,Xingjun Ma,James Bailey
Detecting Backdoor Samples in Contrastive Language Image Pretraining
Contrastive language-image pretraining (CLIP) has been found to be vulnerable to poisoning backdoor attacks where the adversary can achieve an almost perfect attack success rate on CLIP models by poisoning only 0.01\% of the training dataset. This raises security concerns on the current practice of pretraining large-scale models on unscrutinized web data using CLIP. In this work, we analyze the representations of backdoor-poisoned samples learned by CLIP models and find that they exhibit unique characteristics in their local subspace, i.e., their local neighborhoods are far more sparse than that of clean samples. Based on this finding, we conduct a systematic study on detecting CLIP backdoor attacks and show that these attacks can be easily and efficiently detected by traditional density ratio-based local outlier detectors, whereas existing backdoor sample detection methods fail. Our experiments also reveal that an unintentional backdoor already exists in the original CC3M dataset and has been trained into a popular open-source model released by OpenCLIP. Based on our detector, one can clean up a million-scale web dataset (e.g., CC3M) efficiently within 15 minutes using 4 Nvidia A100 GPUs.
****
Nicholas Matthew Boffi,Arthur Jacot,Stephen Tu,Ingvar Ziemann
Shallow diffusion networks provably learn hidden low-dimensional structure
Diffusion-based generative models provide a powerful framework for learning to sample from a complex target distribution. The remarkable empirical success of these models applied to high-dimensional signals, including images and video, stands in stark contrast to classical results highlighting the curse of dimensionality for distribution recovery. In this work, we take a step towards understanding this gap through a careful analysis of learning diffusion models over the Barron space of single hidden layer neural networks. In particular, we show that these shallow models provably adapt to simple forms of low-dimensional structure, such as an unknown linear subspace or hidden independence, thereby avoiding the curse of dimensionality. We combine our results with recent analyses of sampling with diffusions to provide an end-to-end sample complexity bound for learning to sample from structured distributions. Importantly, our results do not require specialized architectures tailored to particular latent structures, and instead rely on the low-index structure of the Barron space to adapt to the underlying distribution.
****
Boaz Carmeli,Ron Meir,Yonatan Belinkov
CtD: Composition through Decomposition in Emergent Communication
Compositionality is a cognitive mechanism that allows humans to systematically combine known concepts in novel ways.
This study demonstrates how artificial neural agents acquire and utilize compositional generalization to describe previously unseen images.
Our method, termed \`\`Composition through Decomposition'', involves two sequential training steps.
In the \'Decompose\' step, the agents learn to decompose an image into basic con

cepts using a codebook acquired during interaction in a multi-target coordinatio
n game.
Subsequently, in the \`Compose\' step, the agents employ this codebook to descri
be novel images by composing basic concepts into complex phrases.
Remarkably, we observe cases where generalization in the `Compose' step is achie
ved zero-shot, without the need for additional training.
****

Mahtab Sarvmaili,Hassan Sajjad,Ga Wu
Data-centric Prediction Explanation via Kernelized Stein Discrepancy
Existing example-based prediction explanation methods often bridge test and trai
ning data points through the model’s parameters or latent representations. While
 these methods offer clues to the causes of model predictions, they often exhibi
t innate shortcomings, such as incurring significant computational overhead or p
roducing coarse-grained explanations. This paper presents a Highly-precise and D
ata-centric Explanation (HD-Explain) prediction explanation method that exploits
 properties of Kernelized Stein Discrepancy (KSD). Specifically, the KSD uniquel
y defines a parameterized kernel function for a trained model that encodes model
-dependent data correlation. By leveraging the kernel function, one can identify
 training samples that provide the best predictive support to a test point effic
iently. We conducted thorough analyses and experiments across multiple classific
ation domains, where we show that HD-Explain outperforms existing methods from v
arious aspects, including 1) preciseness (fine-grained explanation), 2) consiste
ncy, and 3) computation efficiency, leading to a surprisingly simple, effective,
 and robust prediction explanation solution.
****

Enrico Ventura,Beatrice Achilli,Gianluigi Silvestri,Carlo Lucibello,Luca Ambrogi
oni
Manifolds, Random Matrices and Spectral Gaps: The geometric phases of generative
 diffusion
In this paper, we investigate the latent geometry of generative diffusion models
 under the manifold hypothesis. For this purpose, we analyze the spectrum of eig
envalues (and singular values) of the Jacobian of the score function, whose disc
ontinuities (gaps) reveal the presence and dimensionality of distinct sub-manifo
lds. Using a statistical physics approach, we derive the spectral distributions
and formulas for the spectral gaps under several distributional assumptions, and
 we compare these theoretical predictions with the spectra estimated from traine
d networks. Our analysis reveals the existence of three distinct qualitative pha
ses during the generative process: a trivial phase; a manifold coverage phase wh
ere the diffusion process fits the distribution internal to the manifold; a cons
olidation phase where the score becomes orthogonal to the manifold and all parti
cles are projected on the support of the data. This `division of labor' between
different timescales provides an elegant explanation of why generative diffusion
 models are not affected by the manifold overfitting phenomenon that plagues lik
elihood-based models, since the internal distribution and the manifold geometry
are produced at different time points during generation.
****

Shuguang Yu,Wenqian Xu,Xinyi Zhou,Xuechun Wang,Hongtu Zhu,Fan Zhou
Enhancing Prediction Performance through Influence Measure
In the field of machine learning, the pursuit of accurate models is ongoing. A k
ey aspect of improving prediction performance lies in identifying which data poi
nts in the training set should be excluded and which high-quality, potentially u
nlabeled data points outside the training set should be incorporated to improve
the model's performance on unseen data. To accomplish this, an effective metric
is needed to evaluate the contribution of each data point toward enhancing overa
ll model performance.
This paper proposes the use of an influence measure as a metric to assess the im
pact of training data on test set performance. Additionally, we introduce a data
 selection method to optimize the training set as well as a dynamic active learn
ing algorithm driven by the influence measure. The effectiveness of these method
s is demonstrated through extensive simulations and real-world datasets.

****
Shengju Yu,Zhibin Dong,Siwei Wang,Pei Zhang,Yi Zhang,Xinwang Liu,Naiyang Guan,Tiejun Li,Yiu-ming Cheung
Simple yet Effective Incomplete Multi-view Clustering: Similarity-level Imputation and Intra-view Hybrid-group Prototype Construction
Most of incomplete multi-view clustering (IMVC) methods typically choose to ignore the missing samples and only utilize observed unpaired samples to construct bipartite similarity. Moreover, they employ a single quantity of prototypes to extract the information of $\textbf{all}$ views. To eliminate these drawbacks, we present a simple yet effective IMVC approach, SIIHPC, in this work. It firstly transforms partial bipartition learning into original sample form by virtue of reconstruction concept to split out of observed similarity, and then loosens traditional non-negative constraints via regularizing samples to more freely characterize the similarity. Subsequently,
it learns to recover the incomplete parts by utilizing the connection built between the similarity exclusive on respective view and the consensus graph shared for all views. On this foundation, it further introduces a group of hybrid prototype quantities for each individual view to flexibly extract the data features belonging to each view itself. Accordingly, the resulting graphs are with various scales and describe the overall similarity more comprehensively. It is worth mentioning that these all are optimized in one unified learning framework, which makes it possible for them to reciprocally promote. Then, to effectively solve the formulated optimization problem, we design an ingenious auxiliary function that is with theoretically proven monotonic-increasing properties. Finally, the clustering results are obtained by implementing spectral grouping action on the eigenvectors of stacked multi-scale consensus similarity. Experimental results confirm the effectiveness of SIIHPC.
****
Krunoslav Lehman Pavasovic,Giulio Biroli,Levent Sagun
A Differentiable Rank-Based Objective for Better Feature Learning
In this paper, we leverage existing statistical methods to better understand feature learning from data. We tackle this by modifying the model-free variable selection method, Feature Ordering by Conditional Independence (FOCI), which is introduced in Azadkia & Chatterjee (2021). While FOCI is based on a non-parametric coefficient of conditional dependence, we introduce its parametric, differentiable approximation. With this approximate coefficient of correlation, we present a new algorithm called difFOCI, which is applicable to a wider range of machine learning problems thanks to its differentiable nature and learnable parameters. We present difFOCI in three contexts: (1) as a variable selection method with baseline comparisons to FOCI, (2) as a trainable model parametrized with a neural network, and (3) as a generic, widely applicable neural network regularizer, one that improves feature learning with better management of spurious correlations. We evaluate difFOCI on increasingly complex problems ranging from basic variable selection in toy examples to saliency map comparisons in convolutional networks. We then show how difFOCI can be incorporated in the context of fairness to facilitate classifications without relying on sensitive data.
****
Thomas P Cannon,Özgür ■im■ek
Accelerating Task Generalisation with Multi-Level Skill Hierarchies
Developing reinforcement learning agents that can generalise effectively to new tasks is one of the main challenges in AI research. This paper introduces Fracture Cluster Options (FraCOs), a multi-level hierarchical reinforcement learning method designed to improve generalisation performance. FraCOs identifies patterns in agent behaviour and forms temporally-extended actions (options) based on the expected future usefulness of those patterns, enabling rapid adaptation to new tasks. In tabular settings, FraCOs demonstrates effective transfer and improves performance as the depth of the hierarchy increases. In several complex procedurally-generated environments, FraCOs consistently outperforms state-of-the-art deep reinforcement learning algorithms, achieving superior results in both in-distribution and out-of-distribution scenarios.

****
Filip Matzner,František Mráz
Locally Connected Echo State Networks for Time Series Forecasting
Echo State Networks (ESNs) are a class of recurrent neural networks in which only a small readout regression layer is trained, while the weights of the recurrent network, termed the reservoir, are randomly assigned and remain fixed. Our work introduces the Locally Connected ESN (LCESN), a novel ESN variant with a locally connected reservoir, forced memory, and a weight adaptation strategy. LCESN significantly reduces the asymptotic time and space complexities compared to the conventional ESN, enabling substantially larger networks. LCESN also improves the memory properties of ESNs without affecting network stability. We evaluate LCESN's performance on the NARMA10 benchmark task and compare it to state-of-the-art models on nine real-world datasets. Despite the simplicity of our model and its one-shot training approach, LCESN achieves competitive results, even surpassing several state-of-the-art models. LCESN introduces a fresh approach to real-world time series forecasting and demonstrates that large, well-tuned random recurrent networks can rival complex gradient-trained feedforward models. We provide our GPU-based implementation of LCESN as an open-source library.
****
Huimin Zeng,Xiaojie Wang,Anoop Jain,Zhicheng Dou,Dong Wang
A Non-Contrastive Learning Framework for Sequential Recommendation with Preference-Preserving Profile Generation
Contrastive Learning (CL) proves to be effective for learning generalizable user representations in Sequential Recommendation (SR), but it suffers from high computational costs due to its reliance on negative samples. To overcome this limitation, we propose the first Non-Contrastive Learning (NCL) framework for SR, which eliminates computational overhead of identifying and generating negative samples. However, without negative samples, it is challenging to learn uniform representations from only positive samples, which is prone to representation collapse. Furthermore, the alignment of the learned representations may be substantially compromised because existing ad-hoc augmentations can produce positive samples that have inconsistent user preferences. To tackle these challenges, we design a novel preference-preserving profile generation method to produce high-quality positive samples for non-contrastive training. Inspired by differential privacy, our approach creates augmented user profiles that exhibit high diversity while provably retaining consistent user preferences. With larger diversity and consistency of the positive samples, our NCL framework significantly enhances the alignment and uniformity of the learned representations, which contributes to better generalization. The experimental results on various benchmark datasets and model architectures demonstrate the effectiveness of the proposed method. Finally, our investigations reveal that both uniformity and alignment play a vital role in improving generalization for SR. Interestingly, in our data-sparse setting, alignment is usually more important than uniformity.
****
Wei-Hsiang Yu,Yen-Yu Lin,Ming-Hsuan Yang,Yi-Hsuan Tsai
Ranking-aware adapter for text-driven image ordering with CLIP
Recent advances in vision-language models (VLMs) have made significant progress in downstream tasks that require quantitative concepts such as facial age estimation and image quality assessment, enabling VLMs to explore applications like image ranking and retrieval. However, existing studies typically focus on the reasoning based on a single image and heavily depend on text prompting, limiting their ability to learn comprehensive understanding from multiple images. To address this, we propose an effective yet efficient approach that reframes the CLIP model into a learning-to-rank task and introduces a lightweight adapter to augment CLIP for text-guided image ranking. Specifically, our approach incorporates learnable prompts to adapt to new instructions for ranking purposes and an auxiliary branch with ranking-aware attention, leveraging text-conditioned visual differences for additional supervision in image ranking. Our ranking-aware adapter consistently outperforms fine-tuned CLIPs on various tasks and achieves competitive results compared to state-of-the-art models designed for specific tasks like fac

ial age estimation and image quality assessment. Overall, our approach primarily focuses on ranking images with a single instruction, which provides a natural and generalized way of learning from visual differences across images, bypassing the need for extensive text prompts tailored to individual tasks.
****
Yumeng Li,William H. Beluch,Margret Keuper,Dan Zhang,Anna Khoreva
VSTAR: Generative Temporal Nursing for Longer Dynamic Video Synthesis
Despite tremendous progress in the field of text-to-video (T2V) synthesis, open-sourced T2V diffusion models struggle to generate longer videos with dynamically varying and evolving content. They tend to synthesize quasi-static videos, ignoring the necessary visual change-over-time implied in the text prompt. Meanwhile, scaling these models to enable longer, more dynamic video synthesis often remains computationally intractable. To tackle this challenge, we introduce the concept of Generative Temporal Nursing (GTN), where we adjust the generative process on the fly during inference to improve control over the temporal dynamics and enable generation of longer videos. We propose a method for GTN, dubbed VSTAR, which consists of two key ingredients: Video Synopsis Prompting (VSP) and Temporal Attention Regularization (TAR), the latter being our core contribution. Based on a systematic analysis, we discover that the temporal units in pretrained T2V models are crucial to control the video dynamics. Upon this finding, we propose a novel regularization technique to refine the temporal attention, enabling training-free longer video synthesis in a single inference pass. For prompts involving visual progression, we leverage LLMs to generate video synopsis - description of key visual states - based on the original prompt to provide better guidance along the temporal axis. We experimentally showcase the superiority of our method in synthesizing longer, visually appealing videos over open-sourced T2V models.
****
Wen Yang,Junhong Wu,Chen Wang,Chengqing Zong,Jiajun Zhang
Language Imbalance Driven Rewarding for Multilingual Self-improving
Large Language Models (LLMs) have achieved state-of-the-art performance across numerous tasks. However, these advancements have predominantly benefited "first-class" languages such as English and Chinese, leaving many other languages underrepresented. This imbalance, while limiting broader applications, generates a natural preference ranking between languages, offering an opportunity to bootstrap the multilingual capabilities of LLM in a self-improving manner. Thus, we propose $\textit{Language Imbalance Driven Rewarding}$, where the inherent imbalance between dominant and non-dominant languages within LLMs is leveraged as a reward signal. Iterative DPO training demonstrates that this approach not only enhances LLM performance in non-dominant languages but also improves the dominant language's capacity, thereby yielding an iterative reward signal. Fine-tuning Meta-Llama-3-8B-Instruct over two iterations of this approach results in continuous improvements in multilingual performance across instruction-following and arithmetic reasoning tasks, evidenced by an average improvement of 7.46\% win rate on the X-AlpacaEval leaderboard and 13.9\% accuracy on the MGSM benchmark. This work serves as an initial exploration, paving the way for multilingual self-improvement of LLMs.
****
Sueda Taner,Ziyi Wang,Christoph Studer
Cauchy-Schwarz Regularizers
We introduce a novel class of regularization functions, called Cauchy–Schwarz (CS) regularizers, which can be designed to induce a wide range of properties in solution vectors of optimization problems. To demonstrate the versatility of CS regularizers, we derive regularization functions that promote discrete-valued vectors, eigenvectors of a given matrix, and orthogonal matrices. The resulting CS regularizers are simple, differentiable, and can be free of spurious stationary points, making them suitable for gradient-based solvers and large-scale optimization problems. In addition, CS regularizers automatically adapt to the appropriate scale, which is, for example, beneficial when discretizing the weights of neural networks. To demonstrate the efficacy of CS regularizers, we provide results for solving underdetermined systems of linear equations and weight quantization

in neural networks. Furthermore, we discuss specializations, variations, and ge
neralizations, which lead to an even broader class of new and possibly more powe
rful regularizers.
****
Shangyu Chen,Zizheng Pan,Jianfei Cai,Dinh Phung
PaRa: Personalizing Text-to-Image Diffusion via Parameter Rank Reduction
Personalizing a large-scale pretrained Text-to-Image (T2I) diffusion model is ch
al-
lenging as it typically struggles to make an appropriate trade-off between its t
raining
data distribution and the target distribution, i.e., learning a novel concept wi
th only a
few target images to achieve personalization (aligning with the personalized tar
get)
while preserving text editability (aligning with diverse text prompts). In this
paper,
we propose PaRa, an effective and efficient Parameter Rank Reduction approach
for T2I model personalization by explicitly controlling the rank of the diffusio
n
model parameters to restrict its initial diverse generation space into a small a
nd
well-balanced target space. Our design is motivated by the fact that taming a T2
I
model toward a novel concept such as a specific art style implies a small genera
tion
space. To this end, by reducing the rank of model parameters during finetuning,
we
can effectively constrain the space of the denoising sampling trajectories towar
ds
the target. With comprehensive experiments, we show that PaRa achieves great
advantages over existing finetuning approaches on single/multi-subject generatio
n
as well as single-image editing. Notably, compared to the prevailing fine-tuning
technique LoRA, PaRa achieves better parameter efficiency (2× fewer learnable
parameters) and much better target image alignment.
****
Yanzhao Qin,Tao Zhang,Tao Zhang,Yanjun Shen,Wenjing Luo,sunhaoze,Yan Zhang,Yujin
g Qiao,weipeng chen,Zenan Zhou,Wentao Zhang,Bin CUI
SysBench: Can LLMs Follow System Message?
Large Language Models (LLMs) have become instrumental across various application
s, with the customization of these models to specific scenarios becoming increas
ingly critical. System message, a fundamental component of LLMs, is consist of c
arefully crafted instructions that guide the behavior of model to meet intended
goals. Despite the recognized potential of system messages to optimize AI-driven
 solutions, there is a notable absence of a comprehensive benchmark for evaluat
ing how well LLMs follow system messages. To fill this gap, we introduce SysBenc
h, a benchmark that systematically analyzes system message following ability in
terms of three limitations of existing LLMs: constraint violation, instruction m
isjudgement and multi-turn instability. Specifically, we manually construct eval
uation dataset based on six prevalent types of constraints, including 500 tailor
-designed system messages and multi-turn user conversations covering various int
eraction relationships. Additionally, we develop a comprehensive evaluation prot
ocol to measure model performance. Finally, we conduct extensive evaluation acro
ss various existing LLMs, measuring their ability to follow specified constraint
s given in system messages. The results highlight both the strengths and weaknes
ses of existing models, offering key insights and directions for future research
.
****
Jizhou Huang,Brendan Juba
Distribution-Specific Agnostic Conditional Classification With Halfspaces

We study "selective" or "conditional" classification problems under an agnostic setting. Classification tasks commonly focus on modeling the relationship betwee n features and categories that captures the vast majority of data. In contrast t o common machine learning frameworks, conditional classification intends to mode l such relationships only on a subset of the data defined by some selection rule . Most work on conditional classification either solves the problem in a realiza ble setting or does not guarantee the error is bounded compared to an optimal so lution. In this work, we consider selective/conditional classification by sparse linear classifiers for subsets defined by halfspaces, and give both positive as well as negative results for Gaussian feature distributions. On the positive si de, we present the first PAC-learning algorithm for homogeneous halfspace select ors with error guarantee $\tilde{O}(\sqrt{\mathrm{opt}})$, where opt is the smallest conditional classification error over the given class of classif iers and homogeneous halfspaces. On the negative side, we find that, under crypt ographic assumptions, approximating the conditional classification loss within a small additive error is computationally hard even under Gaussian distribution. We prove that approximating conditional classification is at least as hard as ap proximating agnostic classification in both additive and multiplicative form.
****

Dingdong Yang,Yizhi Wang,Konrad Schindler,Ali Mahdavi Amiri,Hao Zhang
GALA: Geometry-Aware Local Adaptive Grids for Detailed 3D Generation
We propose GALA, a novel representation of 3D shapes that (i) excels at capturin g and reproducing complex geometry and surface details, (ii) is computationally efficient, and (iii) lends itself to 3D generative modelling with modern, diffus ion-based schemes. The key idea of GALA is to exploit both the global sparsity o f surfaces within a 3D volume and their local surface properties. *Sparsity* is promoted by covering only the 3D object boundaries, not empty space, with an ens emble of tree root voxels. Each voxel contains an octree to further limit storag e and compute to regions that contain surfaces. *Adaptivity* is achieved by fitt ing one local and geometry-aware coordinate frame in each non-empty leaf node. A djusting the orientation of the local grid, as well as the anisotropic scales of its axes, to the local surface shape greatly increases the amount of detail th at can be stored in a given amount of memory, which in turn allows for quantizat ion without loss of quality. With our optimized C++/CUDA implementation, GALA ca n be fitted to an object in less than 10 seconds. Moreover, the representation c an efficiently be flattened and manipulated with transformer networks. We provid e a cascaded generation pipeline capable of generating 3D shapes with great geom etric detail. For more information, please visit our [project page](https://sant isy.github.io/GALA/).
****

Jeffrey Ouyang-Zhang,Chengyue Gong,Yue Zhao,Philipp Kraehenbuehl,Adam Klivans,Da niel Jesus Diaz
Distilling Structural Representations into Protein Sequence Models
Protein language (or sequence) models, like the popular ESM2, are now widely use d tools for extracting evolution-based protein representations and have achieved significant success on core downstream biological tasks.
A major open problem is how to obtain representations that best capture both the sequence evolutionary history and the atomic structural properties of proteins in general.
We introduce **I**mplicit **S**equence **M**odel, a sequence-only input model wi th structurally-enriched representations that outperforms state-of-the-art seque nce models on several well-studied benchmarks including mutation stability asses sment and structure prediction.
Our key innovations are a microenvironment-based Autoencoder for generating stru cture tokens and a self-supervised training objective that distills these tokens into ESM2's pre-trained model.
Notably, we make ISM's structure-enriched weights easily accessible for any appl ication using the ESM2 framework.
****

Hao-Tang Tsui,Chien-Yao Wang,Hong-Yuan Mark Liao

YOLO-RD: Introducing Relevant and Compact Explicit Knowledge to YOLO by Retriever-Dictionary

Identifying and localizing objects within images is a fundamental challenge, and numerous efforts have been made to enhance model accuracy by experimenting with diverse architectures and refining training strategies. Nevertheless, a prevalent limitation in existing models is overemphasizing the current input while ignoring the information from the entire dataset. We introduce an innovative $\textbf{R}etriever-\textbf{D}ictionary$ (RD) module to address this issue. This architecture enables YOLO-based models to efficiently retrieve features from a Dictionary that contains the insight of the dataset, which is built by the knowledge from Visual Models (VM), Large Language Models (LLM), or Visual Language Models (VLM). The flexible RD enables the model to incorporate such explicit knowledge that enhances the ability to benefit multiple tasks, specifically, segmentation, detection, and classification, from pixel to image level. The experiments show that using the RD significantly improves model performance, achieving more than a 3\% increase in mean Average Precision for object detection with less than a 1\% increase in model parameters. Beyond 1-stage object detection models, the RD module improves the effectiveness of 2-stage models and DETR-based architectures, such as Faster R-CNN and Deformable DETR. Code is released at https://github.com/henrytsui000/YOLO.

****

Zhihai Wang,Jie Wang,Xilin Xia,Dongsheng Zuo,Lei Chen,Yuzhe Ma,Jianye HAO,Mingxuan Yuan,Feng Wu

Computing Circuits Optimization via Model-Based Circuit Genetic Evolution

Optimizing computing circuits such as multipliers and adders is a fundamental challenge in modern integrated circuit design. Recent efforts propose formulating this optimization problem as a reinforcement learning (RL) proxy task, offering a promising approach to search high-speed and area-efficient circuit design solutions. However, we show that the RL-based formulation (proxy task) converges to a local optimal design solution (original task) due to the deceptive reward signals and incrementally localized actions in the RL-based formulation. To address this challenge, we propose a novel model-based circuit genetic evolution (MUTE) framework, which reformulates the problem as a genetic evolution process by proposing a grid-based genetic representation of design solutions. This novel formulation avoids misleading rewards by evaluating and improving generated solutions using the true objective value rather than proxy rewards. To promote globally diverse exploration, MUTE proposes a multi-granularity genetic crossover operator that recombines design substructures at varying column ranges between two grid-based genetic solutions. To the best of our knowledge, MUTE is the first to reformulate the problem as a circuit genetic evolution process, which enables effectively searching for global optimal design solutions. We evaluate MUTE on several fundamental computing circuits, including multipliers, adders, and multiply-accumulate circuits. Experiments on these circuits demonstrate that MUTE significantly Pareto-dominates state-of-the-art approaches in terms of both area and delay. Moreover, experiments demonstrate that circuits designed by MUTE well generalize to large-scale computation-intensive circuits as well.

****

Wanlin Zhang,Weichen Lin,Ruomin Huang,Shihong Song,Hu Ding

To Tackle Adversarial Transferability: A Novel Ensemble Training Method with Fourier Transformation

Ensemble methods are commonly used for enhancing robustness in machine learning. However, due to the ''transferability'' of adversarial examples, the performance of an ensemble model can be
seriously affected even it contains a set of independently trained sub-models. To address this issue, we propose an efficient data transformation method based on a cute ''weakness allocation'' strategy, to diversify non-robust features.
Our approach relies on a fine-grained analysis on the relation between non-robust features and adversarial attack directions.
Moreover, our approach enjoys several other advantages, e.g., it does not require any communication between sub-models and the construction complexity is also

quite low.
We conduct a set of experiments to evaluate the performance of our proposed method and compare it with several popular baselines. The results suggest that our approach can achieve significantly improved robust accuracy over most existing ensemble methods, and meanwhile preserve high clean accuracy.
****
Amirhossein Alimohammadi,Sauradip Nag,Saeid Asgari,Andrea Tagliasacchi,Ghassan Hamarneh,Ali Mahdavi Amiri
SMITE: Segment Me In TimE
Segmenting an object in a video presents significant challenges. Each pixel must be accurately labeled, and these labels must remain consistent across frames. The difficulty increases when the segmentation is with arbitrary granularity, meaning the number of segments can vary arbitrarily, and masks are defined based on only one or a few sample images. In this paper, we address this issue by employing a
pre-trained text to image diffusion model supplemented with an additional tracking mechanism. We demonstrate that our approach can effectively manage various segmentation scenarios and outperforms state-of-the-art alternatives. The project page is available at https://segment-me-in-time.github.io/
****
Ayush Mohanty,Nazal Mohamed,Paritosh Ramanan,Nagi Gebraeel
Federated Granger Causality Learning For Interdependent Clients With State Space Representation
Advanced sensors and IoT devices have improved the monitoring and control of complex industrial enterprises. They have also created an interdependent fabric of geographically distributed process operations (clients) across these enterprises. Granger causality is an effective approach to detect and quantify interdependencies by examining how the state of one client affects the states of others over time. Understanding these interdependencies helps capture how localized events, such as faults and disruptions, can propagate throughout the system, potentially leading to widespread operational impacts. However, the large volume and complexity of industrial data present significant challenges in effectively modeling these interdependencies. This paper develops a federated approach to learning Granger causality. We utilize a linear state space system framework that leverages low-dimensional state estimates to analyze interdependencies. This helps address bandwidth limitations and the computational burden commonly associated with centralized data processing. We propose augmenting the client models with the Granger causality information learned by the server through a Machine
Learning (ML) function. We examine the co-dependence between the augmented client and server models and reformulate the framework as a standalone ML algorithm providing conditions for its sublinear and linear convergence rates. We also study the convergence of the framework to a centralized oracle model. Moreover, we include a differential privacy analysis to ensure data security while preserving causal insights. Using synthetic data, we conduct comprehensive experiments to demonstrate the robustness of our approach to perturbations in causality, the scalability to the size of communication, number of clients, and the dimensions of raw data. We also evaluate the performance on two real-world industrial control system datasets by reporting the volume of data saved by decentralization.
****
Keir Adams,Kento Abeywardane,Jenna Fromer,Connor W. Coley
ShEPhERD: Diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design
Engineering molecules to exhibit precise 3D intermolecular interactions with their environment forms the basis of chemical design. In ligand-based drug design, bioisosteric analogues of known bioactive hits are often identified by virtually screening chemical libraries with shape, electrostatic, and pharmacophore similarity scoring functions. We instead hypothesize that a generative model which learns the joint distribution over 3D molecular structures and their interaction profiles may facilitate 3D interaction-aware chemical design. We specifically design ShEPhERD, an SE(3)-equivariant diffusion model which jointly diffuses/denois

es 3D molecular graphs and representations of their shapes, electrostatic potent
ial surfaces, and (directional) pharmacophores to/from Gaussian noise. Inspired
by traditional ligand discovery, we compose 3D similarity scoring functions to a
ssess ShEPhERD’s ability to conditionally generate novel molecules with desired
interaction profiles. We demonstrate ShEPhERD’s potential for impact via exempla
ry drug design tasks including natural product ligand hopping, protein-blind bio
active hit diversification, and bioisosteric fragment merging.
****
Sunny Duan,Mikail Khona,Abhiram Iyer,Rylan Schaeffer,Ila R Fiete
Uncovering Latent Memories in Large Language Models
Frontier AI systems are making transformative impacts across society, but such b
enefits are not without costs: models trained on web-scale datasets containing p
ersonal and private data raise profound concerns about data privacy and security
. Language models are trained on extensive corpora including potentially sensiti
ve or proprietary information, and the risk of data leakage, where the model res
ponse reveals pieces of such information, remains inadequately understood. Prior
 work has investigated that sequence complexity and the number of repetitions ar
e the primary drivers of memorization. In this work, we examine the most vulnera
ble class of data: highly complex sequences that are presented only once during
training. These sequences often contain the most sensitive information and pose
considerable risk if memorized. By analyzing the progression of memorization for
 these sequences throughout training, we uncover a striking observation: many me
morized sequences persist in the model's memory, exhibiting resistance to catast
rophic forgetting even after just one encounter. Surprisingly, these sequences m
ay not appear memorized immediately after their first exposure but can later be
“uncovered” during training, even in the absence of subsequent exposures - a phe
nomenon we call "latent memorization." Latent memorization presents a serious ch
allenge for data privacy, as sequences that seem hidden at the final checkpoint
of a model may still be easily recoverable. We demonstrate how these hidden sequ
ences can be revealed through random weight perturbations, and we introduce a di
agnostic test based on cross-entropy loss to accurately identify latent memorize
d sequences.
****
Jie Liu,Pan Zhou,Yingjun Du,Ah-Hwee Tan,Cees G. M. Snoek,Jan-Jakob Sonke,Efstrat
ios Gavves
CaPo: Cooperative Plan Optimization for Efficient Embodied Multi-Agent Cooperati
on
In this work, we address the cooperation problem among large language model (LLM
) based embodied agents, where agents must cooperate to achieve a common goal. P
revious methods often execute actions extemporaneously and incoherently, without
 long-term strategic and cooperative planning, leading to redundant steps, fail
ures, and even serious repercussions in complex tasks like search-and-rescue mis
sions where discussion and cooperative plan are crucial. To solve this issue, w
e propose Cooperative Plan Optimization (CaPo) to enhance the cooperation effici
ency of LLM-based embodied agents. Inspired by human cooperation schemes, CaPo i
mproves cooperation efficiency with two phases: 1) meta plan generation, and 2)
 progress-adaptive meta plan and execution. In the first phase, all agents analy
ze the task, discuss, and cooperatively create a meta-plan that decomposes the t
ask into subtasks with detailed steps, ensuring a long-term strategic and cohere
nt plan for efficient coordination. In the second phase, agents execute tasks a
ccording to the meta-plan and dynamically adjust it based on their latest progre
ss (e.g., discovering a target object) through multi-turn discussions. This pro
gress-based adaptation eliminates redundant actions, improving the overall coope
ration efficiency of agents. Experimental results on the ThreeDworld Multi-Agent
 Transport and Communicative Watch-And-Help tasks demonstrate CaPo's much higher
 task completion rate and efficiency compared with state-of-the-arts. The code
is released at https://github.com/jliu4ai/CaPo.
****
Hadas Orgad,Michael Toker,Zorik Gekhman,Roi Reichart,Idan Szpektor,Hadas Kotek,Y
onatan Belinkov

LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations
Large language models (LLMs) often produce errors, including factual inaccuracies, biases, and reasoning failures, collectively referred to as "hallucinations". Recent studies have demonstrated that LLMs' internal states encode information regarding the truthfulness of their outputs, and that this information can be utilized to detect errors. In this work, we show that the internal representations of LLMs encode much more information about truthfulness than previously recognized. We first discover that the truthfulness information is concentrated in specific tokens, and leveraging this property significantly enhances error detection performance. Yet, we show that such error detectors fail to generalize across datasets, implying that---contrary to prior claims---truthfulness encoding is not universal but rather multifaceted. Next, we show that internal representations can also be used for predicting the types of errors the model is likely to make, facilitating the development of tailored mitigation strategies. Lastly, we reveal a discrepancy between LLMs' internal encoding and external behavior: they may encode the correct answer, yet consistently generate an incorrect one. Taken together, these insights deepen our understanding of LLM errors from the model's internal perspective, which can guide future research on enhancing error analysis and mitigation.
****
Zhengyue Zhao,Xiaogeng Liu,Somesh Jha,Patrick McDaniel,Bo Li,Chaowei Xiao
Can Watermarks be Used to Detect LLM IP Infringement For Free?
The powerful capabilities of LLMs stem from their rich training data and high-quality labeled datasets, making the training of strong LLMs a resource-intensive process, which elevates the importance of IP protection for such LLMs. Compared to gathering high-quality labeled data, directly sampling outputs from these fully trained LLMs as training data presents a more cost-effective approach. This practice—where a suspect model is fine-tuned using high-quality data derived from these LLMs, thereby gaining capabilities similar to the target model—can be seen as a form of IP infringement against the original LLM. In recent years, LLM watermarks have been proposed and used to detect whether a text is AI-generated. Intuitively, if data sampled from a watermarked LLM is used for training, the resulting model would also be influenced by this watermark. This raises the question: can we directly use such watermarks to detect IP infringement of LLMs? In this paper, we explore the potential of LLM watermarks for detecting model infringement. We find that there are two issues with direct detection: (1) The queries used to sample output from the suspect LLM have a significant impact on detectability. (2) The watermark that is easily learned by LLMs exhibits instability regarding the watermark's hash key during detection. To address these issues, we propose LIDet, a detection method that leverages available anchor LLMs to select suitable queries for sampling from the suspect LLM. Additionally, it adapts the detection threshold to mitigate detection failures caused by different hash keys. To demonstrate the effectiveness of this approach, we construct a challenging model set containing multiple suspect LLMs on which direct detection methods struggle to yield effective results. Our method achieves over 90\% accuracy in distinguishing between infringing and clean models, demonstrating the feasibility of using LLM watermarks to detect LLM IP infringement.
****
Yong Liu,Guo Qin,Xiangdong Huang,Jianmin Wang,Mingsheng Long
Timer-XL: Long-Context Transformers for Unified Time Series Forecasting
We present Timer-XL, a causal Transformer for unified time series forecasting. To uniformly predict multidimensional time series, we generalize next token prediction, predominantly adopted for 1D token sequences, to multivariate next token prediction. The paradigm formulates various forecasting tasks as a long-context prediction problem. We opt for decoder-only Transformers that capture causal dependencies from varying-length contexts for unified forecasting, making predictions on non-stationary univariate time series, multivariate series with complicated dynamics and correlations, as well as covariate-informed contexts that include exogenous variables. Technically, we propose a universal TimeAttention to captu

re fine-grained intra- and inter-series dependencies of flattened time series to
kens (patches), which is further enhanced by deft position embedding for tempora
l causality and variable equivalence. Timer-XL achieves state-of-the-art perform
ance across task-specific forecasting benchmarks through a unified approach. Bas
ed on large-scale pre-training, Timer-XL achieves state-of-the-art zero-shot per
formance, making it a promising architecture for pre-trained time series models.
 Code is available at this repository: https://github.com/thuml/Timer-XL.
****
Omer Moussa,Dietrich Klakow,Mariya Toneva
Improving Semantic Understanding in Speech Language Models via Brain-tuning
Speech language models align with human brain responses to natural language to a
n impressive degree. However, current models rely heavily on low-level speech fe
atures, indicating they lack brain-relevant semantics which limits their utility
 as model organisms of semantic processing in the brain. In this work, we addres
s this limitation by inducing brain-relevant bias directly into the models via f
ine-tuning with fMRI recordings of people listening to natural stories--a proces
s we name brain-tuning. After testing it on 3 different pretrained model familie
s, we show that brain-tuning not only improves overall alignment with new brain
recordings in semantic language regions, but also reduces the reliance on low-le
vel speech features for this alignment. Excitingly, we further show that brain-t
uning leads to 1) consistent improvements in performance on semantic downstream
tasks and 2) a representational space with increased semantic preference. Our re
sults provide converging evidence, for the first time, that incorporating brain
signals into the training of language models improves the models’ semantic under
standing.
****
Jiewen Hu,Thomas Zhu,Sean Welleck
miniCTX: Neural Theorem Proving with (Long-)Contexts
Real-world formal theorem proving often depends on a wealth of context, includin
g definitions, lemmas, comments, file structure, and other information. We intro
duce $\texttt{miniCTX}$, which tests a model's ability to prove formal mathemati
cal theorems that depend on new context that is not seen during training. $\text
tt{miniCTX}$ contains theorems sourced from real Lean projects and textbooks, ea
ch associated with a context that can span tens of thousands of tokens. Models a
re tasked with proving a theorem given access to code from the theorem's reposit
ory, which contains context that is needed for the proof. As a baseline for $\te
xttt{miniCTX}$, we tested fine-tuning and prompting methods that condition theor
em proving on preceding context. Both approaches substantially outperform tradit
ional methods that rely solely on state information. We found that this ability
to use context is not captured by previous benchmarks such as $\texttt{miniF2F}$
. Alongside $\texttt{miniCTX}$, we offer $\texttt{ntp-toolkit}$ for automaticall
y extracting and annotating theorem proving data, making it easy to add new proj
ects into $\texttt{miniCTX}$ to ensure that contexts are not seen during trainin
g. $\texttt{miniCTX}$ offers a challenging and realistic evaluation of neural th
eorem provers.
****
Zukang Xu,Yuxuan Yue,Xing Hu,Dawei Yang,Zhihang Yuan,Zixu Jiang,Zhixuan Chen,Jia
ngyongYu,XUCHEN,Sifan Zhou
MambaQuant: Quantizing the Mamba Family with Variance Aligned Rotation Methods
Mamba is an efficient sequence model that rivals Transformers and demonstrates s
ignificant potential as a foundational architecture for various tasks. Quantizat
ion is commonly used in neural networks to reduce model size and computational l
atency. However, applying quantization to Mamba remains underexplored, and exist
ing quantization methods, which have been effective for CNN and Transformer mode
ls, appear inadequate for Mamba models (e.g., Quarot suffers a 21% accuracy drop
 on Vim-T$\dagger$ even under W8A8). We have pioneered the exploration of this i
ssue and identified several key challenges. First, significant outliers areprese
nt in gate projections, output projections, and matrix multiplications. Second,
Mamba’s unique parallel scan further amplifies these outliers, leading to uneven
 and heavy-tailed data distributions. Third, even with the application of the Ha

damard transform, the variance across channels in weights and activations still remains inconsistent. To these ends, we propose MambaQuant, a post-training quantization (PTQ) framework consisting of: 1) Karhunen-Lo`eve Transformation (KLT) enhanced rotation, rendering the rotation matrix adaptable to diverse channel distributions. 2) Smooth-Fused rotation, which equalizes channel variances and can merge additional parameters into model weights. Experiments show that MambaQuant can quantize both weights and activations into 8-bit with less than 1% accuracy loss for Mamba-based vision and language tasks. To our knowledge, MambaQuant is the first comprehensive PTQ design for the Mamba family, paving the way for further advancements in its application.
****

Chen Hou,Zhibo Chen

Training-free Camera Control for Video Generation

We propose a training-free and robust solution to offer camera movement control for off-the-shelf video diffusion models. Unlike previous work, our method does not require any supervised finetuning on camera-annotated datasets or self-supervised training via data augmentation. Instead, it is plug-and-play with most pretrained video diffusion models and can generate camera-controllable videos with a single image or text prompt as input. The inspiration for our work comes from the layout prior that intermediate latents encode for the generated results, thus rearranging noisy pixels in them will cause the output content to relocate as well. As camera moving could also be seen as a type of pixel rearrangement caused by perspective change, videos can be reorganized following specific camera motion if their noisy latents change accordingly. Building on this, we propose **CamTrol**, which enables robust camera control for video diffusion models. It is achieved by a two-stage process. First, we model image layout rearrangement through explicit camera movement in 3D point cloud space. Second, we generate videos with camera motion by leveraging the layout prior of noisy latents formed by a series of rearranged images. Extensive experiments have demonstrated its superior performance in both video generation and camera motion alignment compared with other finetuned methods. Furthermore, we show the capability of CamTrol to generalize to various base models, as well as its impressive applications in scalable motion control, dealing with complicated trajectories and unsupervised 3D video generation. Videos available at https://lifedecoder.github.io/CamTrol/.
****

Yufei Huang,Yunshu Liu,Lirong Wu,Haitao Lin,Cheng Tan,Odin Zhang,Zhangyang Gao,Siyuan Li,Zicheng Liu,Yunfan Liu,Tailin Wu,Stan Z. Li

EVA: Geometric Inverse Design for Fast Protein Motif-Scaffolding with Coupled Flow

Motif-scaffolding is a fundamental component of protein design, which aims to construct the scaffold structure that stabilizes motifs conferring desired functions. Recent advances in generative models are promising for designing scaffolds, with two main approaches: training-based and sampling-based methods. Training-based methods are resource-heavy and slow, while training-free sampling-based methods are flexible but require numerous sampling steps and costly, unstable guidance. To speed up and improve sampling-based methods, we analyzed failure cases and found that errors stem from the trade-off between generation and guidance. Thus we proposed to exploit the spatial context and adjust the generative direction to be consistent with guidance to overcome this trade-off. Motivated by this, we formulate motif-scaffolding as a Geometric Inverse Design task inspired by the image inverse problem, and present Evolution-ViA-reconstruction (EVA), a novel sampling-based coupled flow framework on geometric manifolds, which starts with a pretrained flow-based generative model. EVA uses motif-coupled priors to leverage spatial contexts, guiding the generative process along a straighter probability path, with generative directions aligned with guidance in the early sampling steps. EVA is 70× faster than SOTA model RFDiffusion with competitive and even better performance on benchmark tests. Further experiments on real-world cases including vaccine design, multi-motif scaffolding and motif optimal placement searching demonstrate EVA's superior efficiency and effectiveness.
****

Yiwen Chen,Tong He,Di Huang,Weicai Ye,Sijin Chen,Jiaxiang Tang,Zhongang Cai,Lei Yang,Gang Yu,Guosheng Lin,Chi Zhang
MeshAnything: Artist-Created Mesh Generation with Autoregressive Transformers
Recently, 3D assets created via reconstruction and generation have matched the quality of manually crafted assets, highlighting their potential for replacement. However, this potential is largely unrealized because these assets always need to be converted to meshes for 3D industry applications, and the meshes produced by current mesh extraction methods are significantly inferior to Artist-Created Meshes (AMs), i.e., meshes created by human artists.

Specifically, current mesh extraction methods rely on dense faces and ignore geometric features, leading to inefficiencies, complicated post-processing, and lower representation quality.
To address these issues, we introduce MeshAnything, a model that treats mesh extraction as a generation problem, producing AMs aligned with specified shapes.

By converting 3D assets in any 3D representation into AMs, MeshAnything can be integrated with various 3D asset production methods, thereby enhancing their application across the 3D industry.
The architecture of MeshAnything comprises a VQ-VAE and a shape-conditioned decoder-only transformer. We first learn a mesh vocabulary using the VQ-VAE, then train the shape-conditioned decoder-only transformer on this vocabulary for shape-conditioned autoregressive mesh generation. Our extensive experiments show that our method generates AMs with hundreds of times fewer faces, significantly improving storage, rendering, and simulation efficiencies, while achieving precision comparable to previous methods.
****
Yuxin Dai,Qi Wang,Jingsen Zhu,Dianbing Xi,Yuchi Huo,Chen Qian,Ying He
Inverse Rendering using Multi-Bounce Path Tracing and Reservoir Sampling
We introduce MIRReS, a novel two-stage inverse rendering framework that jointly reconstructs and optimizes explicit geometry, materials, and lighting from multi-view images. Unlike previous methods that rely on implicit irradiance fields or oversimplified ray tracing, our method begins with an initial stage that extracts an explicit triangular mesh. In the second stage, we refine this representation using a physically-based inverse rendering model with multi-bounce path tracing and Monte Carlo integration. This enables our method to accurately estimate indirect illumination effects, including self-shadowing and internal reflections, leading to a more precise intrinsic decomposition of shape, material, and lighting. To address the noise issue in Monte Carlo integration, we incorporate reservoir sampling, improving convergence and enabling efficient gradient-based optimization with low sample counts. Through both qualitative and quantitative assessments across various scenarios, especially those with complex shadows, we demonstrate that our method achieves state-of-the-art decomposition performance. Furthermore, our optimized explicit geometry seamlessly integrates with modern graphics engines supporting downstream applications such as scene editing, relighting, and material editing.
****
Jie Yang,Yuwen Wang,Kaixuan Chen,Tongya Zheng,Yihe Zhou,Zhenbang Xiao,Ji Cao,Mingli Song,Shunyu Liu
From GNNs to Trees: Multi-Granular Interpretability for Graph Neural Networks
Interpretable Graph Neural Networks (GNNs) aim to reveal the underlying reasoning behind model predictions, attributing their decisions to specific subgraphs that are informative. However, existing subgraph-based interpretable methods suffer from an overemphasis on local structure, potentially overlooking long-range dependencies within the entire graphs. Although recent efforts that rely on graph coarsening have proven beneficial for global interpretability, they inevitably reduce the graphs to a fixed granularity. Such an inflexible way can only capture graph connectivity at a specific level, whereas real-world graph tasks often exhibit relationships at varying granularities (e.g., relevant interactions in pro

teins span from functional groups, to amino acids, and up to protein domains). In this paper, we introduce a novel Tree-like Interpretable Framework (TIF) for graph classification, where plain GNNs are transformed into hierarchical trees, with each level featuring coarsened graphs of different granularity as tree nodes. Specifically, TIF iteratively adopts a graph coarsening module to compress original graphs (i.e., root nodes of trees) into increasingly coarser ones (i.e., child nodes of trees), while preserving diversity among tree nodes within different branches through a dedicated graph perturbation module. Finally, we propose an adaptive routing module to identify the most informative root-to-leaf paths, providing not only the final prediction but also the multi-granular interpretability for the decision-making process. Extensive experiments on the graph classification benchmarks with both synthetic and real-world datasets demonstrate the superiority of TIF in interpretability, while also delivering a competitive prediction performance akin to the state-of-the-art counterparts.
****
Xiaoran Liu,Ruixiao Li,Zhigeng Liu,Qipeng Guo,Yuerong Song,Kai Lv,Hang Yan,Linlin Li,Qun Liu,Xipeng Qiu
ReAttention: Training-Free Infinite Context with Finite Attention Scope
The long-context capability of the Large Language Models (LLM) has made significant breakthroughs, but \textit{the maximum supported context length in length extrapolation} remains a critical bottleneck limiting their practical applications. The constraint of context length in LLMs arises from the self-attention mechanism, which cannot effectively and efficiently capture the semantic relationships within infinitely long contexts via the limited pre-trained positional information and attention scope. In this work, we propose \textbf{ReAttention}, a training-free approach enabling LLM based on the self-attention mechanism to support an infinite context with a finite attention scope under sufficient memory resources. ReAttention performs the position-agnostic top-k attention before the ordinary position-aware self-attention, freeing LLMs from the length extrapolation issue. We validate the performance of ReAttention on the LongBench, L-Eval, and InfiniteBench and demonstrate that it is on par with traditional methods. Furthermore, we also apply ReAttention on mainstream LLMs, including LLaMA3.1-8B and Mistral-v0.3-7B, enabling them to support context lengths of at least 1M and even expanding the context length of LLaMA3.2-3B-chat by 128$\times$ to 4M without any further training in Needle-In-A-Haystack tests. We also improve the efficiency of ReAttention with Triton and achieve an efficient extrapolation without additional overhead. The code is available at \url{https://github.com/OpenMOSS/ReAttention}.
****
Runyu Lu,Yuanheng Zhu,Dongbin Zhao
Divergence-Regularized Discounted Aggregation: Equilibrium Finding in Multiplayer Partially Observable Stochastic Games
This paper presents Divergence-Regularized Discounted Aggregation (DRDA), a multi-round learning system for solving partially observable stochastic games (POSGs). DRDA is based on action values and applicable to multiplayer POSGs, which can unify normal-form games (NFGs), extensive-form games (EFGs) with perfect recall, and Markov games (MGs). In each single round, DRDA can be viewed as a discounted variant of Follow the Regularized Leader (FTRL) under a general value function for POSGs. While previous studies on discounted FTRL have demonstrated its last-iterate convergence towards quantal response equilibrium (QRE) in NFGs, this paper extends the theoretical results to POSGs under divergence regularization and generalizes the QRE concept of Nash distribution. The linear last-iterate convergence of single-round DRDA to its rest point is proved under the assumption on the hypomonotonicity of the game. When the rest point is unique, it induces the unique Nash distribution defined in the POSG, which has a bounded deviation from Nash equilibrium (NE). Under multiple learning rounds, DRDA keeps replacing the base policy for divergence regularization with the policy at the rest point in the previous round. It is further proved that the limit point of multi-round DRDA must be an exact NE (rather than a QRE). In experiments, discrete-time DRDA can converge to NE at a near-exponential rate in (multiplayer) NFGs and outperfor

m the existing baselines for EFGs, MGs, and typical POSGs.
****
Hongjoon Ahn,Jinu Hyeon,Youngmin Oh,Bosun Hwang,Taesup Moon
Prevalence of Negative Transfer in Continual Reinforcement Learning: Analyses an d a Simple Baseline
We argue that the negative transfer problem occurring when the new task to learn arrives is an important problem that needs not be overlooked when developing ef fective Continual Reinforcement Learning (CRL) algorithms. Through comprehensive experimental validation, we demonstrate that such issue frequently exists in CR L and cannot be effectively addressed by several recent work on either mitigatin g plasticity loss of RL agents or enhancing the positive transfer in CRL scenari o. To that end, we develop Reset & Distill (R&D), a simple yet highly effective baseline method, to overcome the negative transfer problem in CRL. R&D combines a strategy of resetting the agent's online actor and critic networks to learn a new task and an offline learning step for distilling the knowledge from the onli ne actor and previous expert's action probabilities. We carried out extensive ex periments on long sequence of Meta World tasks and show that our simple baseline method consistently outperforms recent approaches, achieving significantly high er success rates across a range of tasks. Our findings highlight the importance of considering negative transfer in CRL and emphasize the need for robust strate gies like R&D to mitigate its detrimental effects.
****
Yubo Wang,Jianting Tang,Chaohu Liu,Linli Xu
Tracking the Copyright of Large Vision-Language Models through Parameter Learnin g Adversarial Images
Large vision-language models (LVLMs) have demonstrated remarkable image understa nding and dialogue capabilities, allowing them to handle a variety of visual que stion answering tasks. However, their widespread availability raises concerns ab out unauthorized usage and copyright infringement, where users or individuals ca n develop their own LVLMs by fine-tuning published models. In this paper, we pro pose a novel method called Parameter Learning Attack (PLA) for tracking the copy right of LVLMs without modifying the original model. Specifically, we construct adversarial images through targeted attacks against the original model, enabling it to generate specific outputs. To ensure these attacks remain effective on po tential fine-tuned models to trigger copyright tracking, we allow the original m odel to learn the trigger images by updating parameters in the opposite directio n during the adversarial attack process. Notably, the proposed method can be app lied after the release of the original model, thus not affecting the model’s per formance and behavior. To simulate real-world applications, we fine-tune the ori ginal model using various strategies across diverse datasets, creating a range o f models for copyright verification. Extensive experiments demonstrate that our method can more effectively identify the original copyright of fine-tuned models compared to baseline methods. Therefore, this work provides a powerful tool for tracking copyrights and detecting unlicensed usage of LVLMs.
****
Junpeng Yue,Xinrun Xu,Börje F. Karlsson,Zongqing Lu
MLLM as Retriever: Interactively Learning Multimodal Retrieval for Embodied Agen ts
MLLM agents demonstrate potential for complex embodied tasks by retrieving multi modal task-relevant trajectory data. However, current retrieval methods primaril y focus on surface-level similarities of textual or visual cues in trajectories, neglecting their effectiveness for the specific task at hand. To address this i ssue, we propose a novel method, MART, which enhances the performance of embodie d agents by utilizing interaction data to fine-tune an MLLM retriever based on p reference learning, such that the retriever fully considers the effectiveness of trajectories and prioritize them for unseen tasks. We also introduce Trajectory Abstraction, a mechanism that leverages MLLMs' summarization capabilities to re present trajectories with fewer tokens while preserving key information, enablin g agents to better comprehend milestones in the trajectory. Experimental results across various environments demonstrate our method significantly improves task

success rates in unseen scenes compared to baseline methods. This work presents a new paradigm for multimodal retrieval in embodied agents, by fine-tuning a general-purpose MLLM as the retriever to assess trajectory effectiveness. All the code for benchmark tasks, simulator modifications and the MLLM retriever is available at https://github.com/PKU-RL/MART.
****

Wei Zhao,Pengxiang Ding,Zhang Min,Zhefei Gong,Shuanghao Bai,Han Zhao,Donglin Wang

VLAS: Vision-Language-Action Model with Speech Instructions for Customized Robot Manipulation

Vision-language-action models (VLAs) have recently become highly prevalent in robot manipulation due to its end-to-end architecture and impressive performance. However, current VLAs are limited to processing human instructions in textual form, neglecting the more natural speech modality for human interaction. A typical approach of incorporating speech modality into VLA necessitates a separate speech recognition system to transcribe spoken instructions into text. Such a cascading pipeline raises two major concerns for robotic systems. First, the entire model grows in size and complexity, potentially resulting in redundant computations and increased memory consumption. Second, the transcription procedure would lose non-semantic information in the raw speech, such as voiceprint, which is crucial for a robot to successfully understand and complete customized tasks. To this end, we propose VLAS, the fisrt end-to-end policy model that seamlessly integrates speech modality for robot manipulation. We present a three-stage speech instruction tuning strategy leveraging multimodal datasets, including our manually curated SQA and CSI datasets. Furthermore, to facilitate personalized operations, we develop a voice retrieval-augmented generation (RAG) approach to enhance the robot's performance in tasks requiring individual-specific knowledge. Experimental results show that the proposed VLAS, following either textual or speech instructions, can achieve performance comparable to traditional VLAs on the CALVIN benchmark. In addition, we created a benchmark consisting of customization tasks, where our VLAS demonstrates absolute superiority by fully leveraging the auxiliary information in speech.
****

Chen Qian,Zihao Xie,YiFei Wang,Wei Liu,Kunlun Zhu,Hanchen Xia,Yufan Dang,Zhuoyun Du,Weize Chen,Cheng Yang,Zhiyuan Liu,Maosong Sun

Scaling Large Language Model-based Multi-Agent Collaboration

Recent breakthroughs in large language model-driven autonomous agents have revealed that multi-agent collaboration often surpasses each individual through collective reasoning. Inspired by the neural scaling law—increasing neurons enhances performance, this study explores whether the continuous addition of collaborative agents can yield similar benefits. Technically, we utilize directed acyclic graphs to organize agents into a multi-agent collaboration network (MacNet), upon which their interactive reasoning is topologically orchestrated for autonomous task solving. Extensive evaluations reveal that it effectively supports collaboration among over a thousand agents, with irregular topologies outperforming regular ones. We also identify a collaborative scaling law—the overall performance follows a logistic growth pattern as agents scale, with collaborative emergence occurring earlier than traditional neural emergence. We speculate this may be because scaling agents catalyzes their multidimensional considerations during interactive reflection and refinement, thereby producing more comprehensive artifacts. The code is available at https://github.com/OpenBMB/ChatDev/tree/macnet.
****

Bilgehan Sel,Ruoxi Jia,Ming Jin

LLMs Can Plan Only If We Tell Them

Large language models (LLMs) have demonstrated significant capabilities in natural language processing and reasoning, yet their effectiveness in autonomous planning has been under debate. While existing studies have utilized LLMs with external feedback mechanisms or in controlled environments for planning, these approaches often involve substantial computational and development resources due to the requirement for careful design and iterative backprompting. Moreover, even the

most advanced LLMs like GPT-4 struggle to match human performance on standard planning benchmarks, such as the Blocksworld, without additional support. This paper investigates whether LLMs can independently generate long-horizon plans that rival human baselines. Our novel enhancements to Algorithm-of-Thoughts (AoT), which we dub AoT+, help achieve state-of-the-art results in planning benchmarks out-competing prior methods and human baselines all autonomously.
****
Yukun Huang,Sanxing Chen,Hongyi Cai,Bhuwan Dhingra
To Trust or Not to Trust? Enhancing Large Language Models' Situated Faithfulness to External Contexts
Large Language Models (LLMs) are often augmented with external contexts, such as those used in retrieval-augmented generation (RAG). However, these contexts can be inaccurate or intentionally misleading, leading to conflicts with the model's internal knowledge. We argue that robust LLMs should demonstrate situated faithfulness, dynamically calibrating their trust in external information based on their confidence in the internal knowledge and the external context to resolve knowledge conflicts. To benchmark this capability, we evaluate LLMs across several QA datasets, including a newly created dataset featuring in-the-wild incorrect contexts sourced from Reddit posts. We show that when provided with both correct and incorrect contexts, both open-source and proprietary models tend to overly rely on external information, regardless of its factual accuracy. To enhance situated faithfulness, we propose two approaches: Self-Guided Confidence Reasoning (SCR) and Rule-Based Confidence Reasoning (RCR). SCR enables models to self-access the confidence of external information relative to their own internal knowledge to produce the most accurate answer. RCR, in contrast, extracts explicit confidence signals from the LLM and determines the final answer using predefined rules.
Our results show that for LLMs with strong reasoning capabilities, such as GPT-4o and GPT-4o mini, SCR outperforms RCR, achieving improvements of up to 24.2\% over a direct input augmentation baseline. Conversely, for a smaller model like Llama-3-8B, RCR outperforms SCR. Fine-tuning SCR with our proposed Confidence Reasoning Direct Preference Optimization (CR-DPO) method improves performance on both seen and unseen datasets, yielding an average improvement of 8.9\% on Llama-3-8B. In addition to quantitative results, we offer insights into the relative strengths of SCR and RCR. Our findings highlight promising avenues for improving situated faithfulness in LLMs.
****
Carlos J Soto,Matthew Reimherr,Aleksandra Slavkovic,Mark Shriver
Gaussian Differentially Private Human Faces Under a Face Radial Curve Representation
In this paper we consider the problem of releasing a Gaussian Differentially Private (GDP) 3D human face. The human face is a complex structure with many features and inherently tied to one's identity. Protecting this data, in a formally private way, is important yet challenging given the dimensionality of the problem. We extend approximate DP techniques for functional data to the GDP framework.
We further propose a novel representation, face radial curves, of a 3D face as a set of functions and then utilize our proposed GDP functional data mechanism. To preserve the shape of the face while injecting noise we rely on tools from shape analysis for our novel representation of the face. We show that our method preserves the shape of the average face and injects less noise than traditional methods for the same privacy budget. Our mechanism consists of two primary components, the first is generally applicable to function value summaries (as are commonly found in nonparametric statistics or functional data analysis) while the second is general to disk-like surfaces and hence more applicable than just to human faces.
****
Han-Jia Ye,Huai-Hong Yin,De-Chuan Zhan,Wei-Lun Chao
Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later
The widespread enthusiasm for deep learning has recently expanded into the domai

n of tabular data. Recognizing that the advancement in deep tabular methods is o
ften inspired by classical methods, e.g., integration of nearest neighbors into
neural networks, we investigate whether these classical methods can be revitaliz
ed with modern techniques.
We revisit a differentiable version of K-nearest neighbors (KNN) --- Neighbour
hood Components Analysis (NCA) --- originally designed to learn a linear project
ion to capture semantic similarities between instances, and seek to gradually ad
d modern deep learning techniques on top. Surprisingly, our implementation of NC
A using SGD and without dimensionality reduction already achieves decent perform
ance on tabular data, in contrast to the results of using existing toolboxes lik
e scikit-learn. Further equipping NCA with deep representations and additional t
raining stochasticity significantly enhances its capability, being on par with t
he leading tree-based method CatBoost and outperforming existing deep tabular mo
dels in both classification and regression tasks on 300 datasets. We conclude ou
r paper by analyzing the factors behind these improvements, including loss funct
ions, prediction strategies, and deep architectures.
The code is available at https://github.com/qile2000/LAMDA-TALENT.
****

Marcello Massimo Negri,Jonathan Aellen,Volker Roth
Injective flows for star-like manifolds
Normalizing Flows (NFs) are powerful and efficient models for density estimation
. When modeling densities on manifolds, NFs can be generalized to injective flow
s but the Jacobian determinant becomes computationally prohibitive. Current appr
oaches either consider bounds on the log-likelihood or rely on some approximatio
ns of the Jacobian determinant. In contrast, we propose injective flows for star
-like manifolds and show that for such manifolds we can compute the Jacobian det
erminant exactly and efficiently. This aspect is particularly relevant for varia
tional inference settings, where no samples are available and only some unnormal
ized target is known. Among many, we showcase the relevance of modeling densitie
s on star-like manifolds in two settings. Firstly, we introduce a novel Objectiv
e Bayesian approach for penalized likelihood models by interpreting level-sets o
f the penalty as star-like manifolds. Secondly, we consider probabilistic mixing
 models and introduce a general method for variational inference by defining the
 posterior of mixture weights on the probability simplex.
****

Weiyuan Gong,Chenyi Zhang,Tongyang Li
Robustness of Quantum Algorithms for Nonconvex Optimization
In this paper, we systematically study quantum algorithms for finding an $\epsil
on$-approximate second-order stationary point (ϵ-SOSP) of a d-dimensi
onal nonconvex function, a fundamental problem in nonconvex optimization, with n
oisy zeroth- or first-order oracles as inputs. We first prove that, up to noise
of $O(\epsilon^{10}/d^5)$, perturbed accelerated gradient descent equipped with
quantum gradient estimation takes $O(\log d/\epsilon^{1.75})$ quantum queries to
 find an ϵ-SOSP. We then prove that standard perturbed gradient descent
 is robust to the noise of $O(\epsilon^6/d^4)$ and $O(\epsilon/d^{0.5+\zeta})$ f
or any $\zeta>0$ on the zeroth- and first-order oracles, respectively, which pro
vides a quantum algorithm with poly-logarithmic query complexity. Furthermore, w
e propose a stochastic gradient descent algorithm using quantum mean estimation
on the Gaussian smoothing of noisy oracles, which is robust to $O(\epsilon^{1.5}
/d)$ and $O(\epsilon/\sqrt{d})$ noise on the zeroth- and first-order oracles, re
spectively. The quantum algorithm takes $O(d^{2.5}/\epsilon^{3.5})$ and $O(d^2/\
epsilon^3)$ queries to the two oracles, giving a polynomial speedup over the cla
ssical counterparts. As a complement, we characterize the domains where quantum
algorithms can find an ϵ-SOSP with poly-logarithmic, polynomial, or exp
onential number of queries in d, or the problem is information-theoretically u
nsolvable even with an infinite number of queries. In addition, we prove an $\Om
ega(\epsilon^{-12/7})$ lower bound on ϵ for any randomized classical an
d quantum algorithm to find an ϵ-SOSP using either noisy zeroth- or fir
st-order oracles.
****

Mufei Li,Siqi Miao,Pan Li
Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation
Large Language Models (LLMs) demonstrate strong reasoning abilities but face limitations such as hallucinations and outdated knowledge. Knowledge Graph (KG)-based Retrieval-Augmented Generation (RAG) addresses these issues by grounding LLM outputs in structured external knowledge from KGs. However, current KG-based RAG frameworks still struggle to optimize the trade-off between retrieval effectiveness and efficiency in identifying a suitable amount of relevant graph information for the LLM to digest. We introduce SubgraphRAG, extending the KG-based RAG framework that retrieves subgraphs and leverages LLMs for reasoning and answer prediction. Our approach innovatively integrates a lightweight multilayer perceptron (MLP) with a parallel triple-scoring mechanism for efficient and flexible subgraph retrieval while encoding directional structural distances to enhance retrieval effectiveness. The size of retrieved subgraphs can be flexibly adjusted to match the query's needs and the downstream LLM's capabilities. This design strikes a balance between model complexity and reasoning power, enabling scalable and generalizable retrieval processes. Notably, based on our retrieved subgraphs, smaller LLMs like Llama3.1-8B-Instruct deliver competitive results with explainable reasoning, while larger models like GPT-4o achieve state-of-the-art accuracy compared with previous baselines—all without fine-tuning. Extensive evaluations on the WebQSP and CWQ benchmarks highlight SubgraphRAG's strengths in efficiency, accuracy, and reliability by reducing hallucinations and improving response grounding.
****
Yuheng Jia,Jianhong Cheng,Hui LIU,Junhui Hou
Towards Calibrated Deep Clustering Network
Deep clustering has exhibited remarkable performance; however, the over confidence problem, i.e., the estimated confidence for a sample belonging to a particular cluster greatly exceeds its actual prediction accuracy, has been over looked in prior research. To tackle this critical issue, we pioneer the development of a calibrated deep clustering framework. Specifically, we propose a novel dual head (calibration head and clustering head) deep clustering model that can effectively calibrate the estimated confidence and the actual accuracy. The calibration head adjusts the overconfident predictions of the clustering head, generating prediction confidence that matches the model learning status. Then, the clustering head dynamically selects reliable high-confidence samples estimated by the calibration head for pseudo-label self-training. Additionally, we introduce an effective network initialization strategy that enhances both training speed and network robustness. The effectiveness of the proposed calibration approach and initialization strategy are both endorsed with solid theoretical guarantees. Extensive experiments demonstrate the proposed calibrated deep clustering model not only surpasses the state-of-the-art deep clustering methods by 5× on average in terms of expected calibration error, but also significantly outperforms them in terms of clustering accuracy. The code is available at https://github.com/ChengJianH/CDC.
****
Vighnesh Subramaniam,Yilun Du,Joshua B. Tenenbaum,Antonio Torralba,Shuang Li,Igor Mordatch
Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains
Large language models (LLMs) have achieved remarkable performance in recent years but are fundamentally limited by the underlying training data. To improve models beyond the training data, recent works have explored how LLMs can be used to generate synthetic data for autonomous self-improvement. However, successive steps of self-improvement can reach a point of diminishing returns. In this work, we propose a complementary approach towards self-improvement where finetuning is applied to a multiagent society of language models. A group of language models, all starting from the same base model, are independently specialized by updating each one using data generated through multiagent interactions among the models. By training each model on independent sets of data, we illustrate how this appr

oach enables specialization across models and diversification over the set of mo
dels. As a result, our overall system is able to preserve diverse reasoning chai
ns and autonomously improve over many more rounds of fine-tuning than single-age
nt self-improvement methods. We quantitatively illustrate the efficacy of the ap
proach across a wide suite of reasoning tasks.
****
Eslam Mohamed BAKR,Liangbing Zhao,Vincent Tao Hu,Matthieu Cord,Patrick Perez,Moh
amed Elhoseiny
ToddlerDiffusion: Interactive Structured Image Generation with Cascaded Schrödin
ger Bridge
Diffusion models break down the challenging task of generating data from high-di
mensional distributions into a series of easier denoising steps. Inspired by thi
s paradigm, we propose a novel approach that extends the diffusion framework int
o modality space, decomposing the complex task of RGB image generation into simp
ler, interpretable stages. Our method, termed {\papernameAbbrev}, cascades modal
ity-specific models, each responsible for generating an intermediate representat
ion, such as contours, palettes, and detailed textures, ultimately culminating i
n a high-quality RGB image.
Instead of relying on the naive LDM concatenation conditioning mechanism to conn
ect the different stages together, we employ Schr\"odinger Bridge to determine t
he optimal transport between different modalities.
Although employing a cascaded pipeline introduces more stages, which could lead
to a more complex architecture, each stage is meticulously formulated for effici
ency and accuracy, surpassing Stable-Diffusion (LDM) performance.
Modality composition not only enhances overall performance but enables emerging
proprieties such as consistent editing, interaction capabilities, high-level int
erpretability, and faster convergence and sampling rate.
Extensive experiments on diverse datasets, including LSUN-Churches, ImageNet, Ce
lebHQ, and LAION-Art, demonstrate the efficacy of our approach, consistently out
performing state-of-the-art methods.
For instance, {\papernameAbbrev} achieves notable efficiency, matching LDM perfo
rmance on LSUN-Churches while operating 2$\times$ faster with a 3$\times$ smalle
r architecture.
The project website is available at:
\href{https://toddlerdiffusion.github.io/website/}{$https://toddlerdiffusion.git
hub.io/website/$}
****
Jiajun Chai,Sicheng Li,Yuqian Fu,Dongbin Zhao,Yuanheng Zhu
Empowering LLM Agents with Zero-Shot Optimal Decision-Making through Q-learning
Large language models (LLMs) are trained on extensive text data to gain general
comprehension capability. Current LLM agents leverage this ability to make zero-
 or few-shot decisions without reinforcement learning (RL) but fail in making op
timal decisions, as LLMs inherently perform next-token prediction rather than ma
ximizing rewards. In contrast, agents trained via RL could make optimal decision
s but require extensive environmental interaction. In this work, we develop an a
lgorithm that combines the zero-shot capabilities of LLMs with the optimal decis
ion-making of RL, referred to as the Model-based LLM Agent with Q-Learning (MLAQ
). MLAQ employs Q-learning to derive optimal policies from transitions within me
mory. However, unlike RL agents that collect data from environmental interaction
s, MLAQ constructs an imagination space fully based on LLM to perform imaginary
interactions for deriving zero-shot policies. Our proposed UCB variant generates
 high-quality imaginary data through interactions with the LLM-based world model
, balancing exploration and exploitation while ensuring a sub-linear regret boun
d. Additionally, MLAQ incorporates a mixed-examination mechanism to filter out i
ncorrect data. We evaluate MLAQ in benchmarks that present significant challenge
s for existing LLM agents. Results show that MLAQ achieves a optimal rate of ove
r 90\% in tasks where other methods struggle to succeed. Additional experiments
are conducted to reach the conclusion that introducing model-based RL into LLM a
gents shows significant potential to improve optimal decision-making ability. Ou
r interactive website is available at http://mlaq.site.

****
Kenshi Abe,Mitsuki Sakamoto,Kaito Ariu,Atsushi Iwasaki
Boosting Perturbed Gradient Ascent for Last-Iterate Convergence in Games
This paper presents a payoff perturbation technique, introducing a strong convex ity to players' payoff functions in games. This technique is specifically design ed for first-order methods to achieve last-iterate convergence in games where th e gradient of the payoff functions is monotone in the strategy profile space, po tentially containing additive noise. Although perturbation is known to facilitat e the convergence of learning algorithms, the magnitude of perturbation requires careful adjustment to ensure last-iterate convergence. Previous studies have pr oposed a scheme in which the magnitude is determined by the distance from a peri odically re-initialized anchoring or reference strategy. Building upon this, we propose Gradient Ascent with Boosting Payoff Perturbation, which incorporates a novel perturbation into the underlying payoff function, maintaining the periodic ally re-initializing anchoring strategy scheme. This innovation empowers us to p rovide faster last-iterate convergence rates against the existing payoff perturb ed algorithms, even in the presence of additive noise.
****
Anna Vettoruzzo,Lorenzo Braccaioli,Joaquin Vanschoren,Marlena Nowaczyk
Unsupervised Meta-Learning via In-Context Learning
Unsupervised meta-learning aims to learn feature representations from unsupervis ed datasets that can transfer to downstream tasks with limited labeled data.
In this paper, we propose a novel approach to unsupervised meta-learning that le verages the generalization abilities of in-context learning observed in transfor mer architectures. Our method reframes meta-learning as a sequence modeling prob lem, enabling the transformer encoder to learn task context from support images and utilize it to predict query images.
At the core of our approach lies the creation of diverse tasks generated using a combination of data augmentations and a mixing strategy that challenges the mod el during training while fostering generalization to unseen tasks at test time.
Experimental results on benchmark datasets showcase the superiority of our appro ach over existing unsupervised meta-learning baselines, establishing it as the n ew state-of-the-art. Remarkably, our method achieves competitive results with su pervised and self-supervised approaches, underscoring its efficacy in leveraging generalization over memorization.
****
Yuqi Lin,Hengjia Li,Wenqi Shao,Zheng Yang,Jun Zhao,Xiaofei He,Ping Luo,Kaipeng Z hang
SAMRefiner: Taming Segment Anything Model for Universal Mask Refinement
In this paper, we explore a principal way to enhance the quality of widely pre-e xisting coarse masks, enabling them to serve as reliable training data for segme ntation models to reduce the annotation cost. In contrast to prior refinement te chniques that are tailored to specific models or tasks in a close-world manner, we propose SAMRefiner, a universal and efficient approach by adapting SAM to the mask refinement task. The core technique of our model is the noise-tolerant pro mpting scheme. Specifically, we introduce a multi-prompt excavation strategy to mine diverse input prompts for SAM (\ie, distance-guided points, context-aware e lastic bounding boxes, and Gaussian-style masks) from initial coarse masks. Thes e prompts can collaborate with each other to mitigate the effect of defects in c oarse masks. In particular, considering the difficulty of SAM to handle the mult i-object case in semantic segmentation, we introduce a split-then-merge (STM) pi peline. Additionally, we extend our method to SAMRefiner++ by introducing an add itional IoU adaption step to further boost the performance of the generic SAMRef iner on the target dataset. This step is self-boosted and requires no additional annotation. The proposed framework is versatile and can flexibly cooperate with existing segmentation methods. We evaluate our mask framework on a wide range o f benchmarks under different settings, demonstrating better accuracy and efficie ncy. SAMRefiner holds significant potential to expedite the evolution of refinem ent tools. Our code is available at https://github.com/linyq2117/SAMRefiner.
****

Long Peng,Wenbo Li,Renjing Pei,Jingjing Ren,Jiaqi Xu,Yang Wang,Yang Cao,Zheng-Jun Zha
Towards Realistic Data Generation for Real-World Super-Resolution
Existing image super-resolution (SR) techniques often fail to generalize effectively in complex real-world settings due to the significant divergence between training data and practical scenarios. To address this challenge, previous efforts have either manually simulated intricate physical-based degradations or utilized learning-based techniques, yet these approaches remain inadequate for producing large-scale, realistic, and diverse data simultaneously. In this paper, we introduce a novel Realistic Decoupled Data Generator (RealDGen), an unsupervised learning data generation framework designed for real-world super-resolution. We meticulously develop content and degradation extraction strategies, which are integrated into a novel content-degradation decoupled diffusion model to create realistic low-resolution images from unpaired real LR and HR images. Extensive experiments demonstrate that RealDGen excels in generating large-scale, high-quality paired data that mirrors real-world degradations, significantly advancing the performance of popular SR models on various real-world benchmarks.
****
Hailey Joren,Jianyi Zhang,Chun-Sung Ferng,Da-Cheng Juan,Ankur Taly,Cyrus Rashtchian
Sufficient Context: A New Lens on Retrieval Augmented Generation Systems
Augmenting LLMs with context leads to improved performance across many applications. Despite much research on Retrieval Augmented Generation (RAG) systems, an open question is whether errors arise because LLMs fail to utilize the context from retrieval or the context itself is insufficient to answer the query. To shed light on this, we develop a new notion of sufficient context, along with a method to classify instances that have enough information to answer the query. We then use sufficient context to analyze several models and datasets. By stratifying errors based on context sufficiency, we find that larger models with higher baseline performance (Gemini 1.5 Pro, GPT 4o, Claude 3.5) excel at answering queries when the context is sufficient, but often output incorrect answers instead of abstaining when the context is not. On the other hand, smaller models with lower baseline performance (Llama 3.1, Mistral 3, Gemma 2) hallucinate or abstain often, even with sufficient context. We further categorize cases when the context is useful, and improves accuracy, even though it does not fully answer the query and the model errs without the context. Building on our findings, we explore ways to reduce hallucinations in RAG systems, including a new selective generation method that leverages sufficient context information for guided abstention. Our method improves the fraction of correct answers among times where the model responds by 2--10% for Gemini, GPT, and Gemma. Code for our selective generation method and the prompts used in our autorater analysis are available on our [github](https://github.com/hljoren/sufficientcontext).
****
Andi Han,Wei Huang,Yuan Cao,Difan Zou
On the Feature Learning in Diffusion Models
The predominant success of diffusion models in generative modeling has spurred significant interest in understanding their theoretical foundations. In this work, we propose a feature learning framework aimed at analyzing and comparing the training dynamics of diffusion models with those of traditional classification models. Our theoretical analysis demonstrates that diffusion models, due to the denoising objective, are encouraged to learn more balanced and comprehensive representations of the data. In contrast, neural networks with a similar architecture trained for classification tend to prioritize learning specific patterns in the data, often focusing on easy-to-learn components. To support these theoretical insights, we conduct several experiments on both synthetic and real-world datasets, which empirically validate our findings and highlight the distinct feature learning dynamics in diffusion models compared to classification.
****
Isaac Lin,Tianye Wang,Shang Gao,Tang Shiming,Tai Sing Lee
Self-Attention-Based Contextual Modulation Improves Neural System Identification

Convolutional neural networks (CNNs) have been shown to be state-of-the-art mode ls for visual cortical neurons. Cortical neurons in the primary visual cortex ar e sensitive to contextual information mediated by extensive horizontal and feedb ack connections. Standard CNNs integrate global contextual information to model contextual modulation via two mechanisms: successive convolutions and a fully co nnected readout layer. In this paper, we find that self-attention (SA), an imple mentation of non-local network mechanisms, can improve neural response predictio ns over parameter-matched CNNs in two key metrics: tuning curve correlation and peak tuning. We introduce peak tuning as a metric to evaluate a model's ability to capture a neuron's top feature preference. We factorize networks to assess ea ch context mechanism, revealing that information in the local receptive field is most important for modeling overall tuning, but surround information is critica lly necessary for characterizing the tuning peak. We find that self-attention ca n replace posterior spatial-integration convolutions when learned incrementally, and is further enhanced in the presence of a fully connected readout layer, sug gesting that the two context mechanisms are complementary. Finally, we find that decomposing receptive field learning and contextual modulation learning in an i ncremental manner may be an effective and robust mechanism for learning surround -center interactions.
****

Robert Ganian,Liana Khazaliya,Fionn Mc Inerney,Mathis Rocton
The Computational Complexity of Positive Non-Clashing Teaching in Graphs
We study the classical and parameterized complexity of computing the positive no n-clashing teaching dimension of a set of concepts, that is, the smallest number of examples per concept required to successfully teach an intelligent learner u nder the considered, previously established model. For any class of concepts, it is known that this problem can be effortlessly transferred to the setting of ba lls in a graph G. We establish (1) the NP-hardness of the problem even when re stricted to instances with positive non-clashing teaching dimension $k=2$ and wh ere all balls in the graph are present, (2) near-tight running time upper and lo wer bounds for the problem on general graphs, (3) fixed-parameter tractability w hen parameterized by the vertex integrity of G, and (4) a lower bound excludin g fixed-parameter tractability when parameterized by the feedback vertex number and pathwidth of G, even when combined with k.
Our results provide a nearly complete understanding of the complexity landscape of computing the positive non-clashing teaching dimension and answer open questi ons from the literature.
****

Xiang Liu,Bin Chen,Zimo Liu,Yaowei Wang,Shu-Tao Xia
An Exploration with Entropy Constrained 3D Gaussians for 2D Video Compression
3D Gaussian Splatting (3DGS) has witnessed its rapid development in novel view s ynthesis, which attains high quality reconstruction and real-time rendering. At the same time, there is still a gap before implicit neural representation (INR) can become a practical compressor due to the lack of stream decoding and real-t ime frame reconstruction on consumer-grade hardware. It remains a question wheth er the fast rendering and partial parameter decoding characteristics of 3DGS are applicable to video compression. To address these challenges, we propose a Toas t-like Sliding Window (TSW) orthographic projection for converting any 3D Gaussi an model into a video representation model. This method efficiently represents v ideo by leveraging temporal redundancy through a sliding window approach. Additi onally, the converted model is inherently stream-decodable and offers a higher r endering frame rate compared to INR methods. Building on TSW, we introduce an en d-to-end trainable video compression method, GSVC, which employs deformable Gaus sian representation and optical flow guidance to capture dynamic content in vide os. Experimental results demonstrate that our method effectively transforms a 3D Gaussian model into a practical video compressor. GSVC further achieves better rate-distortion performance than NeRV on the UVG dataset, while achieving highe r frame reconstruction speed (+30%~40% fps) and stream decoding. Code is availab le at [Github](https://github.com/actcwlf/GSVC)
****

Haomiao Xiong,Zongxin Yang,Jiazuo Yu,Yunzhi Zhuge,Lu Zhang,Jiawen Zhu,Huchuan Lu
Streaming Video Understanding and Multi-round Interaction with Memory-enhanced K nowledge
Recent advances in Large Language Models (LLMs) have enabled the development of Video-LLMs, advancing multimodal learning by bridging video data with language t asks. However, current video understanding models struggle with processing long video sequences, supporting multi-turn dialogues, and adapting to real-world dyn amic scenarios. To address these issues, we propose StreamChat, a training-free framework for streaming video reasoning and conversational interaction. StreamCh at leverages a novel hierarchical memory system to efficiently process and compr ess video features over extended sequences, enabling real-time, multi-turn dialo gue. Our framework incorporates a parallel system scheduling strategy that enhan ces processing speed and reduces latency, ensuring robust performance in real-wo rld applications. Furthermore, we introduce StreamBench, a versatile benchmark t hat evaluates streaming video understanding across diverse media types and inter active scenarios, including multi-turn interactions and complex reasoning tasks.
Extensive evaluations on StreamBench and other public benchmarks demonstrate t hat StreamChat significantly outperforms existing
state-of-the-art models in terms of accuracy and response times, confirming its effectiveness for streaming video understanding. Code is available at StreamChat .
****
Jaskirat Singh,Junshen K Chen,Jonas K Kohler,Michael F Cohen
Storybooth: Training-Free Multi-Subject Consistency for Improved Visual Storytel ling
Consistent text-to-image generation depicting the *same* subjects across differe nt images has gained significant recent attention due to its widespread applicat ions in the fields of visual-storytelling and multiple-shot video generation. W hile remarkable, existing methods often require costly finetuning for each subje ct and struggle to maintain consistency across multiple characters. In this work , we first analyse the reason for these limitations. Our exploration reveals tha t the primary-issue stems from *self-attention leakage*, which is exacerbated wh en trying to ensure consistency across multiple-characters. Motivated by these f indings, we next propose a simple yet effective *training and optimization-free approach* for improving multiple-character consistency. In particular, we first leverage multi-modal *chain-of-thought* reasoning in order to *apriori* localize the different subjects across the storyboard frames. The final storyboard image s are then generated using a modified diffusion model which includes *1) a boun ded cross-attention layer* for ensuring adherence to the initially predicted lay out, and *2) a bounded cross-frame self-attention layer* for reducing inter-cha racter attention leakage. Furthermore, we also propose a novel *cross-frame toke n-merging layer* which allows for improved fine-grain consistency for the storyb oard characters.
Experimental analysis reveals that proposed approach is not only $\times 30$ fa ster than prior training-based methods (*eg, textual inversion, dreambooth-lora*) but also surpasses the prior *state-of-the-art*, exhibiting improved multi-cha racter consistency and text-to-image alignment performance.
****
Haoxin Lin,Yu-Yan Xu,Yihao Sun,Zhilong Zhang,Yi-Chen Li,Chengxing Jia,Junyin Ye, Jiaji Zhang,Yang Yu
Any-step Dynamics Model Improves Future Predictions for Online and Offline Reinf orcement Learning
Model-based methods in reinforcement learning offer a promising approach to enha nce data efficiency by facilitating policy exploration within a dynamics model. However, accurately predicting sequential steps in the dynamics model remains a challenge due to the bootstrapping prediction, which attributes the next state t o the prediction of the current state. This leads to accumulated errors during m odel roll-out. In this paper, we propose the Any-step Dynamics Model (ADM) to mi tigate the compounding error by reducing bootstrapping prediction to direct pred iction. ADM allows for the use of variable-length plans as inputs for predicting

future states without frequent bootstrapping. We design two algorithms, ADMPO-ON and ADMPO-OFF, which apply ADM in online and offline model-based frameworks, respectively. In the online setting, ADMPO-ON demonstrates improved sample efficiency compared to previous state-of-the-art methods. In the offline setting, ADMPO-OFF not only demonstrates superior performance compared to recent state-of-the-art offline approaches but also offers better quantification of model uncertainty using only a single ADM.
****
Qi Zhang,Yi Zhou,Simon Khan,Ashley Prater-Bennette,Lixin Shen,Shaofeng Zou
Revisiting Large-Scale Non-convex Distributionally Robust Optimization
Distributionally robust optimization (DRO) is a powerful technique to train robust machine learning models that perform well under distribution shifts. Compared with empirical risk minimization (ERM), DRO optimizes the expected loss under the worst-case distribution in
an uncertainty set of distributions. This paper revisits the important problem of DRO with non-convex smooth loss functions. For this problem, Jin et al. (2021) showed that its dual problem is generalized (L_0, L_1)-smooth condition and gradient noise satisfies the affine variance condition, designed an algorithm of mini-batch normalized gradient descent with momentum, and proved its convergence and complexity. In this paper, we show that the dual problem and the gradient noise satisfy simpler yet more precise partially generalized smoothness condition and partially affine variance condition by studying the optimization variable and dual variable separately, which further yields much simpler algorithm design and convergence analysis. We develop a double stochastic gradient descent with clipping (D-SGD-C) algorithm that converges to an ϵ-stationary point with $\mathcal O(\epsilon^{-4})$ gradient complexity, which matches with results in Jin et al. (2021). Our algorithm does not need to use momentum, and the proof is much simpler, thanks to the more precise characterization of partially generalized smoothness and partially affine variance noise. We further design a variance-reduced method that achieves a lower gradient complexity of $\mathcal O(\epsilon^{-3})$. Our theoretical results and insights are further verified numerically on a number of tasks, and our algorithms outperform the existing DRO method (Jin et al., 2021).
****
Orr Zohar,Xiaohan Wang,Yonatan Bitton,Idan Szpektor,Serena Yeung-Levy
Video-STaR: Self-Training Enables Video Instruction Tuning with Any Supervision
The performance and reasoning capabilities of Large Multi-modal Models (LMMs) is dependent on the size and quality of their training datasets. However, collecting datasets that support chain-of-thought instruction tuning is highly challenging. Existing video instruction tuning datasets are often derived by prompting large language models with video captions to generate question-answer pairs, which makes them predominantly descriptive rather than reasoning-focused.
Meanwhile, many labeled video datasets with diverse labels and supervision exist -- however, we find that their integration into LMMs is non-trivial.
Herein, we present $\underline{\text{Video}}$ $\underline{\text{S}}\text{elf}$-$\underline{\text{T}}\text{raining}$ with $\underline{\text{a}}\text{ugmented}$ $\underline{\text{R}}\text{easoning}$ (Video-STaR), the first self-training approach for video instruction tuning.
Video-STaR allows the utilization of *any* labeled video dataset for video instruction tuning.
In Video-STaR, an LMM cycles between instruction generation and finetuning, which we show (I) improves general video understanding and (II) adapts LMMs to novel downstream tasks with existing supervision.
During instruction generation, an LMM is prompted to propose an answer. The answers are then filtered only to those that contain the original video labels, and the LMM is then re-trained on the generated dataset.
By training exclusively on generated answers containing the correct video labels, Video-STaR leverages these existing labels as weak supervision for video instruction tuning.
Our results demonstrate that Video-STaR-augmented LMMs achieve notable improveme

nts in (I) general Video QA, where TempCompass performance improved by 6.1%, *and* (II) downstream tasks, with a 9.9% increase in Kinetics700-QA accuracy and a 4.0% improvement in action quality assessment on FineDiving, while also exhibiting better interpretability.
****
Zongzhe Xu,Ritvik Gupta,Wenduo Cheng,Alexander Shen,Junhong Shen,Ameet Talwalkar,Mikhail Khodak
Specialized Foundation Models Struggle to Beat Supervised Baselines
Following its success for vision and text, the "foundation model" (FM) paradigm—pretraining large models on massive data, then fine-tuning on target tasks—has rapidly expanded to domains in the sciences, engineering, healthcare, and beyond. Has this achieved what the original FMs accomplished, i.e. the supplanting of traditional supervised learning in their domains? To answer we look at three modalities—genomics, satellite imaging, and time series—with multiple recent FMs and compare them to a standard supervised learning workflow: model development, hyperparameter tuning, and training, all using only data from the target task. Across these three specialized domains, we find that it is consistently possible to train simple supervised models—no more complicated than a lightly modified wide ResNet or UNet—that match or even outperform the latest foundation models. Our work demonstrates that the benefits of large-scale pretraining have yet to be realized in many specialized areas, reinforces the need to compare new FMs to strong, well-tuned baselines, and introduces two new, easy-to-use, open-source, and automated workflows for doing so.
****
Sidak Pal Singh,Bobby He,Thomas Hofmann,Bernhard Schölkopf
The Directionality of Optimization Trajectories in Neural Networks
The regularity or implicit bias in neural network optimization has been typically studied via the parameter norms or the landscape curvature, often overlooking the trajectory leading to these parameters. However, properties of the trajectory --- particularly its directionality --- capture critical aspects of how gradient descent navigates the landscape to converge to a solution. In this work, we introduce the notion of a Trajectory Map and derive natural complexity measures that highlight the directional characteristics of optimization trajectories. Our comprehensive analysis across vision and language modeling tasks reveals that (a) the trajectory's directionality at the macro-level saturates by the initial phase of training, wherein weight decay and momentum play a crucial but understated role; and (b) in subsequent training, trajectory directionality manifests in micro-level behaviors, such as oscillations, for which we also provide a theoretical analysis. This implies that neural optimization trajectories have, overall, a more linear form than zig-zaggy, as evident by high directional similarity, especially towards the end. To further hone this point, we show that when the trajectory direction gathers such an inertia, optimization proceeds largely unaltered even if the network is severely decapacitated (by freezing >99% of the parameters), --- thereby demonstrating the potential for significant computational and resource savings without compromising performance.
****
Jiankang Chen,Tianke Zhang,Changyi Liu,Haojie Ding,Yaya Shi,cheng.feng,Huihui Xiao,Bin Wen,Fan Yang,Tingting Gao,Di ZHANG
TaskGalaxy: Scaling Multi-modal Instruction Fine-tuning with Tens of Thousands Vision Task Types
Multimodal visual language models are gaining prominence in open-world applications, driven by advancements in model architectures, training techniques, and high-quality data. However, their performance is often limited by insufficient task-specific data, leading to poor generalization and biased outputs. Existing efforts to increase task diversity in fine-tuning datasets are hindered by the labor-intensive process of manual task labeling, which typically produces only a few hundred task types. To address this, we propose TaskGalaxy, a large-scale multimodal instruction fine-tuning dataset comprising 19,227 hierarchical task types and 413,648 samples. TaskGalaxy utilizes GPT-4o to enrich task diversity by expanding from a small set of manually defined tasks, with CLIP and GPT-4o filtering

those that best match open-source images, and generating relevant question-answer pairs. Multiple models are employed to ensure sample quality. This automated process enhances both task diversity and data quality, reducing manual intervention. Incorporating TaskGalaxy into LLaVA-v1.5 and InternVL-Chat-v1.0 models shows substantial performance improvements across 16 benchmarks, demonstrating the critical importance of task diversity. TaskGalaxy is publicly released at https://github.com/Kwai-YuanQi/TaskGalaxy.
****
Keyi Shen,Jiangwei Yu,Jose Barreiros,Huan Zhang,Yunzhu Li
BaB-ND: Long-Horizon Motion Planning with Branch-and-Bound and Neural Dynamics
Neural-network-based dynamics models learned from observational data have shown strong predictive capabilities for scene dynamics in robotic manipulation tasks. However, their inherent non-linearity presents significant challenges for effective planning. Current planning methods, often dependent on extensive sampling or local gradient descent, struggle with long-horizon motion planning tasks involving complex contact events.
In this paper, we present a GPU-accelerated branch-and-bound (BaB) framework for motion planning in manipulation tasks that require trajectory optimization over neural dynamics models. Our approach employs a specialized branching heuristic to divide the search space into sub-domains and applies a modified bound propagation method, inspired by the state-of-the-art neural network verifier α,β-CROWN, to efficiently estimate objective bounds within these sub-domains. The branching process guides planning effectively, while the bounding process strategically reduces the search space.
Our framework achieves superior planning performance, generating high-quality state-action trajectories and surpassing existing methods in challenging, contact-rich manipulation tasks such as non-prehensile planar pushing with obstacles, object sorting, and rope routing in both simulated and real-world settings. Furthermore, our framework supports various neural network architectures, ranging from simple multilayer perceptrons to advanced graph neural dynamics models, and scales efficiently with different model sizes.
****
Kacper Wyrwal,Andreas Krause,Viacheslav Borovitskiy
Residual Deep Gaussian Processes on Manifolds
We propose practical deep Gaussian process models on Riemannian manifolds, similar in spirit to residual neural networks.
With manifold-to-manifold hidden layers and an arbitrary last layer, they can model manifold- and scalar-valued functions, as well as vector fields.
We target data inherently supported on manifolds, which is too complex for shallow Gaussian processes thereon.
For example, while the latter perform well on high-altitude wind data, they struggle with the more intricate, nonstationary patterns at low altitudes.
Our models significantly improve performance in these settings, enhancing prediction quality and uncertainty calibration, and remain robust to overfitting, reverting to shallow models when additional complexity is unneeded.
We further showcase our models on Bayesian optimisation problems on manifolds, using stylised examples motivated by robotics, and obtain substantial improvements in later stages of the optimisation process.
Finally, we show our models to have potential for speeding up inference for non-manifold data, when, and if, it can be mapped to a proxy manifold well enough.
****
Jiafei Duan,Wilbert Pumacay,Nishanth Kumar,Yi Ru Wang,Shulin Tian,Wentao Yuan,Ranjay Krishna,Dieter Fox,Ajay Mandlekar,Yijie Guo
AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation
Robotic manipulation in open-world settings requires not only task execution but also the ability to detect and learn from failures. While recent advances in vision-language models (VLMs) and large language models (LLMs) have improved robots' spatial reasoning and problem-solving abilities, they still struggle with failure recognition, limiting their real-world applicability. We introduce AHA, an

open-source VLM designed to detect and reason about failures in robotic manipula
tion using natural language. By framing failure detection as a free-form reasoni
ng task, AHA identifies failures and provides detailed, adaptable explanations a
cross different robots, tasks, and environments. We fine-tuned AHA using FailGen
, a scalable framework that generates the first large-scale dataset of robotic f
ailure trajectories, the AHA dataset. FailGen achieves this by procedurally pert
urbing successful demonstrations from simulation. Despite being trained solely o
n the AHA dataset, AHA generalizes effectively to real-world failure datasets, r
obotic systems, and unseen tasks. It surpasses the second-best model (GPT-4o in-
context learning) by 10.3% and exceeds the average performance of six compared m
odels including five state-of-the-art VLMs by 35.3% across multiple metrics and
datasets. We integrate AHA into three manipulation frameworks that utilize LLMs/
VLMs for reinforcement learning, task and motion planning, and zero-shot traject
ory generation. AHA’s failure feedback enhances these policies' performances by
refining dense reward functions, optimizing task planning, and improving sub-tas
k verification, boosting task success rates by an average of 21.4% across all th
ree tasks compared to GPT-4 models. Project page: https://aha-vlm.github.io
****

Nathaniel Weir,Bhavana Dalvi Mishra,Orion Weller,Oyvind Tafjord,Sam Hornstein,Al
exander Sabol,Peter Jansen,Benjamin Van Durme,Peter Clark

From Models to Microtheories: Distilling a Model's Topical Knowledge for Grounde
d Question-Answering

Recent reasoning methods (e.g., chain-of-thought) help users understand how lang
uage models (LMs) answer a single question, but they do little to reveal the LM’
s overall understanding, or “theory,” about the question’s topic, making it stil
l hard to trust the model. Our goal is to materialize such theories - here calle
d microtheories (a linguistic analog of logical microtheories) - as a set of sen
tences encapsulating an LM’s core knowledge about a topic. These statements syst
ematically work together to entail answers to a set of questions to both engende
r trust and improve performance. Our approach is to first populate a knowledge s
tore with (model-generated) sentences that entail answers to training questions,
 and then distill those down to a core microtheory which is concise, general, an
d non-redundant. We show that, when added to a general corpus (e.g., Wikipedia),
 microtheories can supply critical information not necessarily present in the co
rpus, improving both a model’s ability to ground its answers to verifiable knowl
edge (i.e., show how answers are systematically entailed by documents in the cor
pus, grounding up to +8% more answers), and the accuracy of those grounded answe
rs (up to +8% absolute). We also show that, in a human evaluation in the medical
 domain, our distilled microtheories contain a significantly higher concentratio
n of topically critical facts than the non-distilled knowledge store. Finally, w
e show we can quantify the coverage of a microtheory for a topic (characterized
by a dataset) using a notion of p-relevance. Together, these suggest that microt
heories are an efficient distillation of an LM’s topic-relevant knowledge, that
they can usefully augment existing corpora, and can provide both performance gai
ns and an interpretable, verifiable window into the model’s knowledge of a topic
.
****

Kaishen Wang,Hengrui Gu,Meijun Gao,Kaixiong Zhou

DAMO: Decoding by Accumulating Activations Momentum for Mitigating Hallucination
s in Vision-Language Models

Large Vision-Language Models (VLMs) exhibit significant potential in multimodal
tasks but often struggle with hallucinations—responses that are plausible yet vi
sually ungrounded. In this work, we investigate the layer-wise prediction tenden
cies of VLMs and conduct an in-depth analysis of their decoding mechanism. We ob
serve that VLMs tend to ``overthink'' during the final stages of decoding, makin
g significant prediction shifts in the last few layers often favoring incorrect
results, which leads to a surge in hallucinative outputs. Leveraging this locali
zed pattern, we propose a novel decoding strategy inspired by the momentum analo
gy used in gradient descent-based optimizers. Our method enforces decoding consi
stency across layers in an adaptive manner during forward passes—an under-explor

ed approach in existing works. This strategy significantly improves the reliabil
ity and performance of VLMs in various multimodal tasks, while introducing only
negligible efficiency overhead.
****

Chuhan ZHANG,Chaoyang Zhu,Pingcheng Dong,Long Chen,Dong Zhang
Cyclic Contrastive Knowledge Transfer for Open-Vocabulary Object Detection
In pursuit of detecting unstinted objects that extend beyond predefined categori
es, prior arts of open-vocabulary object detection (OVD) typically resort to pre
trained vision-language models (VLMs) for base-to-novel category generalization.
However, to mitigate the misalignment between upstream image-text pretraining a
nd downstream region-level perception, additional supervisions are indispensable
, e.g., image-text pairs or pseudo annotations generated via self-training strat
egies. In this work, we propose CCKT-Det trained without any extra supervision.
The proposed framework constructs a cyclic and dynamic knowledge transfer from l
anguage queries and visual region features extracted from VLMs, which forces the
detector to closely align with the visual-semantic space of VLMs. Specifically,
1) we prefilter and inject semantic priors to guide the learning of queries, an
d 2) introduce a regional contrastive loss to improve the awareness of queries o
n novel objects. CCKT-Det can consistently improve performance as the scale of V
LMs increases, all while requiring the detector at a moderate level of computati
on overhead. Comprehensive experimental results demonstrate that our method achi
eves performance gain of +2.9% and +10.2% AP_{50} over previous state-of-the-art
s on the challenging COCO benchmark, both without and with a stronger teacher mo
del.
****

Zhaohui JIANG,Xuening Feng,Paul Weng,Yifei Zhu,Yan Song,Tianze Zhou,Yujing Hu,Ta
ngjie Lv,Changjie Fan
Reinforcement Learning from Imperfect Corrective Actions and Proxy Rewards
In practice, reinforcement learning (RL) agents are often trained with a possibl
y imperfect proxy reward function, which may lead to a human-agent alignment iss
ue (i.e., the learned policy either converges to non-optimal performance with lo
w cumulative rewards, or achieves high cumulative rewards but in an undesired ma
nner). To tackle this issue, we consider a framework where a human labeler can p
rovide additional feedback in the form of corrective actions, which expresses th
e labeler's action preferences although this feedback may possibly be imperfect
as well.
In this setting, to obtain a better-aligned policy guided by both learning signa
ls, we propose a novel value-based deep RL algorithm called **I**terative learni
ng from **Co**rrective actions and **Pro**xy rewards (ICoPro), which cycles thro
ugh three phases:
(1) Solicit sparse corrective actions from a human labeler on the agent's demons
trated trajectories;
(2) Incorporate these corrective actions into the Q-function using a margin loss
to enforce adherence to labeler's preferences;
(3) Train the agent with standard RL losses regularized with a margin loss to le
arn from proxy rewards and propagate the Q-values learned from human feedback. M
oreover, another novel design in our approach is to integrate pseudo-labels from
the target Q-network to reduce human labor and further stabilize training.
We experimentally validate our proposition on a variety of tasks (Atari games an
d autonomous driving on highway). On the one hand, using proxy rewards with diff
erent levels of imperfection, our method can better align with human and is more
sample-efficient than baseline methods. On the other hand, facing corrective ac
tions with different types of imperfection, our method can overcome the non-opti
mality of this feedback thanks to the guidance from proxy rewards.
****

Guorui Zheng,Xidong Wang,Juhao Liang,Nuo Chen,Yuping Zheng,Benyou Wang
Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Languag
e Family Experts
Adapting medical Large Language Models to local languages can reduce barriers t
o accessing healthcare services, but data scarcity remains a significant challen

ge, particularly for low-resource languages. To address this, we first construct a high-quality medical dataset and conduct analysis to ensure its quality. In order to leverage the generalization capability of multilingual LLMs to efficiently scale to more resource-constrained languages, we explore the internal information flow of LLMs from a multilingual perspective using Mixture of Experts (MoE) modularity. Technically, we propose a novel MoE routing method that employs language-specific experts and cross-lingual routing. Inspired by circuit theory, our routing analysis revealed a \textit{``Spread Out in the End``} information flow mechanism: while earlier layers concentrate cross-lingual information flow, the later layers exhibit language-specific divergence. This insight directly led to the development of the Post-MoE architecture, which applies sparse routing only in the later layers while maintaining dense others. Experimental results demonstrate that this approach enhances the generalization of multilingual models to other languages while preserving interpretability. Finally, to efficiently scale the model to 50 languages, we introduce the concept of \textit{language family} experts, drawing on linguistic priors, which enables scaling the number of languages without adding additional parameters.
****
Hen Davidov,Shai Feldman,Gil Shamai,Ron Kimmel,Yaniv Romano
Conformalized Survival Analysis for General Right-Censored Data
We develop a framework to quantify predictive uncertainty in survival analysis, providing a reliable lower predictive bound (LPB) for the true, unknown patient survival time. Recently, conformal prediction has been used to construct such valid LPBs for *type-I right-censored data*, with the guarantee that the bound holds with high probability. Crucially, under the type-I setting, the censoring time is observed for all data points. As such, informative LPBs can be constructed by framing the calibration as an estimation task with covariate shift, relying on the conditionally independent censoring assumption. This paper expands the conformal toolbox for survival analysis, with the goal of handling the ubiquitous *general right-censored setting*, in which either the censoring or survival time is observed, but not both. The key challenge here is that the calibration cannot be directly formulated as a covariate shift problem anymore. Yet, we show how to construct LPBs with distribution-free finite-sample guarantees, under the same assumptions as conformal approaches for type-I censored data. Experiments demonstrate the informativeness and validity of our methods in simulated settings and showcase their practical utility using several real-world datasets.
****
Yuanyuan Yang,Tao Xiao,Bhuvesh Kumar,Jamie Heather Morgenstern
Private Mechanism Design via Quantile Estimation
We investigate the problem of designing differentially private (DP), revenue-maximizing single item auction. Specifically, we consider broadly applicable settings in mechanism design where agents' valuation distributions are **independent**, **non-identical**, and can be either **bounded** or **unbounded**. Our goal is to design such auctions with **pure**, i.e., $(\epsilon,0)$ privacy in polynomial time.

In this paper, we propose two computationally efficient auction learning framework that achieves **pure** privacy under bounded and unbounded distribution settings. These frameworks reduces the problem of privately releasing a revenue-maximizing auction to the private estimation of pre-specified quantiles. Our solutions increase the running time by polylog factors compared to the non-private version. As an application, we show how to extend our results to the multi-round online auction setting with non-myopic bidders. To our best knowledge, this paper is the first to efficiently deliver a Myerson auction with **pure** privacy and near-optimal revenue, and the first to provide such auctions for **unbounded** distributions.
****
Antonio Emanuele Cinà,Francesco Villani,Maura Pintor,Lea Schönherr,Battista Biggio,Marcello Pelillo
σ-zero: Gradient-based Optimization of ℓ_0-norm Adversarial Examples

Evaluating the adversarial robustness of deep networks to gradient-based attacks is challenging.
While most attacks consider ℓ_2- and ℓ_∞-norm constraints to craft input perturbations, only a few investigate sparse ℓ_1- and ℓ_0-norm attacks.
In particular, ℓ_0-norm attacks remain the least studied due to the inherent complexity of optimizing over a non-convex and non-differentiable constraint.
However, evaluating adversarial robustness under these attacks could reveal weaknesses otherwise left untested with more conventional ℓ_2- and ℓ_∞-norm attacks.
In this work, we propose a novel ℓ_0-norm attack, called σ-zero, which leverages a differentiable approximation of the ℓ_0 norm to facilitate gradient-based optimization, and an adaptive projection operator to dynamically adjust the trade-off between loss minimization and perturbation sparsity.
Extensive evaluations using MNIST, CIFAR10, and ImageNet datasets, involving robust and non-robust models, show that σ-zero finds minimum ℓ_0-norm adversarial examples without requiring any time-consuming hyperparameter tuning, and that it outperforms all competing sparse attacks in terms of success rate, perturbation size, and efficiency.
****
Barak Lenz,Opher Lieber,Alan Arazi,Amir Bergman,Avshalom Manevich,Barak Peleg,Ben Aviram,Chen Almagor,Clara Fridman,Dan Padnos,Daniel Gissin,Daniel Jannai,Dor Muhlgay,Dor Zimberg,Edden M. Gerber,Elad Dolev,Eran Krakovsky,Erez Safahi,Erez Schwartz,Gal Cohen,Gal Shachaf,Haim Rozenblum,Hofit Bata,Ido Blass,Inbal Magar,Itay Dalmedigos,Jhonathan Osin,Julie Fadlon,Maria Rozman,Matan Danos,Michael Gokhman,Mor Zusman,Naama Gidron,Nir Ratner,Noam Gat,Noam Rozen,Oded Fried,Ohad Leshno,Omer Antverg,Omri Abend,Or Dagan,Orit Cohavi,Raz Alon,Ro'i Belson,Roi Cohen,Rom Gilad,Roman Glozman,Shahar Lev,Shai Shalev-Shwartz,Shaked Haim Meirom,Tal Delbari,Tal Ness,Tomer Asida,Tom Ben Gal,Tom Braude,Uriya Pumerantz,Josh Cohen,Yonatan Belinkov,Yuval Globerson,Yuval Peleg Levy,Yoav Shoham
Jamba: Hybrid Transformer-Mamba Language Models
We present Jamba, a novel hybrid Transformer-Mamba mixture-of-experts (MoE) architecture. Jamba interleaves blocks of Transformer and Mamba layers, enjoying the benefits of both model families. MoE is added in some of these layers to increase model capacity while keeping active parameter usage manageable. This flexible architecture allows resource- and objective-specific configurations. We implement two configurations: Jamba-1.5-Large, with 94B active parameters, and Jamba-1.5-mini, with 12B active parameters. Built at large scale, Jamba models provide high throughput and small memory footprint compared to vanilla Transformers, especially at long-context tasks, with an effective context length of 256K tokens, the largest amongst open-weight models. At the same time, they are also competitive on standard language modeling and chatbot benchmarks. We study various architectural decisions, such as how to combine Transformer and Mamba layers, and how to mix experts, and show that some of them are crucial in large scale modeling.
To support cost-effective inference, we introduce ExpertsInt8, a novel quantization technique that allows fitting Jamba-1.5-Large on a machine with 8 80GB GPUs when processing 256K-token contexts without loss of quality. We also describe several interesting properties of this architecture that the training and evaluation of Jamba have revealed. The model weights are publicly available.
****
Zhaojing Wen,Qiulin Zhang,Yuan Zhang,Rudan Chen,Xichao Yang,Di Xie,Jiang Zhu
From Decoupling to Adaptive Transformation: a Wider Optimization Space for PTQ
Post-Training low-bit Quantization (PTQ) is useful to accelerate DNNs due to its high efficiency, the current SOTAs of which mostly adopt feature reconstruction with self-distillation finetuning. However, when bitwidth goes to be extremely low, we find the current reconstruction optimization space is not optimal. Considering all possible parameters and the ignored fact that integer weight can be obtained early before actual inference, we thoroughly explore different optimization space by quant-step decoupling, where a wider PTQ optimization space, which consistently makes a better optimum, is found out. Based on these, we propose an

Adaptive Quantization Transformation (AdaQTransform) for PTQ reconstruction, which makes the quantized output feature better fit the FP32 counterpart with adaptive per-channel transformation, thus achieves lower feature reconstruction error. In addition, it incurs negligible extra finetuning cost and no extra inference cost. Based on AdaQTransform, for the first time, we build a general quantization setting paradigm subsuming current PTQs, QATs and other potential forms. Experiments demonstrate AdaQTransform expands the optimization space for PTQ and helps current PTQs find a better optimum over CNNs, ViTs, LLMs and image super-resolution networks, e.g., it improves NWQ by 5.7% on ImageNet for W2A2-MobileNet-v2.
****
Haoge Deng,Ting Pan,Haiwen Diao,Zhengxiong Luo,Yufeng Cui,Huchuan Lu,Shiguang Shan,Yonggang Qi,Xinlong Wang
Autoregressive Video Generation without Vector Quantization
This paper presents a novel approach that enables autoregressive video generation with high efficiency. We propose to reformulate the video generation problem as a non-quantized autoregressive modeling of temporal frame-by-frame prediction and spatial set-by-set prediction. Unlike raster-scan prediction in prior autoregressive models or joint distribution modeling of fixed-length tokens in diffusion models, our approach maintains the causal property of GPT-style models for flexible in-context capabilities, while leveraging bidirectional modeling within individual frames for efficiency. With the proposed approach, we train a novel video autoregressive model without vector quantization, termed NOVA. Our results demonstrate that NOVA surpasses prior autoregressive video models in data efficiency, inference speed, visual fidelity, and video fluency, even with a much smaller model capacity, i.e., 0.6B parameters. NOVA also outperforms state-of-the-art image diffusion models in text-to-image generation tasks, with a significantly lower training cost. Additionally, NOVA generalizes well across extended video durations and enables diverse zero-shot applications in one unified model. Code and models are publicly available at https://github.com/baaivision/NOVA.
****
Fabian Otto,Philipp Becker,Vien Anh Ngo,Gerhard Neumann
Efficient Off-Policy Learning for High-Dimensional Action Spaces
Existing off-policy reinforcement learning algorithms often rely on an explicit state-action-value function representation, which can be problematic in high-dimensional action spaces due to the curse of dimensionality.
This reliance results in data inefficiency as maintaining a state-action-value function in such spaces is challenging.
We present an efficient approach that utilizes only a state-value function as the critic for off-policy deep reinforcement learning.
This approach, which we refer to as Vlearn, effectively circumvents the limitations of existing methods by eliminating the necessity for an explicit state-action-value function.
To this end, we leverage a weighted importance sampling loss for learning deep value functions from off-policy data.
While this is common for linear methods, it has not been combined with deep value function networks.
This transfer to deep methods is not straightforward and requires novel design choices such as robust policy updates, twin value function networks to avoid an optimization bias, and importance weight clipping.
We also present a novel analysis of the variance of our estimate compared to commonly used importance sampling estimators such as V-trace.
Our approach improves sample complexity as well as final performance and ensures consistent and robust performance across various benchmark tasks.
Eliminating the state-action-value function in Vlearn facilitates a streamlined learning process, yielding high-return agents.
****
Erwan Fagnou,Paul Caillon,Blaise Delattre,Alexandre Allauzen
Accelerated training through iterative gradient propagation along the residual path

Despite being the cornerstone of deep learning, backpropagation is criticized fo
r its inherent sequentiality, which can limit the scalability of very deep model
s.
Such models faced convergence issues due to vanishing gradient, later resolved u
sing residual connections. Variants of these are now widely used in modern archi
tectures.
However, the computational cost of backpropagation remains a major burden, accou
nting for most of the training time.
Taking advantage of residual-like architectural designs, we introduce Highway ba
ckpropagation, a parallelizable iterative algorithm that approximates backpropag
ation, by alternatively i) accumulating the gradient estimates along the residua
l path, and ii) backpropagating them through every layer in parallel. This algor
ithm is naturally derived from a decomposition of the gradient as the sum of gra
dients flowing through all paths, and is adaptable to a diverse set of common ar
chitectures, ranging from ResNets and Transformers to recurrent neural networks.
Through an extensive empirical study on a large selection of tasks and models, w
e evaluate Highway-BP and show that major speedups can be achieved with minimal
performance degradation.
****
Xi Jiang,Jian Li,Hanqiu Deng,Yong Liu,Bin-Bin Gao,Yifeng Zhou,Jialin Li,Chengjie
Wang,Feng Zheng
MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industri
al Anomaly Detection
In the field of industrial inspection, Multimodal Large Language Models (MLLMs)
have a high potential to renew the paradigms in practical applications due to th
eir robust language capabilities and generalization abilities. However, despite
their impressive problem-solving skills in many domains, MLLMs' ability in indus
trial anomaly detection has not been systematically studied. To bridge this gap,
we present MMAD, a full-spectrum MLLM benchmark in industrial Anomaly Detection
. We defined seven key subtasks of MLLMs in industrial inspection and designed a
novel pipeline to generate the MMAD dataset with 39,672 questions for 8,366 ind
ustrial images. With MMAD, we have conducted a comprehensive, quantitative evalu
ation of various state-of-the-art MLLMs. The commercial models performed the bes
t, with the average accuracy of GPT-4o models reaching 74.9\%. However, this res
ult falls far short of industrial requirements. Our analysis reveals that curren
t MLLMs still have significant room for improvement in answering questions relat
ed to industrial anomalies and defects. We further explore two training-free per
formance enhancement strategies to help models improve in industrial scenarios,
highlighting their promising potential for future research.
The code and data are available at https://github.com/jam-cc/MMAD.
****
Hanlin Zhang,Depen Morwani,Nikhil Vyas,Jingfeng Wu,Difan Zou,Udaya Ghai,Dean Fos
ter,Sham M. Kakade
How Does Critical Batch Size Scale in Pre-training?
Training large-scale models under given resources requires careful design of par
allelism strategies. In particular, the efficiency notion of critical batch size
(CBS), concerning the compromise between time and compute, marks the threshold
beyond which greater data parallelism leads to diminishing returns. To operation
alize it, we propose a measure of CBS and pre-train a series of auto-regressive
language models, ranging from 85 million to 1.2 billion parameters, on the C4 da
taset. Through extensive hyper-parameter sweeps and careful control of factors s
uch as batch size, momentum, and learning rate along with its scheduling, we sys
tematically investigate the impact of scale on CBS. Then we fit scaling laws wit
h respect to model and data sizes to decouple their effects. Overall, our result
s demonstrate that CBS scales primarily with data size rather than model size, a
finding we justify theoretically through the analysis of infinite-width limits
of neural networks and infinite-dimensional least squares regression. Of indepen
dent interest, we highlight the importance of common hyper-parameter choices and
strategies for studying large-scale pre-training beyond fixed training duration
s.

****
Xingchen Wan,Han Zhou,Ruoxi Sun,Sercan O Arik
From Few to Many: Self-Improving Many-Shot Reasoners Through Iterative Optimizat
ion and Generation
Recent advances in long-context large language models (LLMs) have led to the eme
rging paradigm of many-shot in-context learning (ICL), where it is observed that
 scaling many more demonstrating examples beyond the conventional few-shot setup
 in the context can lead to performance benefits. However, despite its promise,
it is unclear what aspects dominate the benefits and whether simply scaling to m
ore examples is the most effective way of improving many-shot ICL. In this work,
 we first provide an analysis on the factors driving many-shot ICL, and we find
that 1) many-shot performance can still be attributed to often a few disproporti
onately influential examples and 2) identifying such influential examples ("opti
mize") and using them as demonstrations to regenerate new examples ("generate")
can lead to further improvements. Inspired by the findings, we propose BRIDGE, a
n algorithm that alternates between the optimize step with Bayesian optimization
 to discover the influential sets of examples and the generate step to reuse thi
s set to expand the reasoning paths of the examples back to the many-shot regime
 automatically. On Gemini, Claude, and Mistral LLMs of different sizes, we show
BRIDGE led to significant improvements across a diverse set of tasks including s
ymbolic reasoning, numerical reasoning and code generation.
****
Kasia Kobalczyk,Nicolás Astorga,Tennison Liu,Mihaela van der Schaar
Active Task Disambiguation with LLMs
Despite the impressive performance of large language models (LLMs) across variou
s benchmarks, their ability to address ambiguously specified problems—frequent i
n real-world interactions—remains underexplored. To address this gap, we introdu
ce a formal definition of task ambiguity and frame the problem of task disambigu
ation through the lens of Bayesian Experimental Design. By posing clarifying que
stions, LLM agents can acquire additional task specifications, progressively nar
rowing the space of viable solutions and reducing the risk of generating unsatis
factory outputs. Yet, generating effective clarifying questions requires LLM age
nts to engage in a form of meta-cognitive reasoning, an ability LLMs may present
ly lack. Our proposed approach of active task disambiguation enables LLM agents
to generate targeted questions maximizing the information gain. Effectively, thi
s approach shifts the load from implicit to explicit reasoning about the space o
f viable solutions. Empirical results demonstrate that this form of question sel
ection leads to more effective task disambiguation in comparison to approaches r
elying on reasoning solely within the space of questions.
****
Omer Sahin Tas,Royden Wagner
Words in Motion: Extracting Interpretable Control Vectors for Motion Transformer
s
Transformer-based models generate hidden states that are difficult to interpret.
 In this work, we analyze hidden states and modify them at inference, with a foc
us on motion forecasting. We use linear probing to analyze whether interpretable
 features are embedded in hidden states. Our experiments reveal high probing acc
uracy, indicating latent space regularities with functionally important directio
ns. Building on this, we use the directions between hidden states with opposing
features to fit control vectors. At inference, we add our control vectors to hid
den states and evaluate their impact on predictions. Remarkably, such modificati
ons preserve the feasibility of predictions. We further refine our control vecto
rs using sparse autoencoders (SAEs). This leads to more linear changes in predic
tions when scaling control vectors. Our approach enables mechanistic interpretat
ion as well as zero-shot generalization to unseen dataset characteristics with n
egligible computational overhead.
****
Moritz Willig,Tim Tobiasch,Florian Peter Busch,Jonas Seng,Devendra Singh Dhami,K
ristian Kersting
Systems with Switching Causal Relations: A Meta-Causal Perspective

Most work on causality in machine learning assumes that causal relationships are driven by a constant underlying process. However, the flexibility of agents' actions or tipping points in the environmental process can change the qualitative dynamics of the system. As a result, new causal relationships may emerge, while existing ones change or disappear, resulting in an altered causal graph. To analyze these qualitative changes on the causal graph, we propose the concept of meta-causal states, which groups classical causal models into clusters based on equivalent qualitative behavior and consolidates specific mechanism parameterizations. We demonstrate how meta-causal states can be inferred from observed agent behavior, and discuss potential methods for disentangling these states from unlabeled data. Finally, we direct our analysis towards the application of a dynamical system, showing that meta-causal states can also emerge from inherent system dynamics, and thus constitute more than a context-dependent framework in which mechanisms emerge only as a result of external factors.
****

Wenhan Gao,Ruichen Xu,Yuefan Deng,Yi Liu

Discretization-invariance? On the Discretization Mismatch Errors in Neural Operators

In recent years, neural operators have emerged as a prominent approach for learning mappings between function spaces, such as the solution operators of parametric PDEs. A notable example is the Fourier Neural Operator (FNO), which models the integral kernel as a convolution operator and uses the Convolution Theorem to learn the kernel directly in the frequency domain. The parameters are decoupled from the resolution of the data, allowing the FNO to take inputs of different resolutions.

However, training at a lower resolution and inferring at a finer resolution does not guarantee consistent performance, nor can fine details, present only in fine-scale data, be learned solely from coarse data. In this work, we address this misconception by defining and examining the discretization mismatch error: the discrepancy between the outputs of the neural operator when using different discretizations of the input data. We demonstrate that neural operators may suffer from discretization mismatch errors that hinder their effectiveness when inferred on data with resolutions different from that of the training data or when trained on data with varying resolutions. As neural operators underpin many critical cross-resolution scientific tasks, such as climate modeling and fluid dynamics, understanding discretization mismatch errors is essential. Based on our findings, we propose a Cross-Resolution Operator-learning Pipeline that is free of aliasing and discretization mismatch errors, enabling efficient cross-resolution and multi-spatial-scale learning, and resulting in superior performance.
****

Feng Tian,Yixuan Li,Yichao Yan,Shanyan Guan,Yanhao Ge,Xiaokang Yang

PostEdit: Posterior Sampling for Efficient Zero-Shot Image Editing

In the field of image editing, three core challenges persist: controllability, background preservation, and efficiency. Inversion-based methods rely on time-consuming optimization to preserve the features of the initial images, which results in low efficiency due to the requirement for extensive network inference. Conversely, inversion-free methods lack theoretical support for background similarity, as they circumvent the issue of maintaining initial features to achieve efficiency. As a consequence, none of these methods can achieve both high efficiency and background consistency. To tackle the challenges and the aforementioned disadvantages, we introduce PostEdit, a method that incorporates a posterior scheme to govern the diffusion sampling process. Specifically, a corresponding measurement term related to both the initial features and Langevin dynamics is introduced to optimize the estimated image generated by the given target prompt. Extensive experimental results indicate that the proposed PostEdit achieves state-of-the-art editing performance while accurately preserving unedited regions. Furthermore, the method is both inversion- and training-free, necessitating approximately 1.5 seconds and 18 GB of GPU memory to generate high-quality results.
****

Du Chen,Geoffrey A. Chua

Exploiting Hidden Symmetry to Improve Objective Perturbation for DP Linear Learn
ers with a Nonsmooth L1-Norm
Objective Perturbation (OP) is a classic approach to differentially private (DP)
 convex optimization with smooth loss functions but is less understood for nonsm
ooth cases. In this work, we study how to apply OP to DP linear learners under l
oss functions with an implicit ℓ_1-norm structure, such as $\max(0,x)$ as a
 motivating example. We propose to first smooth out the implicit ℓ_1-norm b
y convolution, and then invoke standard OP. Convolution has many advantages that
 distinguish itself from Moreau Envelope, such as approximating from above and a
 higher degree of hyperparameters. These advantages, in conjunction with the sym
metry of ℓ_1-norm, result in tighter pointwise approximation, which further
 facilitates tighter analysis of generalization risks by using pointwise bounds.
 Under mild assumptions on groundtruth distributions, the proposed OP-based algo
rithm is found to be rate-optimal, and can achieve the excess generalization ris
k $\mathcal{O}(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon})$.
Experiments demonstrate the competitive performance of the proposed method to No
isy-SGD.
****

Chenglong Kang,Xiaoyi Liu,Fei Guo
RetroInText: A Multimodal Large Language Model Enhanced Framework for Retrosynth
etic Planning via In-Context Representation Learning
Development of robust and effective strategies for retrosynthetic planning requi
res a deep understanding of the synthesis process. A critical step in achieving
 this goal is accurately identifying synthetic intermediates. Current machine lea
rning-based methods often overlook the valuable context from the overall route,
 focusing only on predicting reactants from the product, requiring cost annotatio
ns for every reaction step, and ignoring the multi-faced nature of molecular, re
sulting in inaccurate synthetic route predictions. Therefore, we introduce Retro
InText, an advanced end-to-end framework based on a multimodal Large Language Mo
del (LLM), featuring in-context learning with TEXT descriptions of synthetic rou
tes. First, RetroInText including ChatGPT presents detailed descriptions of the
 reaction procedure. It learns the distinct compound representations in parallel
 with corresponding molecule encoders to extract multi-modal representations incl
uding 3D features. Subsequently, we propose an attention-based mechanism that of
fers a fusion module to complement these multi-modal representations with in-con
text learning and a fine-tuned language model for a single-step model. As a resu
lt, RetroInText accurately represents and effectively captures the complex relat
ionship between molecules and the synthetic route. In experiments on the USPTO p
athways dataset RetroBench, RetroInText outperforms state-of-the-art methods, ac
hieving up to a 5% improvement in Top-1 test accuracy, particularly for long syn
thetic routes. These results demonstrate the superiority of RetroInText by integ
rating with context information over routes. They also demonstrate its potential
 for advancing pathway design and facilitating the development of organic chemis
try. Code is available at https://github.com/guofei-tju/RetroInText.
****

Ke Wang,Nikolaos Dimitriadis,Alessandro Favero,Guillermo Ortiz-Jimenez,François
Fleuret,Pascal Frossard
LiNeS: Post-training Layer Scaling Prevents Forgetting and Enhances Model Mergin
g
Fine-tuning pre-trained models has become the standard approach to endow them wi
th specialized knowledge, but it poses fundamental challenges. In particular, (i
) fine-tuning often leads to catastrophic forgetting, where improvements on a ta
rget domain degrade generalization on other tasks, and (ii) merging fine-tuned c
heckpoints from disparate tasks can lead to significant performance loss. To add
ress these challenges, we introduce LiNeS, Layer-increasing Network Scaling, a p
ost-training editing technique designed to preserve pre-trained generalization w
hile enhancing fine-tuned task performance. LiNeS scales parameter updates linea
rly based on their layer depth within the network, maintaining shallow layers cl
ose to their pre-trained values to preserve general features while allowing deep
er layers to retain task-specific representations. In multi-task model merging s

cenarios, layer-wise scaling of merged parameters reduces negative task interfer
ence. LiNeS demonstrates significant improvements in both single-task and multi-
task settings across various benchmarks in vision and natural language processin
g. It mitigates forgetting, enhances out-of-distribution generalization, integra
tes seamlessly with existing multi-task model merging baselines improving their
performance across benchmarks and model sizes, and can boost generalization when
 merging LLM policies aligned with different rewards via RLHF. Our method is sim
ple to implement, computationally efficient and complementary to many existing t
echniques. Our source code is available at github.com/wang-kee/LiNeS.
****

Qiujie Xie,Qingqiu Li,Zhuohao Yu,Yuejie Zhang,Yue Zhang,Linyi Yang
An Empirical Analysis of Uncertainty in Large Language Model Evaluations
As LLM-as-a-Judge emerges as a new paradigm for assessing large language models
(LLMs), concerns have been raised regarding the alignment, bias, and stability o
f LLM evaluators. While substantial work has focused on alignment and bias, litt
le research has concentrated on the stability of LLM evaluators. In this paper,
we conduct extensive experiments involving 9 widely used LLM evaluators across 2
 different evaluation settings to investigate the uncertainty in model-based LLM
 evaluations. We pinpoint that LLM evaluators exhibit varying uncertainty based
on model families and sizes. With careful comparative analyses, we find that emp
loying special prompting strategies, whether during inference or post-training,
can alleviate evaluation uncertainty to some extent. By utilizing uncertainty to
 enhance LLM's reliability and detection capability in Out-Of-Distribution (OOD)
 data, we further fine-tune an uncertainty-aware LLM evaluator named ConfiLM usi
ng a human-annotated fine-tuning set and assess ConfiLM's OOD evaluation ability
 on a manually designed test set sourced from the 2024 Olympics. Experimental re
sults demonstrate that incorporating uncertainty as additional information durin
g the fine-tuning phase can largely improve the model's evaluation performance i
n OOD scenarios. The code and data are released at: https://github.com/hasakiXie
123/LLM-Evaluator-Uncertainty.
****

Avrajit Ghosh,Soo Min Kwon,Rongrong Wang,Saiprasad Ravishankar,Qing Qu
Learning Dynamics of Deep Matrix Factorization Beyond the Edge of Stability
Deep neural networks trained using gradient descent with a fixed learning rate $
\eta$ often operate in the regime of ``edge of stability'' (EOS), where the larg
est eigenvalue of the Hessian equilibrates about the stability threshold $2/\eta
$. In this work, we present a fine-grained analysis of the learning dynamics of
(deep) linear networks (DLNs) within the deep matrix factorization loss beyond E
OS. For DLNs, loss oscillations beyond EOS follow a period-doubling route to cha
os. We theoretically analyze the regime of the 2-period orbit and show that the
loss oscillations occur within a small subspace, with the dimension of the subsp
ace precisely characterized by the learning rate. The crux of our analysis lies
in showing that the symmetry-induced conservation law for gradient flow, defined
 as the balancing gap among the singular values across layers, breaks at EOS and
 decays monotonically to zero. Overall, our results contribute to explaining two
 key phenomena in deep networks: (i) shallow models and simple tasks do not alwa
ys exhibit EOS; and (ii) oscillations occur within top features}. We present exp
eriments to support our theory, along with examples demonstrating how these phen
omena occur in nonlinear networks and how they differ from those which have beni
gn landscape such as in DLNs.
****

Nazia Tasnim,Bryan A. Plummer
RECAST: Reparameterized, Compact weight Adaptation for Sequential Tasks
Incremental learning aims to adapt to new sets of categories over time with mini
mal computational overhead. Prior work often addresses this task by training eff
icient task-specific adaptors that modify frozen layer weights or features to ca
pture relevant information without affecting predictions on any previously learn
ed categories. While these adaptors are generally more efficient than finetuning
 the entire network, they still can require tens to hundreds of thousands task-s
pecific trainable parameters even for relatively small networks. This makes it c

hallenging to operate in resource-constrained environments with high communicati on costs, such as edge devices or mobile phones. Thus, we propose Reparameterize d, Compact weight Adaptation for Sequential Tasks (RECAST), a novel method that dramatically reduces the number of task-specific trainable parameters to fewer t han 50 – several orders of magnitude less than competing methods like LoRA. RECA ST accomplishes this efficiency by learning to decompose layer weights into a so ft parameter-sharing framework consisting of a set of shared weight templates an d very few module-specific scaling factors or coefficients. This soft parameter- sharing framework allows for effective task-wise reparameterization by tuning on ly these coefficients while keeping templates frozen. A key innovation of RECAST is the novel weight reconstruction pipeline called Neural Mimicry, which elimin ates the need for pretraining from scratch. This allows for high-fidelity emulat ion of existing pretrained weights within our framework and provides quick adapt ability to any model scale and architecture. Extensive experiments across six di verse datasets demonstrate RECAST outperforms the state-of-the-art by up to ∼ 1. 5% and improves baselines by > 3% across various scales, architectures, and para meter spaces. Moreover, we show that RECAST’s architecture-agnostic nature allow s for seamless integration with existing methods, further boosting performance.
****

Dongzhi Jiang,Renrui Zhang,Ziyu Guo,Yanmin Wu,jiayi lei,Pengshuo Qiu,Pan Lu,Zehu i Chen,Guanglu Song,Peng Gao,Yu Liu,Chunyuan Li,Hongsheng Li

MMSearch: Unveiling the Potential of Large Models as Multi-modal Search Engines

The advent of Large Language Models (LLMs) has paved the way for AI search engin es, e.g., SearchGPT, showcasing a new paradigm in human-internet interaction. Ho wever, most current AI search engines are limited to text-only settings, neglect ing the multimodal user queries and the text-image interleaved nature of website information. Recently, Large Multimodal Models (LMMs) have made impressive stri des. Yet, whether they can function as AI search engines remains under-explored, leaving the potential of LMMs in multimodal search an open question. To this en d, we first design a delicate pipeline, MMSearch-Engine, to empower any LMMs wit h multimodal search capabilities. On top of this, we introduce MMSearch, a compr ehensive evaluation benchmark to assess the multimodal search performance of LMM s. The curated dataset contains 300 manually collected instances spanning 14 sub fields, which involves no overlap with the current LMMs' training data, ensuring the correct answer can only be obtained within searching. By using MMSearch-Eng ine, the LMMs are evaluated by performing three individual tasks (requery, reran k, and summarization), and one challenging end-to-end task with a complete searc hing process. We conduct extensive experiments on closed-source and open-source LMMs. Among all tested models, GPT-4o with MMSearch-Engine achieves the best res ults, which surpasses the commercial product, Perplexity Pro, in the end-to-end task, demonstrating the effectiveness of our proposed pipeline. We further prese nt error analysis to unveil current LMMs still struggle to fully grasp the multi modal search tasks, and conduct ablation study to indicate the potential of scal ing test-time computation for AI search engine. We hope MMSearch may provide uni que insights to guide the future development of multimodal AI search engine.
****

Jeonghoon Shim,Gyuhyeon Seo,Cheongsu Lim,Yohan Jo

ToolDial: Multi-turn Dialogue Generation Method for Tool-Augmented Language Mode ls

Tool-Augmented Language Models (TALMs) leverage external APIs to answer user que ries across various domains. However, existing benchmark datasets for TALM resea rch often feature simplistic dialogues that do not reflect real-world scenarios, such as the need for models to ask clarifying questions or proactively call add itional APIs when essential information is missing. To address these limitations , we construct and release ToolDial, a dataset comprising 11,111 multi-turn dial ogues, with an average of 8.95 turns per dialogue, based on APIs from RapidAPI. ToolDial has two key characteristics. First, the dialogues incorporate 16 user a nd system actions (e.g., request, clarify, fail inform) to capture the rich dyna mics of real-world interactions. Second, we simulate dialogues where the system requests necessary information from the user based on API documentation and seek

s additional APIs if the user fails to provide the required information. To faci
litate this process, we introduce a method for generating an API graph that repr
esents input and output compatibility between APIs. Using ToolDial, we evaluate
a suite of language models on their ability to predict correct actions and extra
ct input parameter values for API calls from the dialogue history. Modern langua
ge models achieve accuracy scores below 70\%, indicating substantial room for im
provement. We provide a detailed analysis of the areas where these models fall s
hort.
****

Yuancheng Xu,Udari Madhushani Sehwag,Alec Koppel,Sicheng Zhu,Bang An,Furong Huan
g,Sumitra Ganesh
GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-Time
Alignment
Large Language Models (LLMs) exhibit impressive capabilities but require careful
 alignment with human preferences. Traditional training-time methods finetune LL
Ms using human preference datasets but incur significant training costs and requ
ire repeated training to handle diverse user preferences. Test-time alignment me
thods address this by using reward models (RMs) to guide frozen LLMs without ret
raining. However, existing test-time approaches rely on trajectory-level RMs whi
ch are designed to evaluate complete responses, making them unsuitable for autor
egressive text generation that requires computing next-token rewards from partia
l responses. To address this, we introduce GenARM, a test-time alignment approac
h that leverages the Autoregressive Reward Model—a novel reward parametrization
designed to predict next-token rewards for efficient and effective autoregressiv
e generation. Theoretically, we demonstrate that this parametrization can provab
ly guide frozen LLMs toward any distribution achievable by traditional RMs withi
n the KL-regularized reinforcement learning framework. Experimental results show
 that GenARM significantly outperforms prior test-time alignment baselines and m
atches the performance of training-time methods. Additionally, GenARM enables ef
ficient weak-to-strong guidance, aligning larger LLMs with smaller RMs without t
he high costs of training larger models. Furthermore, GenARM supports multi-obje
ctive alignment, allowing real-time trade-offs between preference dimensions and
 catering to diverse user preferences without retraining. Our project page is av
ailable at: https://genarm.github.io.
****

Jiawei Huang,Vinzenz Thoma,Zebang Shen,Heinrich H. Nax,Niao He
Learning to Steer Markovian Agents under Model Uncertainty
Designing incentives for an adapting population is a ubiquitous problem in a wid
e array of economic applications and beyond. In this work, we study how to desig
n additional rewards to steer multi-agent systems towards desired policies \emph
{without} prior knowledge of the agents' underlying learning dynamics. Motivated
 by the limitation of existing works, we consider a new and general category of
learning dynamics called \emph{Markovian agents}. We introduce a model-based non
-episodic Reinforcement Learning (RL) formulation for our steering problem. Impo
rtantly, we focus on learning a \emph{history-dependent} steering strategy to ha
ndle the inherent model uncertainty about the agents' learning dynamics. We intr
oduce a novel objective function to encode the desiderata of achieving a good st
eering outcome with reasonable cost. Theoretically, we identify conditions for t
he existence of steering strategies to guide agents to the desired policies. Com
plementing our theoretical contributions, we provide empirical algorithms to app
roximately solve our objective, which effectively tackles the challenge in learn
ing history-dependent strategies. We demonstrate the efficacy of our algorithms
through empirical evaluations.
****

Joy Hsu,Jiayuan Mao,Joshua B. Tenenbaum,Noah Goodman,Jiajun Wu
What Makes a Maze Look Like a Maze?
A unique aspect of human visual understanding is the ability to flexibly interpr
et abstract concepts: acquiring lifted rules explaining what they symbolize, gro
unding them across familiar and unfamiliar contexts, and making predictions or r
easoning about them. While off-the-shelf vision-language models excel at making

literal interpretations of images (e.g., recognizing object categories such as t
ree branches), they still struggle to make sense of such visual abstractions (e.
g., how an arrangement of tree branches may form the walls of a maze). To addres
s this challenge, we introduce Deep Schema Grounding (DSG), a framework that lev
erages explicit structured representations of visual abstractions for grounding
and reasoning. At the core of DSG are schemas—dependency graph descriptions of a
bstract concepts that decompose them into more primitive-level symbols. DSG uses
 large language models to extract schemas, then hierarchically grounds concrete
to abstract components of the schema onto images with vision-language models. Th
e grounded schema is used to augment visual abstraction understanding. We system
atically evaluate DSG and different methods in reasoning on our new Visual Abstr
actions Benchmark, which consists of diverse, real-world images of abstract conc
epts and corresponding question-answer pairs labeled by humans. We show that DSG
 significantly improves the abstract visual reasoning performance of vision-lang
uage models, and is a step toward human-aligned understanding of visual abstract
ions.
****
Maojia Song,Shang Hong Sim,Rishabh Bhardwaj,Hai Leong Chieu,Navonil Majumder,Sou
janya Poria
Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attribut
ions and Learning to Refuse
LLMs are an integral component of retrieval-augmented generation (RAG) systems.
While many studies focus on evaluating the overall quality of end-to-end RAG sys
tems, there is a gap in understanding the appropriateness of LLMs for the RAG ta
sk. To address this, we introduce Trust-Score, a holistic metric that evaluates
the trustworthiness of LLMs within the RAG framework. Our results show that vari
ous prompting methods, such as in-context learning, fail to effectively adapt LL
Ms to the RAG task as measured by Trust-Score. Consequently, we propose Trust-Al
ign, a method to align LLMs for improved Trust-Score performance. 26 out of 27 m
odels aligned using Trust-Align substantially outperform competitive baselines o
n ASQA, QAMPARI, and ELI5. Specifically, in LLaMA-3-8b, Trust-Align outperforms
FRONT on ASQA (↑12.56), QAMPARI (↑36.04), and ELI5 (↑17.69). Trust-Align also si
gnificantly enhances models' ability to correctly refuse and provide quality cit
ations. We also demonstrate the effectiveness of Trust-Align across different op
en-weight models, including the LLaMA series (1b to 8b), Qwen-2.5 series (0.5b t
o 7b), and Phi3.5 (3.8b). We release our code at https://github.com/declare-lab/
trust-align.
****
Lirong Wu,Yunfan Liu,Haitao Lin,Yufei Huang,Guojiang Zhao,Zhifeng Gao,Stan Z. Li
A Simple yet Effective $\Delta\Delta G$ Predictor is An Unsupervised Antibody Op
timizer and Explainer
The proteins that exist today have been optimized over billions of years of natu
ral evolution, during which nature creates random mutations and selects them. Th
e discovery of functionally promising mutations is challenged by the limited evo
lutionary accessible regions, i.e., only a small region on the fitness landscape
 is beneficial. There have been numerous priors used to constrain protein evolut
ion to regions of landscapes with high-fitness variants, among which the change
in binding free energy ($\Delta\Delta G$) of protein complexes upon mutations is
 one of the most commonly used priors. However, the huge mutation space poses tw
o challenges: (1) how to improve the efficiency of $\Delta\Delta G$ prediction f
or fast mutation screening; and (2) how to explain mutation preferences and effi
ciently explore accessible evolutionary regions. To address these challenges, we
 propose a lightweight $\Delta\Delta G$ predictor (Light-DDG), which adopts a st
ructure-aware Transformer as the backbone and enhances it by knowledge distilled
 from existing powerful but computationally heavy $\Delta\Delta G$ predictors. A
dditionally, we augmented, annotated, and released a large-scale dataset contain
ing millions of mutation data for pre-training Light-DDG. We find that such a si
mple yet effective Light-DDG can serve as a good unsupervised antibody optimizer
 and explainer. For the target antibody, we propose a novel Mutation Explainer t
o learn mutation preferences, which accounts for the marginal benefit of each mu

tation per residue. To further explore accessible evolutionary regions, we conduct preference-guided antibody optimization and evaluate antibody candidates quickly using Light-DDG to identify desirable mutations. Extensive experiments have demonstrated the effectiveness of Light-DDG in terms of test generalizability, noise robustness, and inference practicality, e.g., 89.7$\times$ inference acceleration and 15.45\% performance gains over previous state-of-the-art baselines. A case study of SARS-CoV-2 further demonstrates the crucial role of Light-DDG for mutation explanation and antibody optimization.
****
Yunfei Xie,Ce Zhou,Lang Gao,Juncheng Wu,Xianhang Li,Hong-Yu Zhou,Sheng Liu,Lei Xing,James Zou,Cihang Xie,Yuyin Zhou
MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine
This paper introduces MedTrinity-25M, a comprehensive, large-scale multimodal dataset for medicine, covering over 25 million images across 10 modalities with multigranular annotations for more than 65 diseases. These multigranular annotations encompass both global information, such as modality and organ detection, and local information like ROI analysis, lesion texture, and region-wise correlations. Unlike the existing multimodal datasets, which are limited by the availability of image-text pairs, we have developed the first automated pipeline that scales up multimodal data by generating multigranular visual and textual annotations in the form of image-ROI-description triplets without the need for any paired text descriptions. Specifically, data from over 30 different sources have been collected, preprocessed, and grounded using domain-specific expert models to identify ROIs related to abnormal regions. We then build a comprehensive knowledge base and prompt multimodal large language models to perform retrieval-augmented generation with the identified ROIs as guidance, resulting in multigranular textual descriptions. Compared to existing datasets, MedTrinity-25M provides the most enriched annotations, supporting a comprehensive range of multimodal tasks such as captioning and report generation, as well as vision-centric tasks like classification and segmentation. We propose LLaVA-Tri by pretraining LLaVA on MedTrinity-25M, achieving state-of-the-art performance on VQA-RAD, SLAKE, and PathVQA, surpassing representative SOTA multimodal large language models. Furthermore, MedTrinity-25M can also be utilized to support large-scale pre-training of multimodal medical AI models, contributing to the development of future foundation models in the medical domain. We will make our dataset available. The dataset is publicly available at https://yunfeixie233.github.io/MedTrinity-25M/.
****
Puning Zhao,Jiafei Wu,Zhe Liu,Li Shen,Zhikun Zhang,Rongfei Fan,Le Sun,Qingming Li
Enhancing Learning with Label Differential Privacy by Vector Approximation
Label differential privacy (DP) is a framework that protects the privacy of labels in training datasets, while the feature vectors are public. Existing approaches protect the privacy of labels by flipping them randomly, and then train a model to make the output approximate the privatized label. However, as the number of classes K increases, stronger randomization is needed, thus the performances of these methods become significantly worse. In this paper, we propose a vector approximation approach for learning with label local differential privacy, which is easy to implement and introduces little additional computational overhead. Instead of flipping each label into a single scalar, our method converts each label into a random vector with K components, whose expectations reflect class conditional probabilities. Intuitively, vector approximation retains more information than scalar labels. A brief theoretical analysis shows that the performance of our method only decays slightly with K. Finally, we conduct experiments on both synthesized and real datasets, which validate our theoretical analysis as well as the practical performance of our method.
****
Mehdi Azabou,Krystal Xuejing Pan,Vinam Arora,Ian Jarratt Knight,Eva L Dyer,Blake Aaron Richards
Multi-session, multi-task neural decoding from distinct cell-types and brain reg

ions
Recent work has shown that scale is important for improved brain decoding, with more data leading to greater decoding accuracy. However, large-scale decoding across many different datasets is challenging because neural circuits are heterogeneous---each brain region contains a unique mix of cellular sub-types, and the responses to different stimuli are diverse across regions and sub-types. It is unknown whether it is possible to pre-train and transfer brain decoding models between distinct tasks, cellular sub-types, and brain regions. To address these questions, we developed a multi-task transformer architecture and trained it on the entirety of the Allen Institute's Brain Observatory dataset. This dataset contains responses from over 100,000 neurons in 6 areas of the brains of mice, observed with two-photon calcium imaging, recorded while the mice observed different types of visual stimuli. Our results demonstrate that transfer is indeed possible -combining data from different sources is beneficial for a number of downstream decoding tasks. As well, we can transfer the model between regions and sub-types, demonstrating that there is in fact common information in diverse circuits that can be extracted by an appropriately designed model. Interestingly, we found that the model's latent representations showed clear distinctions between different brain regions and cellular sub-types, even though it was never given any information about these distinctions. Altogether, our work demonstrates that training a large-scale neural decoding model on diverse data is possible, and this provides a means of studying the differences and similarities between heterogeneous neural circuits.
****
Zeyu Huang,Zihan Qiu,Zili Wang,Edoardo Ponti,Ivan Titov
Post-hoc Reward Calibration: A Case Study on Length Bias
Reinforcement Learning from Human Feedback aligns the outputs of Large Language Models with human values and preferences. Central to this process is the reward model (RM), which translates human feedback into training signals for optimising LLM behaviour. However, RMs can develop biases by exploiting spurious correlations in their training data, such as favouring outputs based on length or
style rather than true quality. These biases can lead to incorrect output rankings, sub-optimal model evaluations, and the amplification of undesirable behaviours in LLMs alignment. This paper addresses the challenge of correcting such biases without additional data and training, introducing the concept of Post-hoc Reward Calibration. We first propose to use local average reward to estimate the bias term
and, thus, remove it to approximate the underlying true reward. We then extend the approach to a more general and robust form with the Locally Weighted Regression. Focusing on the prevalent length bias, we validate our proposed approaches across three experimental settings, demonstrating consistent improvements: (1) a 3.11 average performance gain across 33 reward models on the RewardBench
dataset; (2) improved agreement of RM produced rankings with GPT-4 evaluations and human preferences based on the AlpacaEval benchmark; and (3) improved Length-Controlled win rate (Dubois et al., 2024) of the RLHF process in multiple LLM–RM combinations. According to our experiments, our method is computationally efficient and generalisable to other types of bias and RMs, offering a scalable and robust solution for mitigating biases in LLM alignment and evaluation.
****
Qiuchen Wang,Zehui Chen,Chenhongyi Yang,Jiaming Liu,Zhenyu Li,Feng Zhao
PseDet: Revisiting the Power of Pseudo Label in Incremental Object Detection
Incremental Objection Detection (IOD) facilitates the expansion of the usage scope of object detectors without forgetting previously acquired knowledge. Current approaches mostly adopt response-level knowledge distillation to overcome forgetting issues, by conducting implicit memory replay from the teacher model on new training data. However, this indirect learning paradigm does not fully leverage the knowledge generated by the teacher model. In this paper, we dive deeper into the mechanism of pseudo-labeling in incremental object detection by investigating three critical problems: (a) the upper bound quality of the pseudo labels is greatly limited by the previous model, (b) fixed score thresholds for label fil

tering, without considering the distribution across categories, and (c) the conf
idence score generated by the model does not well reflect the quality of the loc
alization. Based on these observations, we propose a simple yet effective pseudo
-labeling continual object detection framework, namely PseDet. Specifically, we
introduce the spatio-temporal enhancement module to alleviate the negative effec
ts when learning noisy data from the previous model. Considering the score distr
ibution divergence across different classes, we propose the Categorical Adaptive
 Label Selector with a simple mathematical prior and fast K-Means pre-computatio
n to dynamically determine the class-wise filtering threshold. In order to align
 the label score with the localization quality of the pseudo labels, we project
the score through non-linear mapping to calibrate the distribution and integrate
 it into the new-step supervision. Extensive experiments on the competitive COCO
 benchmarks demonstrate the effectiveness and generalization of PseDet. Notably,
 it achieves 43.5+/41.2+ mAP under the 1/4-step incremental settings, achieving
new state-of-the-art performance.
****
Bertram Højer,Oliver Simon Jarvis,Stefan Heinrich
Improving Reasoning Performance in Large Language Models via Representation Engi
neering
Recent advancements in large language models (LLMs) have resulted in increasingl
y anthropomorphic language concerning the ability of LLMs to reason. Whether \te
xtit{reasoning} in LLMs should be understood to be inherently different is, howe
ver, widely debated. We propose utilizing a representation engineering approach
wherein model activations are read from the residual stream of an LLM when proce
ssing a reasoning task. The activations are used to derive a control vector that
 is applied to the model as an inference-time intervention, modulating the repre
sentational space of the model, to improve performance on the specified task. We
 publish the code for deriving control vectors and analyzing model representatio
ns. The method allows us to improve performance on reasoning benchmarks and asse
ss how control vectors influence the final logit distribution of a model via met
rics such as KL divergence and entropy. We apply control vectors to Mistral-7B-I
nstruct and a range of Pythia models on an inductive, a deductive and mathematic
al reasoning task. We show that an LLM can, to a certain degree, be controlled t
o improve its perceived reasoning ability by modulating activations. The interve
ntion is dependent upon the ability to reliably extract the model's typical stat
e when correctly solving a task. Our results suggest that reasoning performance
can be modulated in the same manner as other information-processing tasks perfor
med by LLMs and demonstrate that we are capable of improving performance on spec
ific tasks via a simple intervention on the residual stream with no additional t
raining.
****
Ammar I Marvi,Nancy Kanwisher,Meenakshi Khosla
Sparse components distinguish visual pathways & their alignment to neural networ
ks
The ventral, dorsal, and lateral streams in high-level human visual cortex are i
mplicated in distinct functional processes. Yet, deep neural networks (DNNs) tra
ined on a single task model the entire visual system surprisingly well, hinting
at common computational principles across these pathways. To explore this incons
istency, we applied a novel sparse decomposition approach to identify the domina
nt components of visual representations within each stream. Consistent with trad
itional neuroscience research, we find a clear difference in component response
profiles across the three visual streams—identifying components selective for fa
ces, places, bodies, text, and food in the ventral stream; social interactions,
implied motion, and hand actions in the lateral stream; and some less interpreta
ble components in the dorsal stream. Building on this, we introduce Sparse Compo
nent Alignment (SCA), a new method for measuring representational alignment betw
een brains and machines that better captures the latent neural tuning of these t
wo visual systems. We find that standard visual DNNs are more aligned with ventr
al than either dorsal or lateral representations. SCA reveals these distinctions
 with greater resolution than conventional population-level geometry, offering a

measure of representational alignment that is sensitive to a system’s underlyin
g axes of neural tuning.
****

Xiaoyang Wu,Lin Lu,Zhaojun Wang,Changliang Zou

Conditional Testing based on Localized Conformal p-values

In this paper, we address conditional testing problems through the conformal inf
erence framework. We define the localized conformal p-values by inverting pred
iction intervals and prove their theoretical properties. These defined p-value
s are then applied to several conditional testing problems to illustrate their p
racticality. Firstly, we propose a conditional outlier detection procedure to te
st for outliers in the conditional distribution with finite-sample false discove
ry rate (FDR) control. We also introduce a novel conditional label screening pro
blem with the goal of screening multivariate response variables and propose a sc
reening procedure to control the family-wise error rate (FWER). Finally, we cons
ider the two-sample conditional distribution test and define a weighted U-statis
tic through the aggregation of localized p-values. Numerical simulations and r
eal-data examples validate the superior performance of our proposed strategies.
****

Weibang Jiang,Yansen Wang,Bao-liang Lu,Dongsheng Li

NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between La
nguage and EEG Signals

Recent advancements for large-scale pre-training with neural signals such as ele
ctroencephalogram (EEG) have shown promising results, significantly boosting the
 development of brain-computer interfaces (BCIs) and healthcare. However, these
pre-trained models often require full fine-tuning on each downstream task to ach
ieve substantial improvements, limiting their versatility and usability, and lea
ding to considerable resource wastage. To tackle these challenges, we propose Ne
uroLM, the first multi-task foundation model that leverages the capabilities of
Large Language Models (LLMs) by regarding EEG signals as a foreign language, end
owing the model with multi-task learning and inference capabilities. Our approac
h begins with learning a text-aligned neural tokenizer through vector-quantized
temporal-frequency prediction, which encodes EEG signals into discrete neural to
kens. These EEG tokens, generated by the frozen vector-quantized (VQ) encoder, a
re then fed into an LLM that learns causal EEG information via multi-channel aut
oregression. Consequently, NeuroLM can understand both EEG and language modaliti
es. Finally, multi-task instruction tuning adapts NeuroLM to various downstream
tasks. We are the first to demonstrate that, by specific incorporation with LLMs
, NeuroLM unifies diverse EEG tasks within a single model through instruction tu
ning. The largest variant NeuroLM-XL has record-breaking 1.7B parameters for EEG
 signal processing, and is pre-trained on a large-scale corpus comprising approx
imately 25,000-hour EEG data. When evaluated on six diverse downstream datasets,
 NeuroLM showcases the huge potential of this multi-task learning paradigm.
****

Shitong Shao,zikai zhou,Bai LiChen,Haoyi Xiong,Zeke Xie

IV-mixed Sampler: Leveraging Image Diffusion Models for Enhanced Video Synthesis

Exploring suitable solutions to improve performance by increasing the computatio
nal cost of inference in visual diffusion models is a highly promising direction
. Sufficient prior studies have demonstrated that correctly scaling up computati
on in the sampling process can successfully lead to improved generation quality,
 enhanced image editing, and compositional generalization. While there have been
 rapid advancements in developing inference-heavy algorithms for improved image
generation, relatively little work has explored inference scaling laws in video
diffusion models (VDMs). Furthermore, existing research shows only minimal perfo
rmance gains that are perceptible to the naked eye. To address this, we design a
 novel training-free algorithm IV-Mixed Sampler that leverages the strengths of
image diffusion models (IDMs) to assist VDMs surpass their current capabilities.
 The core of IV-Mixed Sampler is to use IDMs to significantly enhance the qualit
y of each video frame and VDMs ensure the temporal coherence of the video during
 the sampling process. Our experiments have demonstrated that IV-Mixed Sampler a
chieves state-of-the-art performance on 4 benchmarks including UCF-101-FVD, MSR-

VTT-FVD, Chronomagic-Bench-150/1649, and VBench. For example, the open-source Animatediff with IV-Mixed Sampler reduces the UMT-FVD score from 275.2 to 228.6, closing to 223.1 from the closed-source Pika-2.0.
****
Olivia Wiles,Chuhan Zhang,Isabela Albuquerque,Ivana Kajic,Su Wang,Emanuele Bugliarello,Yasumasa Onoe,Pinelopi Papalampidi,Ira Ktena,Christopher Knutsen,Cyrus Rashtchian,Anant Nawalgaria,Jordi Pont-Tuset,Aida Nematzadeh
Revisiting text-to-image evaluation with Gecko: on metrics, prompts, and human rating
While text-to-image (T2I) generative models have become ubiquitous, they do not necessarily generate images that align with a given prompt.
While many metrics and benchmarks have been proposed to evaluate T2I models and alignment metrics, the impact of the evaluation components (prompt sets, human annotations, evaluation task) has not been systematically measured.
We find that looking at only *one slice of data*, i.e. one set of capabilities or human annotations, is not enough to obtain stable conclusions that generalise to new conditions or slices when evaluating T2I models or alignment metrics.
We address this by introducing an evaluation suite of $>$100K annotations across four human annotation templates that comprehensively evaluates models' capabilities across a range of methods for gathering human annotations and comparing models.
In particular, we propose (1) a carefully curated set of prompts -- *Gecko2K*; (2) a statistically grounded method of comparing T2I models; and (3) how to systematically evaluate metrics under three *evaluation tasks* -- *model ordering, pair-wise instance scoring, point-wise instance scoring*.
Using this evaluation suite, we evaluate a wide range of metrics and find that a metric may do better in one setting but worse in another.
As a result, we introduce a new, interpretable auto-eval metric that is consistently better correlated with human ratings than such existing metrics on our evaluation suite--across different human templates and evaluation settings--and on TIFA160.
****
Tom Sander,Pierre Fernandez,Alain Oliviero Durmus,Teddy Furon,Matthijs Douze
Watermark Anything With Localized Messages
Image watermarking methods are not tailored to handle small watermarked areas.
This restricts applications in real-world scenarios where parts of the image may come from different sources or have been edited.
We introduce a deep-learning model for localized image watermarking, dubbed the Watermark Anything Model (WAM).
The WAM embedder imperceptibly modifies the input image, while the extractor segments the received image into watermarked and non-watermarked areas and recovers one or several hidden messages from the areas found to be watermarked.
The models are jointly trained at low resolution and without perceptual constraints, then post-trained for imperceptibility and multiple watermarks.
Experiments show that WAM is competitive with state-of-the art methods in terms of imperceptibility and robustness, especially against inpainting and splicing, even on high-resolution images.
Moreover, it offers new capabilities: WAM can locate watermarked areas in spliced images and extract distinct 32-bit messages with less than 1 bit error from multiple small regions -- no larger than 10\% of the image surface -- even for small 256×256 images.
Training and inference code and model weights are available at https://github.com/facebookresearch/watermark-anything.
****
Prannay Kaul,Chengcheng Ma,Ismail Elezi,Jiankang Deng
From Attention to Activation: Unraveling the Enigmas of Large Language Models
We study two strange phenomena in auto-regressive Transformers: (1) the dominance of the ■rst token in attention heads; (2) the occurrence of large outlier activations in the hidden states. We ■nd that popular large language models, such as Llama attend maximally to the first token in 98% of attention heads, a behaviou

r we attribute to the softmax function. To mitigate this issue, we propose a ref
ormulation of softmax to softmax-1. Furthermore, we identify adaptive optimisers
, e.g. Adam, as the primary contributor to the large outlier activations and int
roduce OrthoAdam, a novel optimiser that utilises orthogonal matrices to transfo
rm gradients, to address this issue. Finally, not only do our methods prevent th
ese phenomena from occurring, but additionally, they enable Transformers to sust
ain their performance when quantised using basic algorithms, something that stan
dard methods are unable to do. In summary, our methods reduce the attention prop
ortion on the first token from 65% to 3.3%, the activation kurtosis in the hidde
n states from 1657 to 3.1, and perplexity penalty under 4-bit weight quantisatio
n from 3565 to 0.3. Code is available at https://github.com/prannaykaul/OrthoAda
m
****
Niklas Schmidinger,Lisa Schneckenreiter,Philipp Seidl,Johannes Schimunek,Pieter-
Jan Hoedt,Johannes Brandstetter,Andreas Mayr,Sohvi Luukkonen,Sepp Hochreiter,Gün
ter Klambauer
Bio-xLSTM: Generative modeling, representation and in-context learning of biolog
ical and chemical sequences
Language models for biological and chemical sequences enable crucial application
s such as drug discovery, protein engineering, and precision medicine. Currently
, these language models are predominantly based on Transformer architectures. Wh
ile Transformers have yielded impressive results, their quadratic runtime depend
ency on sequence length complicates their use for long genomic sequences and in-
context learning on proteins and chemical sequences. Recently, the recurrent xLS
TM architecture has been shown to perform favorably compared to Transformers and
 modern state-space models (SSMs) in the natural language domain. Similar to SSM
s, xLSTMs have linear runtime dependency and allow for constant-memory decoding
at inference time, which makes them prime candidates for modeling long-range dep
endencies in biological and chemical sequences. In this work, we tailor xLSTM to
wards these domains and we propose a suite of language models called Bio-xLSTM.
Extensive experiments in three large domains, genomics, proteins, and chemistry,
 were performed to assess xLSTM’s ability to model biological and chemical seque
nces. The results show that Bio-xLSTM is a highly proficient generative model fo
r DNA, protein, and chemical sequences, learns rich representations, and can per
form in-context learning for proteins and small molecules.
****
Jan Betley,Xuchan Bao,Martín Soto,Anna Sztyber-Betley,James Chua,Owain Evans
Tell me about yourself: LLMs are aware of their learned behaviors
We study *behavioral self-awareness*, which we define as an LLM's capability to
articulate its behavioral policies without relying on in-context examples. We fi
netune LLMs on examples that exhibit particular behaviors, including (a) making
risk-seeking / risk-averse economic decisions, and (b) making the user say a cer
tain word. Although these examples never contain explicit descriptions of the po
licy (e.g. "I will now take the risk-seeking option"), we find that the finetune
d LLMs can explicitly describe their policies through out-of-context reasoning.
We demonstrate LLMs' behavioral self-awareness across various evaluation tasks,
both for multiple-choice and free-form questions.
Furthermore, we demonstrate that models can correctly attribute different learne
d policies to distinct personas.
Finally, we explore the connection between behavioral self-awareness and the con
cept of backdoors in AI safety, where certain behaviors are implanted in a model
, often through data poisoning, and can be triggered under certain conditions. W
e find evidence that LLMs can recognize the existence of the backdoor-like behav
ior that they have acquired through fine-tuning.
****
Fredrik Carlsson,Fangyu Liu,Daniel Ward,Murathan Kurfali,Joakim Nivre
The Hyperfitting Phenomenon: Sharpening and Stabilizing LLMs for Open-Ended Text
 Generation
This paper introduces the counter-intuitive generalization results of overfittin
g pre-trained large language models (LLMs) on very small datasets. In the settin

g of open-ended text generation, it is well-documented that LLMs tend to generat
e repetitive and dull sequences, a phenomenon that is especially apparent when g
enerating using greedy decoding. This issue persists even with state-of-the-art
LLMs containing billions of parameters, trained via next-token prediction on lar
ge datasets. We find that by further fine-tuning these models to achieve a near-
zero training loss on a small set of samples -- a process we refer to as hyperfi
tting -- the long-sequence generative capabilities are greatly enhanced.
Greedy decoding with these Hyperfitted models even outperform Top-P sampling ove
r long-sequences, both in terms of diversity and human preferences. This phenome
non extends to LLMs of various sizes, different domains, and even autoregressive
 image generation. We further find this phenomena to be distinctly different fro
m that of Grokking and double descent. Surprisingly, our experiments indicate th
at hyperfitted models rarely fall into repeating sequences they were trained on,
 and even explicitly blocking these sequences results in high-quality output. Al
l hyperfitted models produce extremely low-entropy predictions, often allocating
 nearly all probability to a single token.
****
Jianyu Zhang,Niklas Nolte,Ranajoy Sadhukhan,Beidi Chen,Leon Bottou
Memory Mosaics
Memory Mosaics are networks of associative memories working in concert to achiev
e a prediction task of interest. Like transformers, memory mosaics possess compo
sitional capabilities and in-context learning capabilities. Unlike transformers,
 memory mosaics achieve these capabilities in comparatively transparent way (“pr
edictive disentanglement”). We illustrate these capabilities on a toy example an
d also show that memory mosaics perform as well or better than transformers on m
edium-scale language modeling tasks.
****
Buelent Uendes,Shujian Yu,Mark Hoogendoorn
Start Smart: Leveraging Gradients For Enhancing Mask-based XAI Methods
Mask-based explanation methods offer a powerful framework for interpreting deep
learning model predictions across diverse data modalities, such as images and ti
me series, in which the central idea is to identify an instance-dependent mask t
hat minimizes the performance drop from the resulting masked input. Different ob
jectives for learning such masks have been proposed, all of which, in our view,
can be unified under an information-theoretic framework that balances performanc
e degradation of the masked input with the complexity of the resulting masked re
presentation. Typically, these methods initialize the masks either uniformly or
as all-ones.
In this paper, we argue that an effective mask initialization strategy is as imp
ortant as the development of novel learning objectives, particularly in light of
 the significant computational costs associated with existing mask-based explana
tion methods. To this end, we introduce a new gradient-based initialization tech
nique called StartGrad, which is the first initialization method specifically de
signed for mask-based post-hoc explainability methods. Compared to commonly used
 strategies, StartGrad is provably superior at initialization in striking the af
orementioned trade-off. Despite its simplicity, our experiments demonstrate that
 StartGrad enhances the optimization process of various state-of-the-art mask-ex
planation methods by reaching target metrics faster and, in some cases, boosting
 their overall performance.
****
Julian Minder,Kevin Du,Niklas Stoehr,Giovanni Monea,Chris Wendler,Robert West,Ry
an Cotterell
Controllable Context Sensitivity and the Knob Behind It
When making predictions, a language model must trade off how much it relies on i
ts context vs. its prior knowledge.
Choosing how sensitive the model is to its context is a fundamental functionalit
y, as it enables the model to excel at tasks like retrieval-augmented generation
 and question-answering.
In this paper, we search for a knob which controls this sensitivity, determining
 whether language models answer from the context or their prior knowledge.

To guide this search, we design a task for controllable context sensitivity.
In this task, we first feed the model a context ("Paris is in England") and a qu
estion ("Where is Paris?"); we then instruct the model to either use its prior o
r contextual knowledge and evaluate whether it generates the correct answer for
both intents (either "France" or "England").
When fine-tuned on this task, instruct versions of Llama-3.1, Mistral-v0.3, and
Gemma-2 can solve it with high accuracy (85-95%).
Analyzing these high-performing models, we narrow down which layers may be impor
tant to context sensitivity using a novel linear time algorithm.
Then, in each model, we identify a 1-D subspace in a single layer that encodes w
hether the model follows context or prior knowledge.
Interestingly, while we identify this subspace in a fine-tuned model, we find th
at the exact same subspace serves as an effective knob in not only that model bu
t also non-fine-tuned instruct and base models of that model family.
Finally, we show a strong correlation between a model's performance and how dist
inctly it separates context-agreeing from context-ignoring answers in this subsp
ace.
These results suggest a single fundamental subspace facilitates how the model ch
ooses between context and prior knowledge.
****

Shiyang Zhang,Aakash Patel,Syed A Rizvi,Nianchen Liu,Sizhuang He,Amin Karbasi,Em
anuele Zappala,David van Dijk
Intelligence at the Edge of Chaos
We explore the emergence of intelligent behavior in artificial systems by invest
igating how the complexity of rule-based systems influences the capabilities of
models trained to predict these rules. Our study focuses on elementary cellular
automata (ECA), simple yet powerful one-dimensional systems that generate behavi
ors ranging from trivial to highly complex. By training distinct Large Language
Models (LLMs) on different ECAs, we evaluated the relationship between the compl
exity of the rules' behavior and the intelligence exhibited by the LLMs, as refl
ected in their performance on downstream tasks. Our findings reveal that rules w
ith higher complexity lead to models exhibiting greater intelligence, as demonst
rated by their performance on reasoning and chess move prediction tasks. Both un
iform and periodic systems, and often also highly chaotic systems, resulted in p
oorer downstream performance, highlighting a sweet spot of complexity conducive
to intelligence. We conjecture that intelligence arises from the ability to pred
ict complexity and that creating intelligence may require only exposure to compl
exity.
****

Han Sun,Rui Gong,Ismail Nejjar,Olga Fink
DynAlign: Unsupervised Dynamic Taxonomy Alignment for Cross-Domain Segmentation
Current unsupervised domain adaptation (UDA) methods for semantic segmentation t
ypically assume identical class labels between the source and target domains. Th
is assumption ignores the label-level domain gap, which is common in real-world
scenarios, and limits their ability to identify finer-grained or novel categorie
s without requiring extensive manual annotation.
A promising direction to address this limitation lies in recent advancements in
foundation models, which exhibit strong generalization abilities due to their ri
ch prior knowledge. However, these models often struggle with domain-specific nu
ances and underrepresented fine-grained categories.
To address these challenges, we introduce DynAlign, a two-stage framework that i
ntegrates UDA with foundation models to bridge both the image-level and label-le
vel domain gaps. Our approach leverages prior semantic knowledge to align source
 categories with target categories that can be novel, more fine-grained, or name
d differently. (e.g., vehicle to car, truck, bus). Foundation models are then em
ployed for precise segmentation and category reassignment. To further enhance ac
curacy, we propose a knowledge fusion approach that dynamically adapts to varyin
g scene contexts. DynAlign generates accurate predictions in a new target label
space without requiring any manual annotations, allowing seamless adaptation to
new taxonomies through either model retraining or direct inference.

Experiments on the GTA $\rightarrow$ IDD and GTA$\rightarrow$ Mapillary benchmarks validate the effectiveness of our approach, achieving a significant improvement over existing methods. Our code is publically available at https://github.com/hansunhayden/DynAlign.
****
In-Hwan Jin,Haesoo Choo,Seong-Hun Jeong,Park Heemoon,Junghwan Kim,Oh-joon Kwon,Kyeongbo Kong
Optimizing 4D Gaussians for Dynamic Scene Video from Single Landscape Images
To achieve realistic immersion in landscape images, fluids such as water and clouds need to move within the image while revealing new scenes from various camera perspectives. Recently, a field called dynamic scene video has emerged, which combines single image animation with 3D photography. These methods use pseudo 3D space, implicitly represented with Layered Depth Images (LDIs). LDIs separate a single image into depth-based layers, which enables elements like water and clouds to move within the image while revealing new scenes from different camera perspectives. However, as landscapes typically consist of continuous elements, including fluids, the representation of a 3D space separates a landscape image into discrete layers, and it can lead to diminished depth perception and potential distortions depending on camera movement. Furthermore, due to its implicit modeling of 3D space, the output may be limited to videos in the 2D domain, potentially reducing their versatility. In this paper, we propose representing a complete 3D space for dynamic scene video by modeling explicit representations, specifically 4D Gaussians, from a single image. The framework is focused on optimizing 3D Gaussians by generating multi-view images from a single image and creating 3D motion to optimize 4D Gaussians. The most important part of proposed framework is consistent 3D motion estimation, which estimates common motion among multi-view images to bring the motion in 3D space closer to actual motions. As far as we know, this is the first attempt that considers animation while representing a complete 3D space from a single landscape image. Our model demonstrates the ability to provide realistic immersion in various landscape images through diverse experiments and metrics. Extensive experimental results are https://anonymous.4open.science/r/ICLR_3D_MOM-7B9E/README.md.
****
Sayantan Dasgupta,Trevor Cohn
Improving Language Model Distillation through Hidden State Matching
Hidden State Matching is shown to improve knowledge distillation of language models by encouraging similarity between a student and its teacher's hidden states since DistilBERT. This typically uses a cosine loss, which restricts the dimensionality of the student to the teacher's, severely limiting the compression ratio. We present an alternative technique using Centered Kernel Alignment (CKA) to match hidden states of different dimensionality, allowing for smaller students and higher compression ratios. We show the efficacy of our method using encoder--decoder (BART, mBART \& T5) and encoder-only (BERT) architectures across a range of tasks from classification to summarization and translation. Our technique is competitive with the current state-of-the-art distillation methods at comparable compression rates and does not require already pretrained student models. It can scale to students smaller than the current methods, is no slower in training and inference, and is considerably more flexible. The code is available on github.
****
Jasmine Bayrooti,Carl Henrik Ek,Amanda Prorok
Efficient Model-Based Reinforcement Learning Through Optimistic Thompson Sampling
Learning complex robot behavior through interactions with the environment necessitates principled exploration. Effective strategies should prioritize exploring regions of the state-action space that maximize rewards, with optimistic exploration emerging as a promising direction aligned with this idea and enabling sample-efficient reinforcement learning. However, existing methods overlook a crucial aspect: the need for optimism to be informed by a belief connecting the reward and state. To address this, we propose a practical, theoretically grounded appro

ach to optimistic exploration based on Thompson sampling. Our approach is the f
irst that allows for reasoning about _joint_ uncertainty over transitions and re
wards for optimistic exploration. We apply our method on a set of MuJoCo and VMA
S continuous control tasks. Our experiments demonstrate that optimistic explorat
ion significantly accelerates learning in environments with sparse rewards, acti
on penalties, and difficult-to-explore regions. Furthermore, we provide insights
 into when optimism is beneficial and emphasize the critical role of model uncer
tainty in guiding exploration.
****

Masaru Isonuma,Ivan Titov
What's New in My Data? Novelty Exploration via Contrastive Generation
Fine-tuning is widely used to adapt language models for specific goals, often le
veraging real-world data such as patient records, customer-service interactions,
 or web content in languages not covered in pre-training.
These datasets are typically massive, noisy, and often confidential, making thei
r direct inspection challenging.
However, understanding them is essential for guiding model deployment and inform
ing decisions about data cleaning or suppressing any harmful behaviors learned d
uring fine-tuning.
In this study, we introduce the task of novelty discovery through generation, wh
ich aims to identify novel domains of a fine-tuning dataset by generating exampl
es that illustrate these properties.
Our approach ■ Contrastive Generative Exploration (CGE) ■ assumes no direct acce
ss to the data but instead relies on a pre-trained model and the same model afte
r fine-tuning.
By contrasting the predictions of these two models, CGE can generate examples th
at highlight novel domains of the fine-tuning data.
However, this simple approach may produce examples that are too similar to one a
nother, failing to capture the full range of novel domains present in the datase
t.
We address this by introducing an iterative version of CGE, where the previously
 generated examples are used to update the pre-trained model, and this updated m
odel is then contrasted with the fully fine-tuned model to generate the next exa
mple, promoting diversity in the generated outputs.
Our experiments demonstrate the effectiveness of CGE in detecting novel domains,
 such as toxic language, as well as new natural and programming languages.
Furthermore, we show that CGE remains effective even when models are fine-tuned
using differential privacy techniques.
****

Yulong Yang,Felix O'Mahony,Christine Allen-Blanchette
Learning Color Equivariant Representations
In this paper, we introduce group convolutional neural networks (GCNNs) equivari
ant to color variation. GCNNs have been designed for a variety of geometric tran
sformations from 2D and 3D rotation groups, to semi-groups such as scale. Despit
e the improved interpretability, accuracy and generalizability of these architec
tures, GCNNs have seen limited application in the context of perceptual quantiti
es. Notably, the recent CEConv network uses a GCNN to achieve equivariance to hu
e transformations by convolving input images with a hue rotated RGB filter. Howe
ver, this approach leads to invalid RGB values which break equivariance and degr
ade performance. We resolve these issues with a lifting layer that transforms th
e input image directly, thereby circumventing the issue of invalid RGB values an
d improving equivariance error by over three orders of magnitude. Moreover, we e
xtend the notion of color equivariance to include equivariance to saturation and
 luminance shift. Our hue-, saturation-, luminance- and color-equivariant networ
ks achieve strong generalization to out-of-distribution perceptual variations an
d improved sample efficiency over conventional architectures. We demonstrate the
 utility of our approach on synthetic and real world datasets where we consisten
tly outperform competitive baselines.
****

Song Wang,Peng Wang,Tong Zhou,Yushun Dong,Zhen Tan,Jundong Li

CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models
As Large Language Models (LLMs) are increasingly deployed to handle various natu
ral language processing (NLP) tasks, concerns regarding the potential negative s
ocietal impacts of LLM-generated content have also arisen. To evaluate the biase
s exhibited by LLMs, researchers have recently proposed a variety of datasets. H
owever, existing bias evaluation efforts often focus on only a particular type o
f bias and employ inconsistent evaluation metrics, leading to difficulties in co
mparison across different datasets and LLMs. To address these limitations, we co
llect a variety of datasets designed for the bias evaluation of LLMs, and furthe
r propose CEB, a Compositional Evaluation Bechmark that covers different types o
f bias across different social groups and tasks. The curation of CEB is based on
 our newly proposed compositional taxonomy, which characterizes each dataset fro
m three dimensions: bias types, social groups, and tasks. By combining the three
 dimensions, we develop a comprehensive evaluation strategy for the bias in LLMs
. Our experiments demonstrate that the levels of bias vary across these dimensio
ns, thereby providing guidance for the development of specific bias mitigation m
ethods.
****

Tianyi Zhang,Anshumali Shrivastava
LeanQuant: Accurate and Scalable Large Language Model Quantization with Loss-err
or-aware Grid
Large language models (LLMs) have shown immense potential across various domains
, but their high memory requirements and inference costs remain critical challen
ges for deployment. Post-training quantization (PTQ) has emerged as a promising
technique to reduce memory requirements and decoding latency. However, recent ac
curate quantization methods often depend on specialized computations or custom d
ata formats to achieve better model quality, which limits their compatibility wi
th popular frameworks, as they require dedicated inference kernels tailored to s
pecific hardware and software platforms, hindering wider adoption. Furthermore,
many competitive methods have high resource requirements and computational overh
ead for quantizing models, making it challenging to scale them to hundreds of bi
llions of parameters. In response to these challenges, we propose LeanQuant (Los
s-error-aware network Quantization), a novel quantization method that is accurat
e, versatile, and scalable. In the existing popular iterative loss-error-based q
uantization framework, we identify a critical limitation in prior methods: the m
in-max affine quantization grid fails to preserve model quality due to outliers
in inverse Hessian diagonals. To overcome this fundamental issue, we propose lea
rning loss-error-aware grids, instead of using non-adaptive min-max affine grids
. Our approach not only produces quantized models that are more accurate but als
o generalizes to a wider range of quantization types, including affine and non-u
niform quantization, enhancing compatibility with more frameworks. Extensive exp
eriments with recent LLMs demonstrate that LeanQuant is highly accurate, compari
ng favorably against competitive baselines in model quality, and scalable, achie
ving very accurate quantization of Llama-3.1 405B, one of the largest open-sourc
e LLMs to date, using two Quadro RTX 8000-48GB GPUs in 21 hours. Our code is ava
ilable at https://github.com/LeanModels/LeanQuant.
****

Xinyi Wang,Antonis Antoniades,Yanai Elazar,Alfonso Amayuelas,Alon Albalak,Kexun
Zhang,William Yang Wang
Generalization v.s. Memorization: Tracing Language Models’ Capabilities Back to
Pretraining Data
The impressive capabilities of large language models (LLMs) have sparked debate
over whether these models genuinely generalize to unseen tasks or predominantly
rely on memorizing vast amounts of pretraining data. To explore this issue, we i
ntroduce an extended concept of memorization, distributional memorization, which
 measures the correlation between the LLM output probabilities and the pretraini
ng data frequency. To effectively capture task-specific pretraining data frequen
cy, we propose a novel task-gram language model, which is built by counting the
co-occurrence of semantically related n-gram pairs from task inputs and output
s in the pretraining corpus. Using the Pythia models trained on the Pile dataset

, we evaluate four distinct tasks: machine translation, factual question answeri
ng, world knowledge understanding, and math reasoning. Our findings reveal varyi
ng levels of memorization, with the strongest effect observed in factual questio
n answering. Furthermore, while model performance improves across all tasks as L
LM size increases, only factual question answering shows an increase in memoriza
tion, whereas machine translation and reasoning tasks exhibit greater generaliza
tion, producing more novel outputs. This study demonstrates that memorization pl
ays a larger role in simpler, knowledge-intensive tasks, while generalization is
 the key for harder, reasoning-based tasks, providing a scalable method for anal
yzing large pretraining corpora in greater depth.
****

Xinyu Ye,Hao Xiong,Jianhao Huang,Ziang Chen,Jia Wang,Junchi Yan
On Designing General and Expressive Quantum Graph Neural Networks with Applicati
ons to MILP Instance Representation
Graph-structured data is ubiquitous, and graph learning models have recently bee
n extended to address complex problems like mixed-integer linear programming (MI
LP). However, studies have shown that the vanilla message-passing based graph ne
ural networks (GNNs) suffer inherent limitations in learning MILP instance repre
sentation, i.e., GNNs may map two different MILP instance graphs to the same rep
resentation. In this paper, we introduce an expressive quantum graph learning ap
proach, leveraging quantum circuits to recognize patterns that are difficult for
 classical methods to learn. Specifically, the proposed General Quantum Graph Le
arning Architecture (GQGLA) is composed of a node feature layer, a graph message
 interaction layer, and an optional auxiliary layer. Its generality is reflected
 in effectively encoding features of nodes and edges while ensuring node permuta
tion equivariance and flexibly creating different circuit structures for various
 expressive requirements and downstream tasks. GQGLA is well suited for learning
 complex graph tasks like MILP representation. Experimental results highlight th
e effectiveness of GQGLA in capturing and learning representations for MILPs. In
 comparison to traditional GNNs, GQGLA exhibits superior discriminative capabili
ties and demonstrates enhanced generalization across various problem instances,
making it a promising solution for complex graph tasks.
****

Huawen Feng,ZekunYao,Junhao Zheng,Qianli Ma
Training Large Language Models for Retrieval-Augmented Question Answering throug
h Backtracking Correction
Despite recent progress in Retrieval-Augmented Generation (RAG) achieved by larg
e language models (LLMs), retrievers often recall uncorrelated documents, regard
ed as "noise" during subsequent text generation. To address this, some methods t
rain LLMs to distinguish between relevant and irrelevant documents using labeled
 data, enabling them to select the most likely relevant ones as context. However
, they remain sensitive to noise, as LLMs can easily make mistakes when the sele
cted document is noisy. Some approaches increase the number of referenced docume
nts and train LLMs to perform stepwise reasoning when presented with multiple do
cuments. Unfortunately, these methods rely on extensive and diverse annotations
to ensure generalization, which is both challenging and costly. In this paper, w
e propose **Backtracking Correction** to address these limitations. Specifically
, we reformulate stepwise RAG into a multi-step decision-making process. Startin
g from the final step, we optimize the model through error sampling and self-cor
rection, and then backtrack to the previous state iteratively. In this way, the
model's learning scheme follows an easy-to-hard progression: as the target state
 moves forward, the context space decreases while the decision space increases.
Experimental results demonstrate that **Backtracking Correction** enhances LLMs'
 ability to make complex multi-step assessments, improving the robustness of RAG
 in dealing with noisy documents.
****

Alex Nguyen,Gautam Reddy
Differential learning kinetics govern the transition from memorization to genera
lization during in-context learning
Transformers exhibit in-context learning (ICL): the ability to use novel informa

tion presented in the context without additional weight updates. Recent work sho
ws that ICL emerges when models are trained on a sufficiently diverse set of tas
ks and the transition from memorization to generalization is sharp with increasi
ng task diversity. One interpretation is that a network's limited capacity to me
morize favors generalization. Here, we examine the mechanistic underpinnings of
this transition using a small transformer applied to a synthetic ICL task. Using
 theory and experiment, we show that the sub-circuits that memorize and generali
ze can be viewed as largely independent. The relative *rates* at which these sub
-circuits learn explains the transition from memorization to generalization, rat
her than capacity constraints. We uncover a memorization scaling law, which dete
rmines the task diversity threshold at which the network generalizes. The theory
 quantitatively explains a variety of other ICL-related phenomena, including the
 long-tailed distribution of when ICL is acquired, the bimodal behavior of solut
ions close to the task diversity threshold, the influence of contextual and data
 distributional statistics on ICL, and the transient nature of ICL.
****

Zhiyu Zhu,Zhibo Jin,Jiayu Zhang,NAN YANG,Jiahao Huang,Jianlong Zhou,Fang Chen
Narrowing Information Bottleneck Theory for Multimodal Image-Text Representation
s Interpretability
The task of identifying multimodal image-text representations has garnered incre
asing attention, particularly with models such as CLIP (Contrastive Language-Ima
ge Pretraining), which demonstrate exceptional performance in learning complex a
ssociations between images and text. Despite these advancements, ensuring the in
terpretability of such models is paramount for their safe deployment in real-wor
ld applications, such as healthcare. While numerous interpretability methods hav
e been developed for unimodal tasks, these approaches often fail to transfer eff
ectively to multimodal contexts due to inherent differences in the representatio
n structures. Bottleneck methods, well-established in information theory, have b
een applied to enhance CLIP's interpretability. However, they are often hindered
 by strong assumptions or intrinsic randomness. To overcome these challenges, we
 propose the Narrowing Information Bottleneck Theory, a novel framework that fun
damentally redefines the traditional bottleneck approach. This theory is specifi
cally designed to satisfy contemporary attribution axioms, providing a more robu
st and reliable solution for improving the interpretability of multimodal models
. In our experiments, compared to state-of-the-art methods, our approach enhance
s image interpretability by an average of 9\%, text interpretability by an avera
ge of 58.83\%, and accelerates processing speed by 63.95\%. Our code is publicly
 accessible at https://github.com/LMBTough/NIB.
****

Qian Li,Minghui Ouyang,Tian Ding,Yuyi Wang,Qingjiang Shi,Ruoyu Sun
Towards Explaining the Power of Constant-depth Graph Neural Networks for Structu
red Linear Programming
Graph neural networks (GNNs) have recently emerged as powerful tools for solving
 complex optimization problems, often being employed to approximate solution map
pings. Empirical evidence shows that even shallow GNNs (with fewer than ten laye
rs) can achieve strong performance in predicting optimal solutions to linear pro
gramming (LP) problems. This finding is somewhat counter-intuitive, as LPs are g
lobal optimization problems, while shallow GNNs predict based on local informati
on. Although previous theoretical results suggest that GNNs have the expressive
power to solve LPs, they require deep architectures whose depth grows at least p
olynomially with the problem size, and thus leave the underlying principle of th
is empirical phenomenon still unclear. In this paper, we examine this phenomenon
 through the lens of distributed computing and average-case analysis. We establi
sh that the expressive power of GNNs for LPs is closely related to well-studied
distributed algorithms for LPs. Specifically, we show that any d-round distrib
uted LP algorithm can be simulated by a d-depth GNN, and vice versa. In partic
ular, by designing a new distributed LP algorithm and then unrolling it, we prov
e that constant-depth, constant-width GNNs suffice to solve sparse binary LPs ef
fectively. Here, in contrast with previous analyses focusing on worst-case scena
rios, in which we show that GNN depth must increase with problem size by leverag

ing an impossibility result about distributed LP algorithms, our analysis shifts the focus to the average-case performance, and shows that constant GNN depth then becomes sufficient no matter how large the problem size is. Our theory is validated by numerical results.
****
Qixun Wang,Yifei Wang,Xianghua Ying,Yisen Wang
Can In-context Learning Really Generalize to Out-of-distribution Tasks?
In this work, we explore the mechanism of in-context learning (ICL) on out-of-distribution (OOD) tasks that were not encountered during training. To achieve this, we conduct synthetic experiments where the objective is to learn OOD mathematical functions through ICL using a GPT-2 model. We reveal that Transformers may struggle to learn OOD task functions through ICL. Specifically, ICL performance resembles implementing a function within the pretraining hypothesis space and optimizing it with gradient descent based on the in-context examples. Additionally, we investigate ICL's well-documented ability to learn unseen abstract labels in context. We demonstrate that such ability only manifests in the scenarios without distributional shifts and, therefore, may not serve as evidence of new-task-learning ability. Furthermore, we assess ICL's performance on OOD tasks when the model is pretrained on multiple tasks. Both empirical and theoretical analyses demonstrate the existence of the \textbf{low-test-error preference} of ICL, where it tends to implement the pretraining function that yields low test error in the testing context. We validate this through numerical experiments. This new theoretical result, combined with our empirical findings, elucidates the mechanism of ICL in addressing OOD tasks.
****
Junpeng Liu,Tianyue Ou,Yifan Song,Yuxiao Qu,Wai Lam,Chenyan Xiong,Wenhu Chen,Graham Neubig,Xiang Yue
Harnessing Webpage UIs for Text-Rich Visual Understanding
Text-rich visual understanding—the ability to interpret both textual content and visual elements within a scene—is crucial for multimodal large language models (MLLMs) to effectively interact with structured environments. We propose leveraging webpage UIs as a naturally structured and diverse data source to enhance MLLMs’ capabilities in this area. Existing approaches, such as rule-based extraction, multimodal model captioning, and rigid HTML parsing, are hindered by issues like noise, hallucinations, and limited generalization. To overcome these challenges, we introduce MultiUI, a dataset of 7.3 million samples spanning various UI types and tasks, structured using enhanced accessibility trees and task taxonomies. By scaling multimodal instructions from web UIs through LLMs, our dataset enhances generalization beyond web domains, significantly improving performance in document understanding, GUI comprehension, grounding, and advanced agent tasks. This demonstrates the potential of structured web data to elevate MLLMs’ proficiency in processing text-rich visual environments and generalizing across domains.
****
Zhihao He,Hang Yu,Zi Gong,Shizhan Liu,Jianguo Li,Weiyao Lin
Rodimus*: Breaking the Accuracy-Efficiency Trade-Off with Efficient Attentions
Recent advancements in Transformer-based large language models (LLMs) have set new standards in natural language processing. However, the classical softmax attention incurs significant computational costs, leading to a $O(T)$ complexity for per-token generation, where T represents the context length. This work explores reducing LLMs' complexity while maintaining performance by introducing Rodimus and its enhanced version, Rodimus$+$. Rodimus employs an innovative data-dependent tempered selection (DDTS) mechanism within a linear attention-based, purely recurrent framework, achieving significant accuracy while drastically reducing the memory usage typically associated with recurrent models. This method exemplifies semantic compression by maintaining essential input information with fixed-size hidden states. Building on this, Rodimus$+$ combines Rodimus with the innovative Sliding Window Shared-Key Attention (SW-SKA) in a hybrid approach, effectively leveraging the complementary semantic, token, and head compression techniques. Our experiments demonstrate that Rodimus$+$-1.6B, trained on 1 trillion toke

ns, achieves superior downstream performance against models trained on more toke
ns, including Qwen2-1.5B and RWKV6-1.6B, underscoring its potential to redefine
the accuracy-efficiency balance in LLMs. Model code and pre-trained checkpoints
are open-sourced at https://github.com/codefuse-ai/rodimus.
****

Naoya Hasegawa,Issei Sato
Multiplicative Logit Adjustment Approximates Neural-Collapse-Aware Decision Boun
dary Adjustment
Real-world data distributions are often highly skewed. This has spurred a growin
g body of research on long-tailed recognition, aimed at addressing the imbalance
 in training classification models. Among the methods studied, multiplicative lo
git adjustment (MLA) stands out as a simple and effective method. What theoretic
al foundation explains the effectiveness of this heuristic method?
We provide a justification for the effectiveness of MLA with the following two-s
tep process. First, we develop a theory that adjusts optimal decision boundaries
 by estimating feature spread on the basis of neural collapse. Second, we demons
trate that MLA approximates this optimal method. Additionally, through experimen
ts on long-tailed datasets, we illustrate the practical usefulness of MLA under
more realistic conditions. We also offer experimental insights to guide the tuni
ng of MLA hyperparameters.
****

Juyeon Heo,Miao Xiong,Christina Heinze-Deml,Jaya Narain
Do LLMs estimate uncertainty well in instruction-following?
Large language models (LLMs) could be valuable personal AI agents across various
 domains, provided they can precisely follow user instructions. However, recent
studies have shown significant limitations in LLMs' instruction-following capabi
lities, raising concerns about their reliability in high-stakes applications.
Accurately estimating LLMs' uncertainty in adhering to instructions is critical
to mitigating deployment risks. We present, to our knowledge, the first systemat
ic evaluation of the uncertainty estimation abilities of LLMs in the context of
instruction-following.
Our study identifies key challenges with existing instruction-following benchmar
ks, where multiple factors are entangled with uncertainty stems from instruction
-following, complicating the isolation and comparison across methods and models.
To address these issues, we introduce a controlled evaluation setup with two ben
chmark versions of data, enabling a comprehensive comparison of uncertainty esti
mation methods under various conditions.
Our findings show that existing uncertainty methods struggle, particularly when
models make subtle errors in instruction following. While internal model states
provide some improvement, they remain inadequate in more complex scenarios.
The insights from our controlled evaluation setups provide a crucial understandi
ng of LLMs' limitations and potential for uncertainty estimation in instruction-
following tasks, paving the way for more trustworthy AI agents.
****

Shuang Liang,Guido Montufar
Implicit Bias of Mirror Flow for Shallow Neural Networks in Univariate Regressio
n
We examine the implicit bias of mirror flow in least squares error regression wi
th wide and shallow neural networks. For a broad class of potential functions, w
e show that mirror flow exhibits lazy training and has the same implicit bias as
 ordinary gradient flow when the network width tends to infinity. For univariate
 ReLU networks, we characterize this bias through a variational problem in funct
ion space. Our analysis includes prior results for ordinary gradient flow as a s
pecial case and lifts limitations which required either an intractable adjustmen
t of the training data or networks with skip connections. We further introduce \
emph{scaled potentials} and show that for these, mirror flow still exhibits lazy
 training but is not in the kernel regime. For univariate networks with absolute
 value activations, we show that mirror flow with scaled potentials induces a ri
ch class of biases, which generally cannot be captured by an RKHS norm. A takeaw
ay is that whereas the parameter initialization determines how strongly the curv

ature of the learned function is penalized at different locations of the input s
pace, the scaled potential determines how the different magnitudes of the curvat
ure are penalized.
****

Minjae Jeong,Yechan Hwang,Jaejin Lee,Sungyoon Jung,Won Hwa Kim
HGM³: Hierarchical Generative Masked Motion Modeling with Hard Token Mining
Text-to-motion generation has significant potential in a wide range of applicati
ons including animation, robotics, and AR/VR. While recent works on masked motio
n models are promising, the task remains challenging due to the inherent ambigui
ty in text and the complexity of human motion dynamics. To overcome the issues,
we propose a novel text-to-motion generation framework that integrates two key c
omponents: Hard Token Mining (HTM) and a Hierarchical Generative Masked Motion M
odel (HGM³). Our HTM identifies and masks challenging regions in motion sequence
s and directs the model to focus on hard-to-learn components for efficacy. Concu
rrently, the hierarchical model uses a semantic graph to represent sentences at
different granularity, allowing the model to learn contextually feasible motions
. By leveraging a shared-weight masked motion model, it reconstructs the same se
quence under different conditioning levels and facilitates comprehensive learnin
g of complex motion patterns. During inference, the model progressively generate
s motions by incrementally building up coarse-to-fine details. Extensive experim
ents on benchmark datasets, including HumanML3D and KIT-ML, demonstrate that our
 method outperforms existing methods in both qualitative and quantitative measur
es for generating context-aware motions.
****

Yichen Xie,Chenfeng Xu,Chensheng Peng,Shuqi Zhao,Nhat Ho,Alexander T. Pham,Mingy
u Ding,Masayoshi Tomizuka,Wei Zhan
X-Drive: Cross-modality Consistent Multi-Sensor Data Synthesis for Driving Scena
rios
Recent advancements have exploited diffusion models for the synthesis of either
LiDAR point clouds or camera image data in driving scenarios. Despite their succ
ess in modeling single-modality data marginal distribution, there is an under- e
xploration in the mutual reliance between different modalities to describe com-
plex driving scenes. To fill in this gap, we propose a novel framework, X-DRIVE,
 to model the joint distribution of point clouds and multi-view images via a dua
l- branch latent diffusion model architecture. Considering the distinct geometri
cal spaces of the two modalities, X-DRIVE conditions the synthesis of each modal
ity on the corresponding local regions from the other modality, ensuring better
alignment and realism. To further handle the spatial ambiguity during denoising,
 we design the cross-modality condition module based on epipolar lines to adapti
vely learn the cross-modality local correspondence. Besides, X-DRIVE allows for
controllable generation through multi-level input conditions, including text, bo
unding box, image, and point clouds. Extensive results demonstrate the high-fide
lity synthetic results of X-DRIVE for both point clouds and multi-view images, a
dhering to input conditions while ensuring reliable cross-modality consistency.
Our code will be made publicly available at https://github.com/yichen928/X-Drive
.
****

Nikhil Vyas,Depen Morwani,Rosie Zhao,Itai Shapira,David Brandfonbrener,Lucas Jan
son,Sham M. Kakade
SOAP: Improving and Stabilizing Shampoo using Adam for Language Modeling
There is growing evidence of the effectiveness of Shampoo, a higher-order precon
ditioning method, over Adam in deep learning optimization tasks. However, Shampo
o's drawbacks include additional hyperparameters and computational overhead when
 compared to Adam, which only updates running averages of first- and second-mome
nt quantities. This work establishes a formal connection between Shampoo (implem
ented with the 1/2 power) and Adafactor --- a memory-efficient approximation of
Adam --- showing that Shampoo is equivalent to running Adafactor in the eigenbas
is of Shampoo's preconditioner. This insight leads to the design of a simpler an
d computationally efficient algorithm: **S**hampo**O** with **A**dam in the **P*
*reconditioner's eigenbasis (SOAP).

With regards to improving Shampoo's computational efficiency, the most straightf
orward approach would be to simply compute Shampoo's eigendecomposition less fre
quently. Unfortunately, as our empirical results show, this leads to performance
 degradation that worsens with this frequency. SOAP mitigates this degradation b
y continually updating the running average of the second moment, just as Adam do
es, but in the current (slowly changing) coordinate basis. Furthermore, since SO
AP is equivalent to running Adam in a rotated space, it introduces only one addi
tional hyperparameter (the preconditioning frequency) compared to Adam. We empir
ically evaluate SOAP on language model pre-training with 360m and 660m sized mod
els. In the large batch regime, SOAP reduces the number of iterations by over 40
\% and wall clock time by over 35\% compared to AdamW, with approximately 20\% i
mprovements in both metrics compared to Shampoo. An implementation of SOAP is av
ailable at https://github.com/nikhilvyas/SOAP.
****

Junmo Kang,Leonid Karlinsky,Hongyin Luo,Zhen Wang,Jacob A Hansen,James R. Glass,
David Daniel Cox,Rameswar Panda,Rogerio Feris,Alan Ritter
Self-MoE: Towards Compositional Large Language Models with Self-Specialized Expe
rts
We present Self-MoE, an approach that transforms a monolithic LLM into a composi
tional, modular system of self-specialized experts, named MiXSE (MiXture of Self
-specialized Experts). Our approach leverages self-specialization, which constru
cts expert modules using self-generated synthetic data, each equipping a shared
base LLM with distinct domain-specific capabilities, activated via self-optimize
d routing. This allows for dynamic and capability-specific handling of various t
arget tasks, enhancing overall capabilities, without extensive human-labeled dat
a and added parameters. Our empirical results reveal that specializing LLMs may
exhibit potential trade-offs in performances on non-specialized tasks. On the ot
her hand, our Self-MoE demonstrates substantial improvements (6.5%p on average)
over the base LLM across diverse benchmarks such as knowledge, reasoning, math,
and coding. It also consistently outperforms other methods, including instance m
erging and weight merging, while offering better flexibility and interpretabilit
y by design with semantic experts and routing. Our findings highlight the critic
al role of modularity, the applicability of Self-MoE to multiple base LLMs, and
the potential of self-improvement in achieving efficient, scalable, and adaptabl
e systems.
****

Xiaodong Chen,Yuxuan Hu,Jing Zhang,Yanling Wang,Cuiping Li,Hong Chen
Streamlining Redundant Layers to Compress Large Language Models
This paper introduces LLM-Streamline, a pioneer work on layer pruning for large
language models (LLMs). It is based on the observation that different layers hav
e varying impacts on hidden states, enabling the identification of less importan
t layers to be pruned.
LLM-Streamline comprises two parts: layer pruning, which removes consecutive lay
ers with the lowest importance based on target sparsity, and layer replacement,
a novel module that trains a lightweight network to replace the pruned layers to
 mitigate performance loss. Additionally, a new metric called stability is propo
sed to address the limitations of the widely used accuracy metric in evaluating
model compression. Experiments show that LLM-Streamline outperforms both previou
s and concurrent state-of-the-art pruning methods in terms of both performance a
nd training efficiency. Our code is available at \href{https://github.com/RUCKBR
easoning/LLM-Streamline}{this repository}.
****

Yilun Zheng,Xiang Li,Sitao Luan,Xiaojiang Peng,Lihui Chen
Let Your Features Tell The Differences: Understanding Graph Convolution By Featu
re Splitting
Graph Neural Networks (GNNs) have demonstrated strong capabilities in processing
 structured data. While traditional GNNs typically treat each feature dimension
equally important during graph convolution, we raise an important question: **Is
 the graph convolution operation equally beneficial for each feature?** If not,
the convolution operation on certain feature dimensions can possibly lead to har

mful effects, even worse than convolution-free models. Therefore, it is required to distinguish convolution-favored and convolution-disfavored features. Traditional feature selection methods mainly focus on identifying informative features or reducing redundancy, but they are not suitable for structured data as they overlook graph structures. In graph community, some studies have investigated the performance of GNN with respect to node features using feature homophily metrics, which assess feature consistency across graph topology. Unfortunately, these metrics do not effectively align with GNN performance and cannot be reliably used for feature selection in GNNs. To address these limitations, we introduce a novel metric, Topological Feature Informativeness (TFI), to distinguish GNN-favored and GNN-disfavored features, where its effectiveness is validated through both theoretical analysis and empirical observations. Based on TFI, we propose a simple yet effective Graph Feature Selection (GFS) method, which processes GNN-favored and GNN-disfavored features with GNNs and non-GNN models separately. Compared to original GNNs, GFS significantly improves the extraction of useful topological information from each feature with comparable computational costs. Extensive experiments show that after applying GFS to $\textbf{8}$ baseline and state-of-the-art (SOTA) GNN architectures across $\textbf{10}$ datasets, $\textbf{90\%}$ of the GFS-augmented cases show significant performance boosts. Furthermore, our proposed TFI metric outperforms other feature selection methods for GFS. These results verify the effectiveness of both GFS and TFI. Additionally, we demonstrate that GFS's improvements are robust to hyperparameter tuning, highlighting its potential as a universally valid method for enhancing various GNN architectures.
****
Kaiwen Zhou,Chengzhi Liu,Xuandong Zhao,Anderson Compalas,Dawn Song,Xin Eric Wang
Multimodal Situational Safety
Multimodal Large Language Models (MLLMs) are rapidly evolving, demonstrating impressive capabilities as multimodal assistants that interact with both humans and their environments. However, this increased sophistication introduces significant safety concerns. In this paper, we present the first evaluation and analysis of a novel safety challenge termed Multimodal Situational Safety, which explores how safety considerations vary based on the specific situation in which the user or agent is engaged.
We argue that for an MLLM to respond safely—whether through language or action—it often needs to assess the safety implications of a language query within its corresponding visual context.
To evaluate this capability, we develop the Multimodal Situational Safety benchmark (MSSBench) to assess the situational safety performance of current MLLMs.
The dataset comprises 1,960 language query-image pairs, half of which the image context is safe, and the other half is unsafe.
We also develop an evaluation framework that analyzes key safety aspects, including explicit safety reasoning, visual understanding, and, crucially, situational safety reasoning.
Our findings reveal that current MLLMs struggle with this nuanced safety problem in the instruction-following setting and struggle to tackle these situational safety challenges all at once, highlighting a key area for future research.
Furthermore, we develop multi-agent pipelines to coordinately solve safety challenges, which shows consistent improvement in safety over the original MLLM response.
****
Aymane El Firdoussi,Mohamed El Amine Seddik,Soufiane Hayou,Reda ALAMI,Ahmed Alzubaidi,Hakim Hacid
Maximizing the Potential of Synthetic Data: Insights from Random Matrix Theory
Synthetic data has gained attention for training large language models, but poor-quality data can harm performance (see, e.g., Shumailov et al. (2023); Seddik et al. (2024)). A potential solution is data pruning, which retains only high-quality data based on a score function (human or machine feedback). Previous work Feng et al. (2024) analyzed models trained on synthetic data as sample size increases. We extend this by using random matrix theory to derive the performance of a binary classifier trained on a mix of real and pruned synthetic data in a high

dimensional setting. Our findings identify conditions where synthetic data coul
d improve performance, focusing on the quality of the generative model and verif
ication strategy. We also show a smooth phase transition in synthetic label nois
e, contrasting with prior sharp behavior in infinite sample limits. Experiments
with toy models and large language models validate our theoretical results.
****

Paria Rashidinejad,Yuandong Tian
Sail into the Headwind: Alignment via Robust Rewards and Dynamic Labels against
Reward Hacking
Aligning AI systems with human preferences typically suffers from the infamous *
reward hacking* problem, where optimization of an imperfect reward model leads t
o undesired behaviors. In this paper, we investigate reward hacking in offline p
reference optimization, which aims to improve an initial model using a preferenc
e dataset. We identify two types of reward hacking stemming from statistical flu
ctuations in the dataset: Type I Reward Hacking due to subpar choices appearing
more favorable, and Type II Reward Hacking due to decent choices appearing less
desirable. We prove that many (mainstream or theoretical) preference optimizatio
n methods suffer from both types of reward hacking. To mitigate Type I Reward Ha
cking, we propose POWER, a new preference optimization method that combines Guia
su's weighted entropy with a robust reward maximization objective. POWER enjoys
finite-sample guarantees under general function approximation, competing with th
e best covered policy in the data. To mitigate Type II Reward Hacking, we analyz
e the learning dynamics of preference optimization and develop a novel technique
that dynamically updates preference labels toward certain "stationary labels",
resulting in diminishing gradients for untrustworthy samples. Empirically, POWER
with dynamic labels (DL) consistently outperforms state-of-the-art methods on a
lignment benchmarks, achieving improvements of up to **13.0** points on AlpacaEv
al 2 and **11.5** points on Arena-Hard over DPO, while also improving or maintai
ning performance on downstream tasks such as mathematical reasoning. Strong theo
retical guarantees and empirical results demonstrate the promise of POWER-DL in
mitigating reward hacking.
****

Yuhua Jiang,Qihan Liu,Yiqin Yang,Xiaoteng Ma,Dianyu Zhong,Hao Hu,Jun Yang,Bin Li
ang,Bo XU,Chongjie Zhang,Qianchuan Zhao
Episodic Novelty Through Temporal Distance
Exploration in sparse reward environments remains a significant challenge in rei
nforcement learning, particularly in Contextual Markov Decision Processes (CMDPs
), where environments differ across episodes. Existing episodic intrinsic motiva
tion methods for CMDPs primarily rely on count-based approaches, which are ineff
ective in large state spaces, or on similarity-based methods that lack appropria
te metrics for state comparison. To address these shortcomings, we propose Episo
dic Novelty Through Temporal Distance (ETD), a novel approach that introduces te
mporal distance as a robust metric for state similarity and intrinsic reward com
putation. By employing contrastive learning, ETD accurately estimates temporal d
istances and derives intrinsic rewards based on the novelty of states within the
current episode. Extensive experiments on various benchmark tasks demonstrate t
hat ETD significantly outperforms state-of-the-art methods, highlighting its eff
ectiveness in enhancing exploration in sparse reward CMDPs.
****

Taiwo Adebiyi,Bach Do,Ruda Zhang
Optimizing Posterior Samples for Bayesian Optimization via Rootfinding
Bayesian optimization devolves the global optimization of a costly objective fun
ction to the global optimization of a sequence of acquisition functions. This in
ner-loop optimization can be catastrophically difficult if it involves posterior
sample paths, especially in higher dimensions. We introduce an efficient global
optimization strategy for posterior samples based on global rootfinding. It pro
vides gradient-based optimizers with two sets of judiciously selected starting p
oints, designed to combine exploration and exploitation. The number of starting
points can be kept small without sacrificing optimization quality. Remarkably, e
ven with just one point from each set, the global optimum is discovered most of

the time. The algorithm scales practically linearly to high dimensions, breaking the curse of dimensionality. For Gaussian process Thompson sampling (GP-TS), we demonstrate remarkable improvement in both inner- and outer-loop optimization, surprisingly outperforming alternatives like EI and GP-UCB in most cases. Our approach also improves the performance of other posterior sample-based acquisition functions, such as variants of entropy search. Furthermore, we propose a sample-average formulation of GP-TS, which has a parameter to explicitly control exploitation and can be computed at the cost of one posterior sample.
****

Youngseok Kim,Sunwook Hwang,Hyung-Sin Kim,Saewoong Bahk
ConcreTizer: Model Inversion Attack via Occupancy Classification and Dispersion Control for 3D Point Cloud Restoration
The growing use of 3D point cloud data in autonomous vehicles (AVs) has raised serious privacy concerns, particularly due to the sensitive information that can be extracted from 3D data. While model inversion attacks have been widely studied in the context of 2D data, their application to 3D point clouds remains largely unexplored. To fill this gap, we present the first in-depth study of model inversion attacks aimed at restoring 3D point cloud scenes. Our analysis reveals the unique challenges, the inherent sparsity of 3D point clouds and the ambiguity between empty and non-empty voxels after voxelization, which are further exacerbated by the dispersion of non-empty voxels across feature extractor layers. To address these challenges, we introduce ConcreTizer, a simple yet effective model inversion attack designed specifically for voxel-based 3D point cloud data. ConcreTizer incorporates Voxel Occupancy Classification to distinguish between empty and non-empty voxels and Dispersion-Controlled Supervision to mitigate non-empty voxel dispersion. Extensive experiments on widely used 3D feature extractors and benchmark datasets, such as KITTI and Waymo, demonstrate that ConcreTizer concretely restores the original 3D point cloud scene from disrupted 3D feature data. Our findings highlight both the vulnerability of 3D data to inversion attacks and the urgent need for robust defense strategies.
****

Samuel Marks,Can Rager,Eric J Michaud,Yonatan Belinkov,David Bau,Aaron Mueller
Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models
We introduce methods for discovering and applying **sparse feature circuits**. These are causally implicated subnetworks of human-interpretable features for explaining language model behaviors. Circuits identified in prior work consist of polysemantic and difficult-to-interpret units like attention heads or neurons, rendering them unsuitable for many downstream applications. In contrast, sparse feature circuits enable detailed understanding of unanticipated mechanisms in neural networks. Because they are based on fine-grained units, sparse feature circuits are useful for downstream tasks: We introduce SHIFT, where we improve the generalization of a classifier by ablating features that a human judges to be task-irrelevant. Finally, we demonstrate an entirely unsupervised and scalable interpretability pipeline by discovering thousands of sparse feature circuits for automatically discovered model behaviors.
****

Huajian Xin,Z.Z. Ren,Junxiao Song,Zhihong Shao,Wanjia Zhao,Haocheng Wang,Bo Liu,Liyue Zhang,Xuan Lu,Qiushi Du,Wenjun Gao,Haowei Zhang,Qihao Zhu,Dejian Yang,Zhibin Gou,Z.F. Wu,Fuli Luo,Chong Ruan
DeepSeek-Prover-V1.5: Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search
Lean is an advanced proof assistant designed to facilitate formal theorem proving by providing a variety of interactive feedback. In this paper, we explore methodologies to leverage proof assistant feedback to augment the capabilities of large language models in constructing formal proofs. First, we deploy online reinforcement learning using Lean verification outcomes as the reward signal to improve the proof completion policy. This straightforward approach shows great promise in enhancing the model's alignment with the formal verification system. In addition, we propose RMaxTS, a variant of Monte-Carlo tree search that employs an i

ntrinsic-reward-driven exploration strategy to generate diverse proof paths. The tree structure is organized to represent the transitions of intermediate tactic states, extracted from the compilation messages given by Lean's tactic mode. The intrinsic reward is constructed to incentivize the discovery of novel tactic states, which helps to to mitigate the sparse-reward problem inherent in proof search. These techniques lead to a more efficient planning scheme for formal proof generation, achieving new state-of-the-art results on both miniF2F and ProofNet benchmarks.
****
Kai Ye,Liangcai Su,Chenxiong Qian
How Far Are We from True Unlearnability?
High-quality data plays an indispensable role in the era of large models, but the use of unauthorized data for model training greatly damages the interests of data owners. To overcome this threat, several unlearnable methods have been proposed, which generate unlearnable examples (UEs) by compromising the training availability of data. Clearly, due to unknown training purposes and the powerful representation learning capabilities of existing models, these data are expected to be unlearnable for various task models, i.e., they will not help improve the model's performance. However, unexpectedly, we find that on the multi-task dataset Taskonomy, UEs still perform well in tasks such as semantic segmentation, failing to exhibit \textit{cross-task unlearnability}. This phenomenon leads us to question: \textit{How far are we from attaining truly unlearnable examples?} We attempt to answer this question from the perspective of model optimization. To this end, we observe the difference in the convergence process between clean and poisoned models using a simple model architecture. Subsequently, from the loss landscape we find that only a part of the critical parameter optimization paths show significant differences, implying a close relationship between the loss landscape and unlearnability. Consequently, we employ the loss landscape to explain the underlying reasons for UEs and propose Sharpness-Aware Learnability (SAL) to quantify the unlearnability of parameters based on this explanation. Furthermore, we propose an Unlearnable Distance (UD) to measure the unlearnability of data based on the SAL distribution of parameters in clean and poisoned models. Finally, we conduct benchmark tests on mainstream unlearnable methods using the proposed UD, aiming to promote community awareness of the capability boundaries of existing unlearnable methods.
****
Yunfeng Diao,Baiqi Wu,Ruixuan Zhang,Ajian Liu,Xiaoshuai Hao,Xingxing Wei,Meng Wang,He Wang
TASAR: Transfer-based Attack on Skeletal Action Recognition
Skeletal sequence data, as a widely employed representation of human actions, are crucial in Human Activity Recognition (HAR). Recently, adversarial attacks have been proposed in this area, which exposes potential security concerns, and more importantly provides a good tool for model robustness test. Within this research, transfer-based attack is an important tool as it mimics the real-world scenario where an attacker has no knowledge of the target model, but is under-explored in Skeleton-based HAR (S-HAR). Consequently, existing S-HAR attacks exhibit weak adversarial transferability and the reason remains largely unknown. In this paper, we investigate this phenomenon via the characterization of the loss function. We find that one prominent indicator of poor transferability is the low smoothness of the loss function. Led by this observation, we improve the transferability by properly smoothening the loss when computing the adversarial examples. This leads to the first Transfer-based Attack on Skeletal Action Recognition, TASAR. TASAR explores the smoothened model posterior of pre-trained surrogates, which is achieved by a new post-train Dual Bayesian optimization strategy. Furthermore, unlike existing transfer-based methods which overlook the temporal coherence within sequences, TASAR incorporates motion dynamics into the Bayesian attack, effectively disrupting the spatial-temporal coherence of S-HARs. For exhaustive evaluation, we build the first large-scale robust S-HAR benchmark, comprising 7 S-HAR models, 10 attack methods, 3 S-HAR datasets and 2 defense models. Extensive results demonstrate the superiority of TASAR. Our benchmark enables easy com

parisons for future studies, with the code available in the https://github.com/yunfengdiao/Skeleton-Robustness-Benchmark.
****
Ying-yee Ava Lau,Zhiwen Shao,Dit-Yan Yeung
Fast and Slow Streams for Online Time Series Forecasting Without Information Leakage
Current research in online time series forecasting (OTSF) faces two significant issues. The first is information leakage, where models make predictions and are then evaluated on historical time steps that have already been used in backpropagation for parameter updates. The second is practicality: while forecasting in real-world applications typically emphasizes looking ahead and anticipating future uncertainties, prediction sequences in this setting include only one future step with the remaining being observed time points. This necessitates a redefinition of the OTSF setting, focusing on predicting unknown future steps and evaluating unobserved data points.
Following this new setting, challenges arise in leveraging incomplete pairs of ground truth and predictions for backpropagation, as well as in generalizing accurate information without overfitting to noise from recent data streams. To address these challenges, we propose a novel dual-stream framework for online forecasting (DSOF): a slow stream that updates with complete data using experience replay, and a fast stream that adapts to recent data through temporal difference learning. This dual-stream approach updates a teacher-student model learned through a residual learning strategy, generating predictions in a coarse-to-fine manner. Extensive experiments demonstrate its improvement in forecasting performance in changing environments. Our code is publicly available at https://github.com/yyalau/iclr2025_dsof.
****
Zhongwei Wan,Xinjian Wu,Yu Zhang,Yi Xin,Chaofan Tao,Zhihong Zhu,Xin Wang,Siqi Luo,Jing Xiong,Longyue Wang,Mi Zhang
D_{2}O: Dynamic Discriminative Operations for Efficient Long-Context Inference of Large Language Models
Efficient generative inference in Large Language Models (LLMs) is impeded by the growing memory demands of Key-Value (KV) cache, especially for longer sequences. Traditional KV Cache eviction strategies, which discard less critical KV-pairs based on attention scores, often degrade generation quality, leading to issues such as context loss or hallucinations. To address this, we introduce **D**ynamic **D**iscriminative **O**perations ($\mathbf{D_2 O}$), a novel method that optimizes KV cache size dynamically and discriminatively at two levels without fine-tuning, while preserving essential context. At **layer-level**, by observing the varying densities of attention weights between shallow and deep layers, we dynamically determine which layers should avoid excessive eviction via our proposed ***dynamic allocation strategy*** to minimize information loss. At **token-level**, for the eviction strategy in each layer, $\mathbf{D_2 O}$ innovatively incorporates a ***compensation mechanism*** that maintains a similarity threshold to re-discriminate the importance of currently discarded tokens, determining whether they should be recalled and merged with similar tokens. Extensive experiments on various benchmarks and LLM architectures have shown that $\mathbf{D_2 O}$ not only achieves significant memory savings and enhances inference throughput by more than 3$\times$ but also maintains high-quality long-text generation.
****
Zhenyu Yang,Yuhang Hu,Zemin Du,Dizhan Xue,Shengsheng Qian,Jiahong Wu,Fan Yang,Weiming Dong,Changsheng Xu
SVBench: A Benchmark with Temporal Multi-Turn Dialogues for Streaming Video Understanding
Despite the significant advancements of Large Vision-Language Models (LVLMs) on established benchmarks, there remains a notable gap in suitable evaluation regarding their applicability in the emerging domain of long-context streaming video understanding. Current benchmarks for video understanding typically emphasize isolated single-instance text inputs and fail to evaluate the capacity to sustain temporal reasoning throughout the entire duration of video streams. To address t

hese limitations, we introduce SVBench, a pioneering benchmark with temporal mul
ti-turn question-answering chains specifically designed to thoroughly assess the
 capabilities of streaming video understanding of current LVLMs. We design a sem
i-automated annotation pipeline to obtain 49,979 Question-Answer (QA) pairs of 1
,353 streaming videos, which includes generating QA chains that represent a seri
es of consecutive multi-turn dialogues over video segments and constructing temp
oral linkages between successive QA chains. Our experimental results, obtained f
rom 14 models in dialogue and streaming evaluations, reveal that while the close
d-source GPT-4o outperforms others, most open-source LVLMs struggle with long-co
ntext streaming video understanding. We also construct a StreamingChat model, wh
ich significantly outperforms open-source LVLMs on our SVBench and achieves comp
arable performance on diverse vision-language benchmarks. We expect SVBench to a
dvance the research of streaming video understanding by providing a comprehensiv
e and in-depth analysis of current LVLMs. Our benchmark and model can be accesse
d at https://yzy-bupt.github.io/SVBench.
****
Harshit Varma,Dheeraj Mysore Nagaraj,Karthikeyan Shanmugam
Glauber Generative Model: Discrete Diffusion Models via Binary Classification
We introduce the Glauber Generative Model (GGM), a new class of discrete diffusi
on models, to obtain new samples from a distribution given samples from a discre
te space. GGM deploys a discrete Markov chain called the heat bath dynamics (or
the Glauber dynamics) to denoise a sequence of noisy tokens to a sample from a j
oint distribution of discrete tokens. Our novel conceptual framework provides an
 exact reduction of the task of learning the denoising Markov chain to solving a
 class of binary classification tasks. More specifically, the model learns to cl
assify a given token in a noisy sequence as signal or noise. In contrast, prior
works on discrete diffusion models either solve regression problems to learn imp
ortance ratios, or minimize loss functions given by variational approximations.
 We apply GGM to language modeling and image generation, where images are discre
tized using image tokenizers like VQGANs. We show that it outperforms existing d
iscrete diffusion models in language generation, and demonstrates strong perform
ance for image generation without using dataset-specific image tokenizers. We al
so show that our model is capable of performing well in zero-shot control settin
gs like text and image infilling.
****
Wenkai Yang,Shiqi Shen,Guangyao Shen,Wei Yao,Yong Liu,Gong Zhi,Yankai Lin,Ji-Ron
g Wen
Super(ficial)-alignment: Strong Models May Deceive Weak Models in Weak-to-Strong
 Generalization
Superalignment, where humans act as weak supervisors for superhuman models, has
become a crucial problem with the rapid development of Large Language Models (LL
Ms). Recent work has preliminarily studied this problem by using weak models to
supervise strong models, and discovered that weakly supervised strong students c
an consistently outperform weak teachers towards the alignment target, leading t
o a weak-to-strong generalization phenomenon. However, we are concerned that beh
ind such a promising phenomenon, whether there exists an issue of weak-to-strong
 deception, where strong models deceive weak models by exhibiting well-aligned i
n areas known to weak models but producing misaligned behaviors in cases weak mo
dels do not know. We take an initial step towards exploring this security issue
in a specific but realistic multi-objective alignment case, where there may be s
ome alignment targets conflicting with each other (e.g., helpfulness v.s. harmle
ssness). We aim to explore whether, in such cases, strong models might deliberat
ely make mistakes in areas known to them but unknown to weak models within one a
lignment dimension, in exchange for a higher reward in another dimension. Throug
h extensive experiments in both the reward modeling and preference optimization
scenarios, we find: (1) The weak-to-strong deception phenomenon exists across al
l settings. (2) The deception intensifies as the capability gap between weak and
 strong models increases. (3) Bootstrapping with an intermediate model can mitig
ate the deception to some extent, though its effectiveness remains limited. Our
work highlights the urgent need to pay more attention to the true reliability of

superalignment.
****
Kai Li,Wendi Sang,Chang Zeng,Runxuan Yang,Guo Chen,Xiaolin Hu
SonicSim: A customizable simulation platform for speech processing in moving sound source scenarios
Systematic evaluation of speech separation and enhancement models under moving sound source conditions requires extensive and diverse data. However, real-world datasets often lack sufficient data for training and evaluation, and synthetic datasets, while larger, lack acoustic realism. Consequently, neither effectively meets practical needs. To address this issue, we introduce SonicSim, a synthetic toolkit based on the embodied AI simulation platform Habitat-sim, designed to generate highly customizable data for moving sound sources. SonicSim supports multi-level adjustments—including scene-level, microphone-level, and source-level—enabling the creation of more diverse synthetic data. Leveraging SonicSim, we constructed a benchmark dataset called SonicSet, utilizing LibriSpeech, Freesound Dataset 50k (FSD50K), Free Music Archive (FMA), and 90 scenes from Matterport3D to evaluate speech separation and enhancement models. Additionally, to investigate the differences between synthetic and real-world data, we selected 5 hours of raw, non-reverberant data from the SonicSet validation set and recorded a real-world speech separation dataset, providing a reference for comparing SonicSet with other synthetic datasets. For speech enhancement, we utilized the real-world dataset RealMAN to validate the acoustic gap between SonicSet and existing synthetic datasets. The results indicate that models trained on SonicSet generalize better to real-world scenarios compared to other synthetic datasets. Code is publicly available at ***https://cslikai.cn/SonicSim/***.
****
Junfeng Fang,Houcheng Jiang,Kun Wang,Yunshan Ma,Jie Shi,Xiang Wang,Xiangnan He,Tat-Seng Chua
AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models
Large language models (LLMs) often exhibit hallucinations, producing incorrect or outdated knowledge. Hence, model editing methods have emerged to enable targeted knowledge updates. To achieve this, a prevailing paradigm is the locating-then-editing approach, which first locates influential parameters and then edits them by introducing a perturbation. While effective, current studies have demonstrated that this perturbation inevitably disrupt the originally preserved knowledge within LLMs, especially in sequential editing scenarios.
To address this, we introduce AlphaEdit, a novel solution that projects perturbation onto the null space of the preserved knowledge before applying it to the parameters. We theoretically prove that this projection ensures the output of post-edited LLMs remains unchanged when queried about the preserved knowledge, thereby mitigating the issue of disruption.
Extensive experiments on various LLMs, including LLaMA3, GPT2-XL, and GPT-J, show that AlphaEdit boosts the performance of most locating-then-editing methods by an average of 36.7% with a single line of additional code for projection solely.
****
Litu Rout,Yujia Chen,Nataniel Ruiz,Constantine Caramanis,Sanjay Shakkottai,Wen-Sheng Chu
Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations
Generative models transform random noise into images, while their inversion aims to reconstruct structured noise for recovery and editing.
This paper addresses two key tasks: (i) *inversion* and (ii) *editing* of real images using stochastic equivalents of rectified flow models (e.g., Flux).
While Diffusion Models (DMs) dominate the field of generative modeling for images, their inversion suffers from faithfulness and editability challenges due to nonlinear drift and diffusion.
Existing DM inversion methods require costly training of additional parameters or test-time optimization of latent variables.
Rectified Flows (RFs) offer a promising alternative to DMs, yet their inversion

remains underexplored.
We propose RF inversion using dynamic optimal control derived via a linear quadr
atic regulator, and prove that the resulting vector field is equivalent to a rec
tified stochastic differential equation.
We further extend our framework to design a stochastic sampler for Flux.
Our method achieves state-of-the-art performance in zero-shot inversion and edit
ing, surpassing prior works in stroke-to-image synthesis and semantic image edit
ing, with large-scale human evaluations confirming user preference.
See our project page https://rf-inversion.github.io/ for code and demo.
****

Tianhao Walter Wu,Jing Yang,Zhilin Guo,Jingyi Wan,Fangcheng Zhong,Cengiz Oztirel
i
Gaussian Head & Shoulders: High Fidelity Neural Upper Body Avatars with Anchor G
aussian Guided Texture Warping
The ability to reconstruct realistic and controllable upper body avatars from ca
sual monocular videos is critical for various applications in communication and
entertainment. By equipping the most recent 3D Gaussian Splatting representation
 with head 3D morphable models (3DMM), existing methods manage to create head av
atars with high fidelity. However, most existing methods only reconstruct a head
 without the body, substantially limiting their application scenarios. We found
that naively applying Gaussians to model the clothed chest and shoulders tends t
o result in blurry reconstruction and noisy floaters under novel poses. This is
because of the fundamental limitation of Gaussians and point clouds -- each Gaus
sian or point can only have a single directional radiance without spatial varian
ce, therefore an unnecessarily large number of them is required to represent com
plicated spatially varying texture, even for simple geometry. In contrast, we pr
opose to model the body part with a neural texture that consists of coarse and p
ose-dependent fine colors. To properly render the body texture for each view and
 pose without accurate geometry nor UV mapping, we optimize another sparse set o
f Gaussians as anchors that constrain the neural warping field that maps image p
lane coordinates to the texture space. We demonstrate that Gaussian Head & Shoul
ders can fit the high-frequency details on the clothed upper body with high fide
lity and potentially improve the accuracy and fidelity of the head region. We ev
aluate our method with casual phone-captured and internet videos and show our me
thod archives superior reconstruction quality and robustness in both self and cr
oss reenactment tasks. To fully utilize the efficient rendering speed of Gaussia
n splatting, we additionally propose an accelerated inference method of our trai
ned model without Multi-Layer Perceptron (MLP) queries and reach a stable render
ing speed of around 130 FPS for any subjects.
****

Armin W Thomas,Rom Parnichkun,Alexander Amini,Stefano Massaroli,Michael Poli
STAR: Synthesis of Tailored Architectures
Iterative improvement of model architectures is fundamental to deep learning: Tr
ansformers first enabled scaling, and recent advances in model hybridization hav
e pushed the quality-efficiency frontier. However, optimizing architectures rema
ins challenging and expensive, with a variety of automated or manual approaches
that fall short, due to limited progress in the design of search spaces and due
to the simplicity of resulting patterns and heuristics. In this work, we propose
 a new approach for the synthesis of tailored architectures (STAR). Our approach
 combines a novel search space based on the theory of linear input-varying syste
ms, supporting a hierarchical numerical encoding into architecture genomes. STAR
 genomes are automatically refined and recombined with gradient-free, evolutiona
ry algorithms to optimize for multiple model quality and efficiency metrics. Usi
ng STAR, we optimize large populations of new architectures, leveraging diverse
computational units and interconnection patterns, improving over highly-optimize
d Transformers and striped hybrid models on the frontier of quality, parameter s
ize, and inference cache for autoregressive language modeling.
****

Runjia Li,Qiwei Di,Quanquan Gu
Unified Convergence Analysis for Score-Based Diffusion Models with Deterministic

Samplers
Score-based diffusion models have emerged as powerful techniques for generating samples from high-dimensional data distributions. These models involve a two-phase process: first, injecting noise to transform the data distribution into a known prior distribution, and second, sampling to recover the original data distribution from noise. Among the various sampling methods, deterministic samplers stand out for their enhanced efficiency. However, analyzing these deterministic samplers presents unique challenges, as they preclude the use of established techniques such as Girsanov's theorem, which are only applicable to stochastic samplers. Furthermore, existing analysis for deterministic samplers usually focuses on specific examples, lacking a generalized approach for general forward processes and various deterministic samplers. Our paper addresses these limitations by introducing a unified convergence analysis framework. To demonstrate the power of our framework, we analyze the variance-preserving (VP) forward process with the exponential integrator (EI) scheme, achieving iteration complexity of $\tilde{O}(d^2/\epsilon)$.
Additionally, we provide a detailed analysis of Denoising Diffusion Implicit Models (DDIM)-type samplers, which have been underexplored in previous research, achieving polynomial iteration complexity.
****
Yinghui Li,Haojing Huang,Jiayi Kuang,Yangning Li,Shu-Yu Guo,Chao Qu,Xiaoyu Tan,Hai-Tao Zheng,Ying Shen,Philip S. Yu
Refine Knowledge of Large Language Models via Adaptive Contrastive Learning
How to alleviate the hallucinations of Large Language Models (LLMs) has always been the fundamental goal pursued by the LLMs research community. Looking through numerous hallucination-related studies, a mainstream category of methods is to reduce hallucinations by optimizing the knowledge representation of LLMs to change their output. Considering that the core focus of these works is the knowledge acquired by models, and knowledge has long been a central theme in human societal progress, we believe that the process of models refining knowledge can greatly benefit from the way humans learn. In our work, by imitating the human learning process, we design an Adaptive Contrastive Learning strategy. Our method flexibly constructs different positive and negative samples for contrastive learning based on LLMs' actual mastery of knowledge. This strategy helps LLMs consolidate the correct knowledge they already possess, deepen their understanding of the correct knowledge they have encountered but not fully grasped, forget the incorrect knowledge they previously learned, and honestly acknowledge the knowledge they lack. Extensive experiments and detailed analyses on widely used datasets demonstrate the effectiveness and competitiveness of our method.
****
Aritra Bhowmik,Pascal Mettes,Martin R. Oswald,Cees G. M. Snoek
Union-over-Intersections: Object Detection beyond Winner-Takes-All
This paper revisits the problem of predicting box locations in object detection architectures. Typically, each box proposal or box query aims to directly maximize the intersection-over-union score with the ground truth, followed by a winner-takes-all non-maximum suppression where only the highest scoring box in each region is retained. We observe that both steps are sub-optimal: the first involves regressing proposals to the entire ground truth, which is a difficult task even with large receptive fields, and the second neglects valuable information from boxes other than the top candidate. Instead of regressing proposals to the whole ground truth, we propose a simpler approach—regress only to the area of intersection between the proposal and the ground truth. This avoids the need for proposals to extrapolate beyond their visual scope, improving localization accuracy. Rather than adopting a winner-takes-all strategy, we take the union over the regressed intersections of all boxes in a region to generate the final box outputs. Our plug-and-play method integrates seamlessly into proposal-based, grid-based, and query-based detection architectures with minimal modifications, consistently improving object localization and instance segmentation. We demonstrate its broad applicability and versatility across various detection and segmentation tasks.

****
Gouki Minegishi,Hiroki Furuta,Yusuke Iwasawa,Yutaka Matsuo
Rethinking Evaluation of Sparse Autoencoders through the Representation of Polysemous Words
Sparse autoencoders (SAEs) have gained a lot of attention as a promising tool to improve the interpretability of large language models (LLMs) by mapping the complex superposition of *polysemantic* neurons into *monosemantic* features and composing a sparse dictionary of words.

However, traditional performance metrics like Mean Squared Error and L_{0} sparsity ignore the evaluation of the semantic representational power of SAEs - whether they can acquire interpretable monosemantic features while preserving the semantic relationship of words.For instance, it is not obvious whether a learned sparse feature could distinguish different meanings in one word.

In this paper, we propose a suite of evaluations for SAEs to analyze the quality of monosemantic features by focusing on polysemous words.
Our findings reveal that SAEs developed to improve the MSE-L_0 Pareto frontier may confuse interpretability, which does not necessarily enhance the extraction of monosemantic features.
The analysis of SAEs with polysemous words can also figure out the internal mechanism of LLMs; deeper layers and the Attention module contribute to distinguishing polysemy in a word.

Our semantics-focused evaluation offers new insights into the polysemy and the existing SAE objective and contributes to the development of more practical SAEs.
****
Yufan Zhou,Zhaobo Qi,Lingshuai Lin,Junqi Jing,Tingting Chai,Beichen Zhang,Shuhui Wang,Weigang Zhang
Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos
In this paper, we address the challenge of procedure planning in instructional videos, aiming to generate coherent and task-aligned action sequences from start and end visual observations. Previous work has mainly relied on text-level supervision to bridge the gap between observed states and unobserved actions, but it struggles with capturing intricate temporal relationships among actions. Building on these efforts, we propose the Masked Temporal Interpolation Diffusion (MTID) model that introduces a latent space temporal interpolation module within the diffusion model. This module leverages a learnable interpolation matrix to generate intermediate latent features, thereby augmenting visual supervision with richer mid-state details. By integrating this enriched supervision into the model, we enable end-to-end training tailored to task-specific requirements, significantly enhancing the model's capacity to predict temporally coherent action sequences. Additionally, we introduce an action-aware mask projection mechanism to restrict the action generation space, combined with a task-adaptive masked proximity loss to prioritize more accurate reasoning results close to the given start and end states over those in intermediate steps. Simultaneously, it filters out task-irrelevant action predictions, leading to contextually aware action sequences. Experimental results across three widely used benchmark datasets demonstrate that our MTID achieves promising action planning performance on most metrics.
****
Peng Xia,Siwei Han,Shi Qiu,Yiyang Zhou,Zhaoyang Wang,Wenhao Zheng,Zhaorun Chen,Chenhang Cui,Mingyu Ding,Linjie Li,Lijuan Wang,Huaxiu Yao
MMIE: Massive Multimodal Interleaved Comprehension Benchmark for Large Vision-Language Models
Interleaved multimodal comprehension and generation, enabling models to produce and interpret both images and text in arbitrary sequences, have become a pivotal area in multimodal learning. Despite significant advancements, the evaluation of this capability remains insufficient. Existing benchmarks suffer from limitations in data scale, scope, and evaluation depth, while current evaluation metrics

are often costly or biased, lacking in reliability for practical applications. To address these challenges, we introduce MMIE, a large-scale knowledge-intensive benchmark for evaluating interleaved multimodal comprehension and generation in Large Vision-Language Models (LVLMs). MMIE comprises 20K meticulously curated multimodal queries, spanning 3 categories, 12 fields, and 102 subfields, including mathematics, coding, physics, literature, health, and arts. It supports both interleaved inputs and outputs, offering a mix of multiple-choice and open-ended question formats to evaluate diverse competencies. Moreover, we propose a reliable automated evaluation metric, leveraging a scoring model fine-tuned with human-annotated data and systematic evaluation criteria, aimed at reducing bias and improving evaluation accuracy. Extensive experiments demonstrate the effectiveness of our benchmark and metrics in providing a comprehensive evaluation of interleaved LVLMs. Specifically, we evaluate eight LVLMs, revealing that even the best models show significant room for improvement, with most achieving only moderate results. We believe MMIE will drive further advancements in the development of interleaved LVLMs.
****
Paul Albert,Frederic Z. Zhang,Hemanth Saratchandran,Cristian Rodriguez-Opazo,Anton van den Hengel,Ehsan Abbasnejad
RandLoRA: Full rank parameter-efficient fine-tuning of large models
Low-Rank Adaptation (LoRA) and its variants have shown impressive results in reducing the number of trainable parameters and memory requirements of large transformer networks while maintaining fine-tuning performance. The low-rank nature of the weight update inherently limits the representation power of fine-tuned models, however, thus potentially compromising performance on complex tasks. This raises a critical question: when a performance gap between LoRA and standard fine-tuning is observed, is it due to the reduced number of train-
able parameters or the rank deficiency? This paper aims to answer this question by introducing RandLoRA, a parameter-efficient method that performs full-rank updates using a learned linear combinations of low-rank, non-trainable random matrices. Our method limits the number of trainable parameters by restricting optimization to diagonal scaling matrices applied to the fixed random matrices. This allows us to effectively overcome the low-rank limitations while maintaining parameter and memory efficiency during training. Through extensive experimen-
tation across vision, language, and vision-language benchmarks, we systematically evaluate the limitations of LoRA and existing random basis methods. Our findings reveal that full-rank updates are beneficial across vision and language tasks individually, and even more so for vision-language tasks, where RandLoRA significantly reduces—and sometimes eliminates—the performance gap between standard fine-tuning and LoRA, demonstrating its efficacy.
****
Cong Wei,Zheyang Xiong,Weiming Ren,Xeron Du,Ge Zhang,Wenhu Chen
OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision
Instruction-guided image editing methods have demonstrated significant potential by training diffusion models on automatically synthesized or manually annotated image editing pairs. However, these methods remain far from practical, real-life applications. We identify three primary challenges contributing to this gap. Firstly, existing models have limited editing skills due to the biased synthesis process. Secondly, these methods are trained with datasets with a high volume of noise and artifacts. This is due to the application of simple filtering methods like CLIP-score. Thirdly, all these datasets are restricted to a single low resolution and fixed aspect ratio, limiting the versatility to handle real-world use cases.
In this paper, we present OmniEdit, which is an omnipotent editor to handle seven different image editing tasks with any aspect ratio seamlessly. Our contribution is in four folds: (1) OmniEdit is trained by utilizing the supervision from seven different specialist models to ensure task coverage. (2) we utilize importance sampling based on the scores provided by large multimodal models (like GPT-4o) instead of CLIP-score to improve the data quality. (3) we propose a new editi

ng architecture called EditNet to greatly boost the editing success rate, (4) we provide images with different aspect ratios to ensure that our model can handle any image in the wild. We have curated a test set containing images of differen t aspect ratios, accompanied by diverse instructions to cover different tasks. B oth automatic evaluation and human evaluations demonstrate that OmniEdit can sig nificantly outperforms all the existing models.
****

Fantine Huot,Reinald Kim Amplayo,Jennimaria Palomaki,Alice Shoshana Jakobovits,E lizabeth Clark,Mirella Lapata
Agents' Room: Narrative Generation through Multi-step Collaboration
Writing compelling fiction is a multifaceted process combining elements such as crafting a plot, developing interesting characters, and using evocative language . While large language models (LLMs) show promise for story writing, they curren tly rely heavily on intricate prompting, which limits their use. We propose Agen ts' Room, a generation framework inspired by narrative theory, that decomposes n arrative writing into subtasks tackled by specialized agents. To illustrate our method, we introduce Tell Me A Story, a high-quality dataset of complex writing prompts and human-written stories, and a novel evaluation framework designed spe cifically for assessing long narratives. We show that Agents' Room generates sto ries that are preferred by expert evaluators over those produced by baseline sys tems by leveraging collaboration and specialization to decompose the complex sto ry writing task into tractable components. We provide extensive analysis with au tomated and human-based metrics of the generated output.
****

Eunsu Baek,Sung-hwan Han,Taesik Gong,Hyung-Sin Kim
Adaptive Camera Sensor for Vision Models
Domain shift remains a persistent challenge in deep-learning-based computer visi on, often requiring extensive model modifications or large labeled datasets to a ddress. Inspired by human visual perception, which adjusts input quality through corrective lenses rather than over-training the brain, we propose Lens, a novel camera sensor control method that enhances model performance by capturing high-quality images from the model’s perspective, rather than relying on traditional human-centric sensor control. Lens is lightweight and adapts sensor parameters t o specific models and scenes in real-time. At its core, Lens utilizes VisiT, a t raining-free, model-specific quality indicator that evaluates individual unlabel ed samples at test time using confidence scores, without additional adaptation c osts. To validate Lens, we introduce ImageNet-ES Diverse, a new benchmark datase t capturing natural perturbations from varying sensor and lighting conditions. E xtensive experiments on both ImageNet-ES and our new ImageNet-ES Diverse show th at Lens significantly improves model accuracy across various baseline schemes fo r sensor control and model modification, while maintaining low latency in image captures. Lens effectively compensates for large model size differences and inte grates synergistically with model improvement techniques. Our code and dataset a re available at github.com/Edw2n/Lens.git.
****

Qi Li,Zhenyu Zhang,Lei Yao,Zhaoxia Li,Tianyi Zhong,Yong Zhang
Diffusion-based Decoupled Deterministic and Uncertain Framework for Probabilisti c Multivariate Time Series Forecasting
Diffusion-based denoising models have demonstrated impressive performance in pro babilistic forecasting for multivariate time series (MTS). Nonetheless, existing approaches often model the entire data distribution, neglecting the variability in uncertainty across different components of the time series. This paper intro duces a Diffusion-based Decoupled Deterministic and Uncertain ($\mathrm{D^3U}$) framework for probabilistic MTS forecasting. The framework integrates non-probab ilistic forecasting with conditional diffusion generation, enabling both accurat e point predictions and probabilistic forecasting. $\mathrm{D^3U}$ utilizes a po int forecasting model to non-probabilistically model high-certainty components i n the time series, generating embedded representations that are conditionally in jected into a diffusion model. To better model high-uncertainty components, a pa tch-based denoising network (PatchDN) is designed in the conditional diffusion m

odel. Designed as a plug-and-play framework, $\mathrm{D^3U}$ can be seamlessly i
ntegrated into existing point forecasting models to provide probabilistic foreca
sting capabilities. It can also be applied to other conditional diffusion method
s that incorporate point forecasting models. Experiments on six real-world datas
ets demonstrate that our method achieves over a 20\% improvement in both point a
nd probabilistic forecasting performance in MTS long-term forecasting compared t
o state-of-the-art (SOTA) probabilistic forecasting methods. Additionally, exten
sive ablation studies further validate the effectiveness of the $\mathrm{D^3U}$
framework.
****
Alex Lewandowski,Micha■ Bortkiewicz,Saurabh Kumar,András György,Dale Schuurmans,
Mateusz Ostaszewski,Marlos C. Machado
Learning Continually by Spectral Regularization
Loss of plasticity is a phenomenon where neural networks can become more difficu
lt to train over the course of learning. Continual learning algorithms seek to m
itigate this effect by sustaining good performance while maintaining network tra
inability. We develop a new technique for improving continual learning inspired
by the observation that the singular values of the neural network parameters at
initialization are an important factor for trainability during early phases of l
earning. From this perspective, we derive a new spectral regularizer for continu
al learning that better sustains these beneficial initialization properties thro
ughout training. In particular, the regularizer keeps the maximum singular value
 of each layer close to one. Spectral regularization directly ensures that gradi
ent diversity is maintained throughout training, which promotes continual traina
bility, while minimally interfering with performance in a single task. We presen
t an experimental analysis that shows how the proposed spectral regularizer can
sustain trainability and performance across a range of model architectures in co
ntinual supervised and reinforcement learning settings. Spectral regularization
is less sensitive to hyperparameters while demonstrating better training in indi
vidual tasks, sustaining trainability as new tasks arrive, and achieving better
generalization performance..
****
Martin Pawelczyk,Jimmy Z. Di,Yiwei Lu,Gautam Kamath,Ayush Sekhari,Seth Neel
Machine Unlearning Fails to Remove Data Poisoning Attacks
We revisit the efficacy of several practical methods for approximate machine unl
earning developed for large-scale deep learning. In addition to complying with d
ata deletion requests, one often-cited potential application for unlearning meth
ods is to remove the effects of poisoned data. We experimentally demonstrate tha
t, while existing unlearning methods have been demonstrated to be effective in a
 number of settings, they fail to remove the effects of data poisoning across a
variety of types of poisoning attacks (indiscriminate, targeted, and a newly-int
roduced Gaussian poisoning attack) and models (image classifiers and LLMs); even
 when granted a relatively large compute budget. In order to precisely character
ize unlearning efficacy, we introduce new evaluation metrics for unlearning base
d on data poisoning. Our results suggest that a broader perspective, including a
 wider variety of evaluations, are required to avoid a false sense of confidence
 in machine unlearning procedures for deep learning without provable guarantees.
 Moreover, while unlearning methods show some signs of being useful to efficient
ly remove poisoned data without having to retrain, our work suggests that these
methods are not yet ``ready for prime time,'' and currently provide limited bene
fit over retraining.
****
Nikhila Ravi,Valentin Gabeur,Yuan-Ting Hu,Ronghang Hu,Chaitanya Ryali,Tengyu Ma,
Haitham Khedr,Roman Rädle,Chloe Rolland,Laura Gustafson,Eric Mintun,Junting Pan,
Kalyan Vasudev Alwala,Nicolas Carion,Chao-Yuan Wu,Ross Girshick,Piotr Dollar,Chr
istoph Feichtenhofer
SAM 2: Segment Anything in Images and Videos
We present Segment Anything Model 2 (SAM 2), a foundation model towards solving
promptable visual segmentation in images and videos. We build a data engine, whi
ch improves model and data via user interaction, to collect the largest video se

gmentation dataset to date. Our model is a simple transformer architecture with streaming memory for real-time video processing. SAM 2 trained on our data provides strong performance across a wide range of tasks. In video segmentation, we observe better accuracy, using 3x fewer interactions than prior approaches. In image segmentation, our model is more accurate and 6x faster than the Segment Anything Model (SAM). We believe that our data, model, and insights will serve as a significant milestone for video segmentation and related perception tasks. We are releasing our main model, the dataset, an interactive demo and code.
****

Dongkwan Kim,Alice Oh

Generalizing Weisfeiler-Lehman Kernels to Subgraphs

Subgraph representation learning has been effective in solving various real-world problems. However, current graph neural networks (GNNs) produce suboptimal results for subgraph-level tasks due to their inability to capture complex interactions within and between subgraphs. To provide a more expressive and efficient alternative, we propose WLKS, a Weisfeiler-Lehman (WL) kernel generalized for subgraphs by applying the WL algorithm on induced k-hop neighborhoods. We combine kernels across different k-hop levels to capture richer structural information that is not fully encoded in existing models. Our approach can balance expressiveness and efficiency by eliminating the need for neighborhood sampling. In experiments on eight real-world and synthetic benchmarks, WLKS significantly outperforms leading approaches on five datasets while reducing training time, ranging from 0.01x to 0.25x compared to the state-of-the-art.
****

Ziyi Ye,Xiangsheng Li,Qiuchi Li,Qingyao Ai,Yujia Zhou,Wei Shen,Dong Yan,Yiqun LIU

Learning LLM-as-a-Judge for Preference Alignment

Learning from preference feedback is a common practice for aligning large language models (LLMs) with human value. Conventionally, preference data is learned and encoded into a scalar reward model that connects a value head with an LLM to produce a scalar score as preference. However, scalar models lack interpretability and are known to be susceptible to biases in datasets. This paper investigates leveraging LLM itself to learn from such preference data and serve as a judge to address both limitations in one shot. Specifically, we prompt the pre-trained LLM to generate initial judgment pairs with contrastive preference in natural language form. The self-generated contrastive judgment pairs are used to train the LLM-as-a-Judge with Direct Preference Optimization (DPO) and incentivize its reasoning capability as a judge. This proposal of learning the LLMas-a-Judge using self-generated Contrastive judgments (Con-J) ensures natural interpretability through the generated rationales supporting the judgments, and demonstrates higher robustness against bias compared to scalar models. Experimental results show that Con-J outperforms the scalar reward model trained on the same collection of preference data, and outperforms a series of open-source and closed-source generative LLMs. We open-source the training process and model weights of Con-J at https://github.com/YeZiyi1998/Con-J.
****

Haotian Zhang,Mingfei Gao,Zhe Gan,Philipp Dufter,Nina Wenzel,Forrest Huang,Dhruti Shah,Xianzhi Du,Bowen Zhang,Yanghao Li,Sam Dodge,Keen You,Zhen Yang,Aleksei Timofeev,Mingze Xu,Hong-You Chen,Jean-Philippe Fauconnier,Zhengfeng Lai,Haoxuan You,Zirui Wang,Afshin Dehghan,Peter Grasch,Yinfei Yang

MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning

We present MM1.5, a new family of multimodal large language models (MLLMs) designed to enhance capabilities in text-rich image understanding, visual referring and grounding, and multi-image reasoning. Building upon the MM1 architecture, MM1.5 adopts a data-centric approach to model training, systematically exploring the impact of diverse data mixtures across the entire model training lifecycle. This includes high-quality OCR data and synthetic captions for continual pre-training, as well as an optimized visual instruction-tuning data mixture for supervised fine-tuning. Our models range from 1B to 30B parameters, encompassing both dense and mixture-of-experts (MoE) variants, and demonstrate that careful data cur

ation and training strategies can yield strong performance even at small scales (1B and 3B). Additionally, we introduce two specialized variants: MM1.5-Video, d esigned for video understanding, and MM1.5-UI, tailored for mobile UI understand ing. Through extensive empirical studies and ablations, we provide detailed insi ghts into the training processes and decisions that inform our final designs, of fering valuable guidance for future research in MLLM development.
****

Youssef Allouah,Joshua Kazdan,Rachid Guerraoui,Sanmi Koyejo
The Utility and Complexity of In- and Out-of-Distribution Machine Unlearning
Machine unlearning, the process of selectively removing data from trained models , is increasingly crucial for addressing privacy concerns and knowledge gaps pos t-deployment. Despite this importance, existing approaches are often heuristic a nd lack formal guarantees. In this paper, we analyze the fundamental utility, ti me, and space complexity trade-offs of approximate unlearning, providing rigorou s certification analogous to differential privacy. For in-distribution forget da ta—data similar to the retain set—we show that a surprisingly simple and general procedure, empirical risk minimization with output perturbation, achieves tight unlearning-utility-complexity trade-offs, addressing a previous theoretical gap on the separation from unlearning ``for free" via differential privacy, which i nherently facilitates the removal of such data. However, such techniques fail wi th out-of-distribution forget data—data significantly different from the retain set—where unlearning time complexity can exceed that of retraining, even for a s ingle sample. To address this, we propose a new robust and noisy gradient descen t variant that provably amortizes unlearning time complexity without compromisin g utility.
****

Lukas Aichberger,Kajetan Schweighofer,Mykyta Ielanskyi,Sepp Hochreiter
Improving Uncertainty Estimation through Semantically Diverse Language Generatio n
Large language models (LLMs) can suffer from hallucinations when generating text . These hallucinations impede various applications in society and industry by ma king LLMs untrustworthy. Current LLMs generate text in an autoregressive fashion by predicting and appending text tokens. When an LLM is uncertain about the sem antic meaning of the next tokens to generate, it is likely to start hallucinatin g. Thus, it has been suggested that predictive uncertainty is one of the main ca uses of hallucinations. We introduce Semantically Diverse Language Generation (S DLG) to quantify predictive uncertainty in LLMs. SDLG steers the LLM to generate semantically diverse yet likely alternatives for an initially generated text. T his approach provides a precise measure of aleatoric semantic uncertainty, detec ting whether the initial text is likely to be hallucinated. Experiments on quest ion-answering tasks demonstrate that SDLG consistently outperforms existing meth ods while being the most computationally efficient, setting a new standard for u ncertainty estimation in LLMs.
****

Sicong Liu,Yang Shu,Chenjuan Guo,Bin Yang
Learning Generalizable Skills from Offline Multi-Task Data for Multi-Agent Coope ration
Learning cooperative multi-agent policy from offline multi-task data that can ge neralize to unseen tasks with varying numbers of agents and targets is an attrac tive problem in many scenarios. Although aggregating general behavior patterns a mong multiple tasks as skills to improve policy transfer is a promising approach , two primary challenges hinder the further advancement of skill learning in off line multi-task MARL. Firstly, extracting general cooperative behaviors from var ious action sequences as common skills lacks bringing cooperative temporal knowl edge into them. Secondly, existing works only involve common skills and can not adaptively choose independent knowledge as task-specific skills in each task for fine-grained action execution. To tackle these challenges, we propose Hierarchi cal and Separate Skill Discovery (HiSSD), a novel approach for generalizable off line multi-task MARL through skill learning. HiSSD leverages a hierarchical fram ework that jointly learns common and task-specific skills. The common skills lea

rn cooperative temporal knowledge and enable in-sample exploitation for offline multi-task MARL. The task-specific skills represent the priors of each task and achieve a task-guided fine-grained action execution. To verify the advancement of our method, we conduct experiments on multi-agent MuJoCo and SMAC benchmarks. After training the policy using HiSSD on offline multi-task data, the empirical results show that HiSSD assigns effective cooperative behaviors and obtains superior performance in unseen tasks.
****
Isha Chaudhary,Qian Hu,Manoj Kumar,Morteza Ziyadi,Rahul Gupta,Gagandeep Singh
Certifying Counterfactual Bias in LLMs
Large Language Models (LLMs) can produce biased responses that can cause representational harms. However, conventional studies are insufficient to thoroughly evaluate biases across LLM responses for different demographic groups (a.k.a. counterfactual bias), as they do not scale to large number of inputs and do not provide guarantees. Therefore, we propose the first framework, LLMCert-B that certifies LLMs for counterfactual bias on distributions of prompts. A certificate consists of high-confidence bounds on the probability of unbiased LLM responses for any set of counterfactual prompts - prompts differing by demographic groups, sampled from a distribution. We illustrate counterfactual bias certification for distributions of counterfactual prompts created by applying prefixes sampled from prefix distributions, to a given set of prompts. We consider prefix distributions consisting random token sequences, mixtures of manual jailbreaks, and perturbations of jailbreaks in LLM’s embedding space. We generate non-trivial certificates for SOTA LLMs, exposing their vulnerabilities over distributions of prompts generated from computationally inexpensive prefix distributions.
****
Fan Yin,Philippe Laban,XIANGYU PENG,Yilun Zhou,Yixin Mao,Vaibhav Vats,Linnea Ross,Divyansh Agarwal,Caiming Xiong,Chien-Sheng Wu
BingoGuard: LLM Content Moderation Tools with Risk Levels
Malicious content generated by large language models (LLMs) can pose varying degrees of harm.
Although existing LLM-based moderators can detect harmful content, they struggle to assess risk levels and may miss lower-risk outputs.
Accurate risk assessment allows platforms with different safety thresholds to tailor content filtering and rejection. In this paper, we introduce per-topic severity rubrics for 11 harmful topics and build BingoGuard, an LLM-based moderation system designed to predict both binary safety labels and severity levels.
To address the lack of annotations on levels of severity, we propose a scalable generate-then-filter framework that first generates responses across different severity levels and then filters out low-quality responses. Using this framework, we create BingoGuardTrain, a training dataset with 54,897 examples covering a variety of topics, response severity, styles, and BingoGuardTest, a test set with 988 examples explicitly labeled based on our severity rubrics that enables fine-grained analysis on model behaviors on different severity levels. Our BingoGuard-8B, trained on BingoGuardTrain, achieves the state-of-the-art performance on several moderation benchmarks, including WildGuardTest and HarmBench, as well as BingoGuardTest, outperforming best public models, WildGuard, by 4.3\%. Our analysis demonstrates that incorporating severity levels into training significantly enhances detection performance and enables the model to effectively gauge the severity of harmful responses. Warning: this paper includes red-teaming examples that may be harmful in nature.
****
Chunlei Li,Yilei Shi,Jingliang Hu,Xiao Xiang Zhu,Lichao Mou
Scale-Aware Contrastive Reverse Distillation for Unsupervised Medical Anomaly Detection
Unsupervised anomaly detection using deep learning has garnered significant rese

arch attention due to its broad applicability, particularly in medical imaging w
here labeled anomalous data are scarce. While earlier approaches leverage genera
tive models like autoencoders and generative adversarial networks (GANs), they o
ften fall short due to overgeneralization. Recent methods explore various strate
gies, including memory banks, normalizing flows, self-supervised learning, and k
nowledge distillation, to enhance discrimination. Among these, knowledge distill
ation, particularly reverse distillation, has shown promise. Following this para
digm, we propose a novel scale-aware contrastive reverse distillation model that
 addresses two key limitations of existing reverse distillation methods: insuffi
cient feature discriminability and inability to handle anomaly scale variations.
 Specifically, we introduce a contrastive student-teacher learning approach to d
erive more discriminative representations by generating and exploring out-of-nor
mal distributions. Further, we design a scale adaptation mechanism to softly wei
ght contrastive distillation losses at different scales to account for the scale
 variation issue. Extensive experiments on benchmark datasets demonstrate state-
of-the-art performance, validating the efficacy of the proposed method. The code
 will be made publicly available.
****

Hao Liu,Wilson Yan,Matei Zaharia,Pieter Abbeel
World Model on Million-Length Video And Language With Blockwise RingAttention
Enabling long-context understanding remains a key challenge in scaling existing
sequence models -- a crucial component in developing generally intelligent model
s that can process and operate over long temporal horizons that potentially cons
ist of millions of tokens. In this paper, we aim to address these challenges by
providing a comprehensive exploration of the full development process for produc
ing 1M context language models and video-language models, setting new benchmarks
 in language retrieval and new capabilities in long video understanding. We deta
il our long context data curation process, progressive context extension from 4K
 to 1M tokens, and present an efficient open-source implementation for scalable
training on long sequences. Additionally, we open-source a family of 7B paramete
r models capable of processing long text documents and videos exceeding 1M token
s.
****

Zhiyuan Zhou,Andy Peng,Qiyang Li,Sergey Levine,Aviral Kumar
Efficient Online Reinforcement Learning Fine-Tuning Need Not Retain Offline Data
The modern paradigm in machine learning involves pre-training on diverse data, f
ollowed by task-specific fine-tuning. In reinforcement learning (RL), this trans
lates to learning via offline RL on a diverse historical dataset, followed by ra
pid online RL fine-tuning using interaction data. Most RL fine-tuning methods re
quire continued training on offline data for stability and performance. However,
 this is undesirable because training on diverse offline data is slow and expens
ive for large datasets, and should, in principle, also limit the performance imp
rovement possible because of constraints or pessimism on offline data. In this p
aper, we show that retaining offline data is unnecessary as long as we use a pro
perly-designed online RL approach for fine-tuning offline RL initializations. To
 build this approach, we start by analyzing the role of retaining offline data i
n online fine-tuning. We find that continued training on offline data is mostly
useful for preventing a sudden divergence in the value function at the onset of
fine-tuning, caused by a distribution mismatch between the offline data and onli
ne rollouts. This divergence typically results in unlearning and forgetting the
benefits of offline pre-training. Our approach, Warm-start RL (WSRL), mitigates
the catastrophic forgetting of pre-trained initializations using a very simple i
dea. WSRL employs a warmup phase that seeds the online RL run with a very small
number of rollouts from the pre-trained policy to do fast online RL. The data co
llected during warmup bridges the distribution mismatch, and helps ``recalibrate
'' the offline Q-function to the online distribution, allowing us to completely
discard offline data without destabilizing the online RL fine-tuning. We show th
at WSRL is able to fine-tune without retaining any offline data, and is able to
learn faster and attains higher performance than existing algorithms irrespectiv
e of whether they do or do not retain offline data.

****
Dezhan Tu,Danylo Vashchilenko,Yuzhe Lu,Panpan Xu
VL-Cache: Sparsity and Modality-Aware KV Cache Compression for Vision-Language Model Inference Acceleration
Vision-Language Models (VLMs) have demonstrated impressive performance across a versatile set of tasks. A key challenge in accelerating VLMs is storing and accessing the large Key-Value (KV) cache that encodes long visual contexts, such as images or videos. While existing KV cache compression methods are effective for Large Language Models (LLMs), directly migrating them to VLMs yields suboptimal accuracy and speedup. To bridge the gap, we propose VL-Cache, a novel KV cache compression recipe tailored for accelerating VLM inference. In this paper, we first investigate the unique sparsity pattern of VLM attention by distinguishing visual and text tokens in prefill and decoding phases. Based on these observations, we introduce a layer-adaptive sparsity-aware cache budget allocation method that effectively distributes the limited cache budget across different layers, further reducing KV cache size without compromising accuracy. Additionally, we develop a modality-aware token scoring policy to better evaluate the token importance. Empirical results on multiple benchmark datasets demonstrate that retaining only 10% of KV cache achieves accuracy comparable to that with full cache. In a speed benchmark, our method accelerates end-to-end latency of generating 100 tokens by up to 2.33x and speeds up decoding by up to 7.08x, while reducing the memory footprint of KV cache in GPU by 90%.
****
Mingyuan Zhou,Zhendong Wang,Huangjie Zheng,Hai Huang
Guided Score identity Distillation for Data-Free One-Step Text-to-Image Generation
Diffusion-based text-to-image generation models trained on extensive text-image pairs have demonstrated the ability to produce photorealistic images aligned with textual descriptions. However, a significant limitation of these models is their slow sample generation process, which requires iterative refinement through the same network. To overcome this, we introduce a data-free guided distillation method that enables the efficient distillation of pretrained Stable Diffusion models without access to the real training data, often restricted due to legal, privacy, or cost concerns. This method enhances Score identity Distillation (SiD) with Long and Short Classifier-Free Guidance (LSG), an innovative strategy that applies Classifier-Free Guidance (CFG) not only to the evaluation of the pretrained diffusion model but also to the training and evaluation of the fake score network. We optimize a model-based explicit score matching loss using a score-identity-based approximation alongside our proposed guidance strategies for practical computation. By exclusively training with synthetic images generated by its one-step generator, our data-free distillation method rapidly improves FID and CLIP scores, achieving state-of-the-art FID performance while maintaining a competitive CLIP score. Notably, the one-step distillation of Stable Diffusion 1.5 achieves an FID of **8.15** on the COCO-2014 validation set, a record low value under the data-free setting. Our code and checkpoints are available at https://github.com/mingyuanzhou/SiD-LSG.
****
Kanika Madan,Alex Lamb,Emmanuel Bengio,Glen Berseth,Yoshua Bengio
Towards Improving Exploration through Sibling Augmented GFlowNets
Exploration is a key factor for the success of an active learning agent, especially when dealing with sparse extrinsic terminal rewards and long trajectories. We introduce Sibling Augmented Generative Flow Networks (SA-GFN), a novel framework designed to enhance exploration and training efficiency of Generative Flow Networks (GFlowNets). SA-GFN uses a decoupled dual network architecture, comprising of a main Behavior Network and an exploratory Sibling Network, to enable a diverse exploration of the underlying distribution using intrinsic rewards. Inspired by the ideas on exploration from reinforcement learning, SA-GFN provides a general-purpose exploration and learning paradigm that integrates with multiple GFlowNet training objectives and is especially helpful for exploration over a wide range of sparse or low reward distributions and task structures. An extensive se

t of experiments across a diverse range of tasks, reward structures and trajecto
ry lengths, along with a thorough set of ablations, demonstrate the superior per
formance of SA-GFN in terms of exploration efficacy and convergence speed as com
pared to the existing methods. In addition, SA-GFN's versatility and compatibili
ty with different GFlowNet training objectives and intrinsic reward methods unde
rscores its broad applicability in various problem domains.
****

Guang Zhao,Byung-Jun Yoon,Gilchan Park,Shantenu Jha,Shinjae Yoo,Xiaoning Qian
Pareto Prompt Optimization
Natural language prompt optimization, or prompt engineering, has emerged as a po
werful technique to unlock the potential of Large Language Models (LLMs) for var
ious tasks. While existing methods primarily focus on maximizing a single task-s
pecific performance metric for LLM outputs, real-world applications often requir
e considering trade-offs between multiple objectives. In this work, we address t
his limitation by proposing an effective technique for multi-objective prompt op
timization for LLMs. Specifically, we propose **ParetoPrompt**, a reinforcement
learning~(RL) method that leverages dominance relationships between prompts to d
erive a policy model for prompts optimization using preference-based loss functi
ons. By leveraging multi-objective dominance relationships, ParetoPrompt enables
 efficient exploration of the entire Pareto front without the need for a predefi
ned scalarization of multiple objectives. Our experimental results show that Par
etoPrompt consistently outperforms existing algorithms that use specific objecti
ve values. ParetoPrompt also yields robust performances when the objective metri
cs differ between training and testing.
****

Ze Yang,Shichao Dong,Ruibo Li,Nan Song,Guosheng Lin
ADAPT: Attentive Self-Distillation and Dual-Decoder Prediction Fusion for Contin
ual Panoptic Segmentation
Panoptic segmentation, which unifies semantic and instance segmentation into a s
ingle task, has witnessed considerable success on predefined tasks. However, tra
ditional methods tend to struggle with catastrophic forgetting and poor generali
zation when learning from a continuous stream of new tasks. While continual lear
ning aims to mitigate these challenges, our study reveals that existing continua
l panoptic segmentation (CPS) methods often suffer from efficiency or scalabilit
y issues. To address these limitations, we propose an efficient adaptation frame
work that incorporates attentive self-distillation and dual-decoder prediction f
usion to efficiently preserve prior knowledge while facilitating model generaliz
ation. Specifically, we freeze the majority of model weights, enabling a shared
forward pass between the teacher and student models during distillation. Attenti
ve self-distillation then adaptively distills useful knowledge from the old clas
ses without being distracted from non-object regions, which effectively enhances
 knowledge retention. Additionally, query-level fusion (QLF) is devised to seaml
essly integrate the output of the dual decoders without incurring scale inconsis
tency. Our method achieves state-of-the-art performance on ADE20K and COCO bench
marks. Code is available at https://github.com/Ze-Yang/ADAPT.
****

Jiani Huang,Ziyang Li,Mayur Naik,Ser-Nam Lim
LASER: A Neuro-Symbolic Framework for Learning Spatio-Temporal Scene Graphs with
 Weak Supervision
Supervised approaches for learning spatio-temporal scene graphs (STSG) from vide
o are greatly hindered due to their reliance on STSG-annotated videos, which are
 labor-intensive to construct at scale. Is it feasible to instead use readily av
ailable video captions as weak supervision? To address this question, we propose
 LASER, a neuro-symbolic framework to enable training STSG generators using only
 video captions. LASER employs large language models to first extract logical sp
ecifications with rich spatio-temporal semantic information from video captions.
 LASER then trains the underlying STSG generator to align the predicted STSG wit
h the specification. The alignment algorithm overcomes the challenges of weak su
pervision by leveraging a differentiable symbolic reasoner and using a combinati
on of contrastive, temporal, and semantics losses. The overall approach efficien

tly trains low-level perception models to extract a fine-grained STSG that confo
rms to the video caption. In doing so, it enables a novel methodology for learni
ng STSGs without tedious annotations. We evaluate our method on three video data
sets: OpenPVSG, 20BN, and MUGEN. Our approach demonstrates substantial improveme
nts over fully-supervised baselines, achieving a unary predicate prediction accu
racy of 27.78% (+12.65%) and a binary recall@5 of 0.42 (+0.22) on OpenPVSG. Add
itionally, LASER exceeds baselines by 7% on 20BN and 5.2% on MUGEN in terms of o
verall predicate prediction accuracy.
****
Yun Zhu,Jia-Chen Gu,Caitlin Sikora,Ho Ko,Yinxiao Liu,Chu-Cheng Lin,Lei Shu,Liang
chen Luo,Lei Meng,Bang Liu,Jindong Chen
Accelerating Inference of Retrieval-Augmented Generation via Sparse Context Sele
ction
Large language models (LLMs) augmented with retrieval exhibit robust performance
and extensive versatility by incorporating external contexts. However, the inpu
t length grows linearly in the number of retrieved documents, causing a dramatic
increase in latency. In this paper, we propose a novel paradigm named Sparse RA
G, which seeks to cut computation costs through sparsity. Speci■cally, Sparse RA
G encodes retrieved documents in parallel, which eliminates latency introduced b
y long-range attention of retrieved documents. Then, LLMs selectively decode the
output by only attending to highly relevant caches auto-regressively, which are
chosen via prompting LLMs with special control tokens. It is notable that Spars
e RAG combines the assessment of each individual document and the generation of
the response into a single process. The designed sparse mechanism in a RAG syste
m can facilitate the reduction of the number of documents loaded during decoding
for accelerating the inference of the RAG system. Additionally, ■ltering out un
desirable contexts enhances the model’s focus on relevant context, inherently im
proving its generation quality. Evaluation results on four datasets show that Sp
arse RAG can be used to strike an optimal balance between generation quality and
computational ef■ciency, demonstrating its generalizability across tasks.
****
Yi Zhang,Siwei Wang,Jiyuan Liu,Shengju Yu,Zhibin Dong,Suyuan Liu,Xinwang Liu,En
Zhu
DLEFT-MKC: Dynamic Late Fusion Multiple Kernel Clustering with Robust Tensor Lea
rning via Min-Max Optimization
Recent advancements in multiple kernel clustering (MKC) have highlighted the eff
ectiveness of late fusion strategies, particularly in enhancing computational ef
ficiency to near-linear complexity while achieving promising clustering performa
nce. However, existing methods encounter three significant limitations: (1) reli
ance on fixed base partition matrices that do not adaptively optimize during the
clustering process, thereby constraining their performance to the inherent repr
esentational capabilities of these matrices; (2) a focus on adjusting kernel wei
ghts to explore inter-view consistency and complementarity, which often neglects
the intrinsic high-order correlations among views, thereby limiting the extract
ion of comprehensive multiple kernel information; (3) a lack of adaptive mechani
sms to accommodate varying distributions within the data, which limits robustnes
s and generalization. To address these challenges, this paper proposes a novel a
lgorithm termed Dynamic Late Fusion Multiple Kernel Clustering with Robust {Tens
or Learning via min-max optimization (DLEFT-MKC), which effectively overcomes th
e representational bottleneck of base partition matrices and facilitates the lea
rning of meaningful high-order cross-view information. Specifically, it is the f
irst to incorporate a min-max optimization paradigm into tensor-based MKC, enhan
cing algorithm robustness and generalization. Additionally, it dynamically recon
structs decision layers to enhance representation capabilities and subsequently
stacks the reconstructed representations for tensor learning that promotes the c
apture of high-order associations and cluster structures across views, ultimatel
y yielding consensus clustering partitions. To solve the resultant optimization
problem, we innovatively design a strategy that combines reduced gradient descen
t with the alternating direction method of multipliers, ensuring convergence to
local optima while maintaining high computational efficiency. Extensive experime

ntal results across various benchmark datasets validate the superior effectivene
ss and efficiency of the proposed DLEFT-MKC.
****

Jiachen T. Wang,Prateek Mittal,Dawn Song,Ruoxi Jia
Data Shapley in One Training Run
Data Shapley offers a principled framework for attributing the contribution of d
ata within machine learning contexts. However, the traditional notion of Data Sh
apley requires re-training models on various data subsets, which becomes computa
tionally infeasible for large-scale models. Additionally, this retraining-based
definition cannot evaluate the contribution of data for a specific model trainin
g run, which may often be of interest in practice. This paper introduces a novel
 concept, In-Run Data Shapley, which eliminates the need for model retraining an
d is specifically designed for assessing data contribution for a particular mode
l of interest. In-Run Data Shapley calculates the Shapley value for each gradien
t update iteration and accumulates these values throughout the training process.
 We present several techniques that allow the efficient scaling of In-Run Data S
hapley to the size of foundation models. In its most optimized implementation, o
ur method adds negligible runtime overhead compared to standard model training.
This dramatic efficiency improvement makes it possible to perform data attributi
on for the foundation model pretraining stage. We present several case studies t
hat offer fresh insights into pretraining data's contribution and discuss their
implications for copyright in generative AI and pretraining data curation.
****

Yusuf H Roohani,Andrew H. Lee,Qian Huang,Jian Vora,Zachary Steinhart,Kexin Huang
,Alexander Marson,Percy Liang,Jure Leskovec
BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments
Agents based on large language models have shown great potential in accelerating
 scientific discovery by leveraging their rich background knowledge and reasonin
g capabilities. In this paper, we introduce BioDiscoveryAgent, an agent that des
igns new experiments, reasons about their outcomes, and efficiently navigates th
e hypothesis space to reach desired solutions. We demonstrate our agent on the p
roblem of designing genetic perturbation experiments, where the aim is to find a
 small subset out of many possible genes that, when perturbed, result in a speci
fic phenotype (e.g., cell growth). Utilizing its biological knowledge, BioDiscov
eryAgent can uniquely design new experiments without the need to train a machine
 learning model or explicitly design an acquisition function as in Bayesian opti
mization. Moreover, BioDiscoveryAgent using Claude 3.5 Sonnet achieves an averag
e of 21% improvement in predicting relevant genetic perturbations across six dat
asets, and a 46% improvement in the harder task of non-essential gene perturbati
on, compared to existing Bayesian optimization baselines specifically trained fo
r this task. Our evaluation includes one dataset that is unpublished, ensuring i
t is not part of the language model's training data. Additionally, BioDiscoveryA
gent predicts gene combinations to perturb more than twice as accurately as a ra
ndom baseline, a task so far not explored in the context of closed-loop experime
nt design. The agent also has access to tools for searching the biomedical liter
ature, executing code to analyze biological datasets, and prompting another agen
t to critically evaluate its predictions. Overall, BioDiscoveryAgent is interpre
table at every stage, representing an accessible new paradigm in the computation
al design of biological experiments with the potential to augment scientists' ef
ficacy.
****

Shiyuan Zhang,Weitong Zhang,Quanquan Gu
Energy-Weighted Flow Matching for Offline Reinforcement Learning
This paper investigates energy guidance in generative modeling, where the target
 distribution is defined as $q(\mathbf x) \propto p(\mathbf x)\exp(-\beta \mathc
al E(\mathbf x))$, with $p(\mathbf x)$ being the data distribution and $\mathcal
 E(\mathbf x)$ as the energy function. To comply with energy guidance, existing
methods often require auxiliary procedures to learn intermediate guidance during
 the diffusion process. To overcome this limitation, we explore energy-guided fl
ow matching, a generalized form of the diffusion process. We introduce energy-we

ighted flow matching (EFM), a method that directly learns the energy-guided flow without the need for auxiliary models. Theoretical analysis shows that energy-weighted flow matching accurately captures the guided flow. Additionally, we extend this methodology to energy-weighted diffusion models and apply it to offline reinforcement learning (RL) by proposing the Q-weighted Iterative Policy Optimization (QIPO). Empirically, we demonstrate that the proposed QIPO algorithm improves performance in offline RL tasks. Notably, our algorithm is the first energy-guided diffusion model that operates independently of auxiliary models and the first exact energy-guided flow matching model in the literature.
****

Mohammad Shahab Sepehri,Zalan Fabian,Maryam Soltanolkotabi,Mahdi Soltanolkotabi
MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models
Multimodal Large Language Models (MLLMs) have tremendous potential to improve the accuracy, availability, and cost-effectiveness of healthcare by providing automated solutions or serving as aids to medical professionals. Despite promising first steps in developing medical MLLMs in the past few years, their capabilities and limitations are not well understood. Recently, many benchmark datasets have been proposed that test the general medical knowledge of such models across a variety of medical areas. However, the systematic failure modes and vulnerabilities of such models are severely underexplored with most medical benchmarks failing to expose the shortcomings of existing models in this safety-critical domain. In this paper, we introduce MediConfusion, a challenging medical Visual Question Answering (VQA) benchmark dataset, that probes the failure modes of medical MLLMs from a vision perspective. We reveal that state-of-the-art models are easily confused by image pairs that are otherwise visually dissimilar and clearly distinct for medical experts. Strikingly, all available models (open-source or proprietary) achieve performance below random guessing on MediConfusion, raising serious concerns about the reliability of existing medical MLLMs for healthcare deployment. We also extract common patterns of model failure that may help the design of a new generation of more trustworthy and reliable MLLMs in healthcare.
****

Anming Gu,Edward Chien,Kristjan Greenewald
Partially Observed Trajectory Inference using Optimal Transport and a Dynamics Prior
Trajectory inference seeks to recover the temporal dynamics of a population from snapshots of its (uncoupled) temporal marginals, i.e. where observed particles are \emph{not} tracked over time. Prior works addressed this challenging problem under a stochastic differential equation (SDE) model with a gradient-driven drift in the observed space, introducing a minimum entropy estimator relative to the Wiener measure and a practical grid-free mean-field Langevin (MFL) algorithm using Schr\"odinger bridges. Motivated by the success of observable state space models in the traditional paired trajectory inference problem (e.g. target tracking), we extend the above framework to a class of latent SDEs in the form of \emph{observable state space models}.
In this setting, we use partial observations to infer trajectories in the latent space under a specified dynamics model (e.g. the constant velocity/acceleration models from target tracking). We introduce the PO-MFL algorithm to solve this latent trajectory inference problem and provide theoretical guarantees to the partially observed setting. Experiments validate the robustness of our method and the exponential convergence of the MFL dynamics, and demonstrate significant outperformance over the latent-free baseline in key scenarios.
****

Liangliang Shi,Zhengyan Shi,Junchi Yan
SelKD: Selective Knowledge Distillation via Optimal Transport Perspective
Knowledge Distillation (KD) has been a popular paradigm for training a (smaller) student model from its teacher model. However, little research has been done on the practical scenario where only a subset of the teacher's knowledge needs to be distilled, which we term selective KD (SelKD). This demand is especially pronounced in the era of foundation models, where the teacher model can be significa

ntly larger than the student model. To address this issue, we propose to rethink the knowledge distillation problem from the perspective of Inverse Optimal Transport (IOT). Previous Bayesian frameworks mapped each sample to the probabilities of corresponding labels in an end-to-end manner, which fixed the number of classification categories and hindered effective partial knowledge transfer. In contrast, IOT calculates from the standpoint of transportation or matching, allowing for the flexible selection of samples and their quantities for matching. Traditional logit-based KD can be viewed as a special case within the IOT framework. Building on this IOT foundation, we formalize this setting in the context of classification, where only selected categories from the teacher's category space are required to be recognized by the student in the context of closed-set recognition, which we call closed-set SelKD, enhancing the student's performance on specific subtasks. Furthermore, we extend the closed-set SelKD, introducing an open-set version of SelKD, where the student model is required to provide a "not selected" response for categories outside its assigned task. Experimental results on standard benchmarks demonstrate the superiority of our approach. The source code is available at: \href{https://github.com/machoshi/SelKD}
****

Xuanyu Zhang,Jiarui Meng,Zhipei Xu,Shuzhou Yang,Yanmin Wu,Ronggang Wang,Jian Zhang

SecureGS: Boosting the Security and Fidelity of 3D Gaussian Splatting Steganography

3D Gaussian Splatting (3DGS) has emerged as a premier method for 3D representation due to its real-time rendering and high-quality outputs, underscoring the critical need to protect the privacy of 3D assets. Traditional NeRF steganography methods fail to address the explicit nature of 3DGS since its point cloud files are publicly accessible. Existing GS steganography solutions mitigate some issues but still struggle with reduced rendering fidelity, increased computational demands, and security flaws, especially in the security of the geometric structure of the visualized point cloud. To address these demands, we propose a \textbf{SecureGS}, a secure and efficient 3DGS steganography framework inspired by Scaffold-GS's anchor point design and neural decoding. SecureGS uses a hybrid decoupled Gaussian encryption mechanism to embed offsets, scales, rotations, and RGB attributes of the hidden 3D Gaussian points in anchor point features, retrievable only by authorized users through privacy-preserving neural networks. To further enhance security, we propose a density region-aware anchor growing and pruning strategy that adaptively locates optimal hiding regions without exposing hidden information. Extensive experiments show that SecureGS significantly surpasses existing GS steganography methods in rendering fidelity, speed, and security.
****

Haitao Yang,Yuan Dong,Hanwen Jiang,Dejia Xu,Georgios Pavlakos,Qixing Huang

Atlas Gaussians Diffusion for 3D Generation

Using the latent diffusion model has proven effective in developing novel 3D generation techniques. To harness the latent diffusion model, a key challenge is designing a high-fidelity and efficient representation that links the latent space and the 3D space. In this paper, we introduce Atlas Gaussians, a novel representation for feed-forward native 3D generation. Atlas Gaussians represent a shape as the union of local patches, and each patch can decode 3D Gaussians. We parameterize a patch as a sequence of feature vectors and design a learnable function to decode 3D Gaussians from the feature vectors. In this process, we incorporate UV-based sampling, enabling the generation of a sufficiently large, and theoretically infinite, number of 3D Gaussian points. The large amount of 3D Gaussians enables the generation of high-quality details. Moreover, due to local awareness of the representation, the transformer-based decoding procedure operates on a patch level, ensuring efficiency. We train a variational autoencoder to learn the Atlas Gaussians representation, and then apply a latent diffusion model on its latent space for learning 3D Generation. Experiments show that our approach outperforms the prior arts of feed-forward native 3D generation. Project page: https://yanghtr.github.io/projects/atlas_gaussians.
****

Alexey Gorbatovski,Boris Shaposhnikov,Alexey Malakhov,Nikita Surnachev,Yaroslav Aksenov,Ian Maksimov,Nikita Balagansky,Daniil Gavrilov
Learn Your Reference Model for Real Good Alignment
Despite the fact that offline methods for Large Language Models (LLMs) alignment do not require a direct reward model, they remain susceptible to overoptimization. This issue arises when the trained model deviates excessively from the reference policy, leading to a decrease in sample quality. We propose a novel approach of offline alignment methods, called Trust Region (including variants TR-DPO, TR-IPO, TR-KTO), which dynamically updates the reference policy throughout the training process. Our results show that TR alignment methods effectively mitigate overoptimization, enabling models to maintain strong performance even when substantially deviating from the initial reference policy. We demonstrate the efficacy of these approaches not only through toy examples that exhibit reduced overoptimization, but also through direct, side-by-side comparisons in specific tasks such as helpful and harmless dialogue, as well as summarization, where they surpass conventional methods. Additionally, we report significant improvements in general-purpose assistant setups with the Llama3 model on the AlpacaEval 2 and Arena-Hard benchmarks, highlighting the advantages of Trust Region methods over classical approaches.
****
Lingjie Yi,Jiachen Yao,Weimin Lyu,Haibin Ling,Raphael Douady,Chao Chen
Geometry of Long-Tailed Representation Learning: Rebalancing Features for Skewed Distributions
Deep learning has achieved significant success by training on balanced datasets. However, real-world data often exhibit long-tailed distributions. Empirical studies have revealed that long-tailed data skew data representations, where head classes dominate the feature space. Many methods have been proposed to empirically rectify the skewed representations. However, a clear understanding of the underlying cause and extent of this skew remains lacking. In this study, we provide a comprehensive theoretical analysis to elucidate how long-tailed data affect feature distributions, deriving the conditions under which centers of tail classes shrink together or even collapse into a single point. This results in overlapping feature distributions of tail classes, making features in the overlapping regions inseparable. Moreover, we demonstrate that merely empirically correcting the skewed representations of the training data is insufficient to separate the overlapping features due to distribution shifts between the training and real data. To address these challenges, we propose a novel long-tailed representation learning method, FeatRecon. It reconstructs the feature space in order to arrange features from different classes into symmetricial and linearly separable regions. This, in turn, enhances the model’s robustness to long-tailed data. We validate the effectiveness of our method through extensive experiments on the CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 datasets.
****
Xiao FU,Xian Liu,Xintao Wang,Sida Peng,Menghan Xia,Xiaoyu Shi,Ziyang Yuan,Pengfei Wan,Di ZHANG,Dahua Lin
3DTrajMaster: Mastering 3D Trajectory for Multi-Entity Motion in Video Generation
This paper aims to manipulate multi-entity 3D motions in video generation. Previous methods on controllable video generation primarily leverage 2D control signals to manipulate object motions and have achieved remarkable synthesis results. However, 2D control signals are inherently limited in expressing the 3D nature of object motions. To overcome this problem, we introduce 3DTrajMaster, a robust controller that regulates multi-entity dynamics in 3D space, given user-desired 6DoF pose (location and rotation) sequences of entities. At the core of our approach is a plug-and-play 3D-motion grounded object injector that fuses multiple input entities with their respective 3D trajectories through a gated self-attention mechanism. In addition, we exploit an injector architecture to preserve the video diffusion prior, which is crucial for generalization ability. To mitigate video quality degradation, we introduce a domain adaptor during training and employ an annealed sampling strategy during inference. To address the lack of suitab

le training data, we construct a 360-Motion Dataset, which first correlates collected 3D human and animal assets with GPT-generated trajectory and then captures their motion with 12 evenly-surround cameras on diverse 3D UE platforms. Extensive experiments show that 3DTrajMaster sets a new state-of-the-art in both accuracy and generalization for controlling multi-entity 3D motions. Project page: http://fuxiao0719.github.io/projects/3dtrajmaster
****
Liu Ziyin,Yizhou Xu,Isaac L. Chuang
Remove Symmetries to Control Model Expressivity and Improve Optimization
When symmetry is present in the loss function, the model is likely to be trapped in a low-capacity state that is sometimes known as a ``collapse." Being trapped in these low-capacity states can be a major obstacle to training across many scenarios where deep learning technology is applied. We first prove two concrete mechanisms through which symmetries lead to reduced capacities and ignored features during training and inference. We then propose a simple and theoretically justified algorithm, \textit{syre}, to remove almost all symmetry-induced low-capacity states in neural networks. When this type of entrapment is especially a concern, removing symmetries with the proposed method is shown to correlate well with improved optimization or performance. A remarkable merit of the proposed method is that it is model-agnostic and does not require any knowledge of the symmetry.
****
Federico Barbero,Alex Vitvitskyi,Christos Perivolaropoulos,Razvan Pascanu,Petar Veli■kovi■
Round and Round We Go! What makes Rotary Positional Encodings useful?
Positional Encodings (PEs) are a critical component of Transformer-based Large Language Models (LLMs), providing the attention mechanism with important sequence-position information. One of the most popular types of encoding used today in LLMs are Rotary Positional Encodings (RoPE), that rotate the queries and keys based on their relative distance. A common belief is that RoPE is useful because it helps to decay token dependency as relative distance increases. In this work, we argue that this is unlikely to be the core reason. We study the internals of a trained Gemma 7B model to understand how RoPE is being used at a mechanical level. We find that Gemma learns to use RoPE to construct robust `positional' attention patterns by exploiting the highest frequencies. We also find that, in general, Gemma greatly prefers to use the lowest frequencies of RoPE, which we suspect are used to carry semantic information. We mathematically prove interesting behaviours of RoPE and conduct experiments to verify our findings, proposing a modification of RoPE that fixes some highlighted issues and improves performance. We believe that this work represents an interesting step in better understanding PEs in LLMs, which we believe holds crucial value for scaling LLMs to large sizes and context lengths.
****
Zui Chen,Tianqiao Liu,Tongqing,Mi Tian,Weiqi Luo,Zitao Liu
Advancing Mathematical Reasoning in Language Models: The Impact of Problem-Solving Data, Data Synthesis Methods, and Training Stages
Mathematical reasoning remains a challenging area for large language models (LLMs), prompting the development of math-specific LLMs such as LLEMMA, DeepSeekMath, and Qwen2-Math, among others. These models typically follow a two-stage training paradigm: pre-training with math-related corpora and post-training with problem datasets for supervised fine-tuning (SFT). Despite these efforts, the improvements in mathematical reasoning achieved through continued pre-training (CPT) are often less significant compared to those obtained via SFT. This study addresses this discrepancy by exploring alternative strategies during the pre-training phase, focusing on the use of problem-solving data over general mathematical corpora.
We investigate three primary research questions: (1) Can problem-solving data enhance the model's mathematical reasoning capabilities more effectively than general mathematical corpora during CPT? (2) Are synthetic data from the same source equally effective, and which synthesis methods are most efficient? (3) How do t

he capabilities developed from the same problem-solving data differ between the CPT and SFT stages, and what factors contribute to these differences?
Our findings indicate that problem-solving data significantly enhances the model's mathematical capabilities compared to general mathematical corpora. We also identify effective data synthesis methods, demonstrating that the tutorship amplification synthesis method achieves the best performance. Furthermore, while SFT facilitates instruction-following abilities, it underperforms compared to CPT with the same data, which can be partially attributed to its poor learning capacity for more challenging problem-solving data. These insights provide valuable guidance for optimizing the mathematical reasoning capabilities of LLMs, culminating in our development of a powerful mathematical base model called MathGPT-8B.
****

Yinuo Liao,Yuanshen Guan,Ruikang Xu,Jiacheng Li,Shida Sun,Zhiwei Xiong
Learning Gain Map for Inverse Tone Mapping
For a more compatible and consistent high dynamic range (HDR) viewing experience, a new image format with a double-layer structure has been developed recently, which incorporates an auxiliary Gain Map (GM) within a standard dynamic range (SDR) image for adaptive HDR display. This new format motivates us to introduce a new task termed Gain Map-based Inverse Tone Mapping (GM-ITM), which focuses on learning the corresponding GM of an SDR image instead of directly estimating its HDR counterpart, thereby enabling a more effective up-conversion by leveraging the advantages of GM. The main challenge in this task, however, is to accurately estimate regional intensity variation with the fluctuating peak value. To this end, we propose a dual-branch network named GMNet, consisting of a Local Contrast Restoration (LCR) branch and a Global Luminance Estimation (GLE) branch to capture pixel-wise and image-wise information for GM estimation. Moreover, to facilitate the future research of the GM-ITM task, we build both synthetic and real-world datasets for comprehensive evaluations: synthetic SDR-GM pairs are generated from existing HDR resources, and real-world SDR-GM pairs are captured by mobile devices. Extensive experiments on these datasets demonstrate the superiority of our proposed GMNet over existing HDR-related methods both quantitatively and qualitatively. The codes and datasets are available at https://github.com/qtlark/GMNet.
****

Yogesh Verma,Ayush Bharti,Vikas Garg
Robust Simulation-Based Inference under Missing Data via Neural Processes
Simulation-based inference (SBI) methods typically require fully observed data to infer parameters of models with intractable likelihood functions. However, datasets often contain missing values due to incomplete observations, data corruptions (common in astrophysics), or instrument limitations (e.g., in high-energy physics applications). In such scenarios, missing data must be imputed before applying any SBI method. We formalize the problem of missing data in SBI and demonstrate that naive imputation methods can introduce bias in the estimation of SBI posterior. We also introduce a novel amortized method that addresses this issue by jointly learning the imputation model and the inference network within a neural posterior estimation (NPE) framework. Extensive empirical results on SBI benchmarks show that our approach provides robust inference outcomes compared to standard baselines for varying levels of missing data. Moreover, we demonstrate the merits of our imputation model on two real-world bioactivity datasets (Adrenergic and Kinase assays). Code is available at https://github.com/Aalto-QuML/RISE.
****

Hossein Mirzaei,Mackenzie W Mathis
Adversarially Robust Out-of-Distribution Detection Using Lyapunov-Stabilized Embeddings
Despite significant advancements in out-of-distribution (OOD) detection, existing methods still struggle to maintain robustness against adversarial attacks, compromising their reliability in critical real-world applications. Previous studies have attempted to address this challenge by exposing detectors to auxiliary OOD datasets alongside adversarial training. However, the increased data complexity inherent in adversarial training, and the myriad of ways that OOD samples can

arise during testing, often prevent these approaches from establishing robust de
cision boundaries. To address these limitations, we propose AROS, a novel approa
ch leveraging neural ordinary differential equations (NODEs) with Lyapunov stabi
lity theorem in order to obtain robust embeddings for OOD detection. By incorpor
ating a tailored loss function, we apply Lyapunov stability theory to ensure tha
t both in-distribution (ID) and OOD data converge to stable equilibrium points w
ithin the dynamical system. This approach encourages any perturbed input to retu
rn to its stable equilibrium, thereby enhancing the model’s robustness against a
dversarial perturbations. To not use additional data, we generate fake OOD embed
dings by sampling from low-likelihood regions of the ID data feature space, appr
oximating the boundaries where OOD data are likely to reside. To then further en
hance robustness, we propose the use of an orthogonal binary layer following the
 stable feature space, which maximizes the separation between the equilibrium po
ints of ID and OOD samples. We validate our method through extensive experiments
 across several benchmarks, demonstrating superior performance, particularly und
er adversarial attacks. Notably, our approach improves robust detection performa
nce from 37.8% to 80.1% on CIFAR-10 vs. CIFAR-100 and from 29.0% to 67.0% on CIF
AR-100 vs. CIFAR-10. Code and pre-trained models are available at https://github
.com/AdaptiveMotorControlLab/AROS.
****
Haitian Jiang,Renjie Liu,Zengfeng Huang,Yichuan Wang,Xiao Yan,Zhenkun Cai,Minjie
 Wang,David Wipf
MuseGNN: Forming Scalable, Convergent GNN Layers that Minimize a Sampling-Based
Energy
Among the many variants of graph neural network (GNN) architectures capable of m
odeling data with cross-instance relations, an important subclass involves layer
s designed such that the forward pass iteratively reduces a graph-regularized en
ergy function of interest. In this way, node embeddings produced at the output l
ayer dually serve as both predictive features for solving downstream tasks (e.g.
, node classification) and energy function minimizers that inherit transparent,
exploitable inductive biases and interpretability. However, scaling GNN architec
tures constructed in this way remains challenging, in part because the convergen
ce of the forward pass may involve models with considerable depth. To tackle thi
s limitation, we propose a sampling-based energy function and scalable GNN layer
s that iteratively reduce it, guided by convergence guarantees in certain settin
gs. We also instantiate a full GNN architecture based on these designs, and the
model achieves competitive accuracy and scalability when applied to the largest
publicly-available node classification benchmark exceeding 1TB in size. Our sour
ce code is available at https://github.com/haitian-jiang/MuseGNN.
****
Yuanhao Ban,Ruochen Wang,Tianyi Zhou,Boqing Gong,Cho-Jui Hsieh,Minhao Cheng
The Crystal Ball Hypothesis in diffusion models: Anticipating object positions f
rom initial noise
Diffusion models have achieved remarkable success in text-to-image generation ta
sks, yet the influence of initial noise remains largely unexplored. In this stud
y, we identify specific regions within the initial noise image, termed trigger p
atches, that play a key role in inducing object generation in the resulting imag
es. Notably, these patches are **universal** and can be generalized across vario
us positions, seeds, and prompts. To be specific, extracting these patches from
one noise and injecting them into another noise leads to object generation in ta
rgeted areas. To identify the trigger patches even before the image has been gen
erated, just like consulting the crystal ball to foresee fate, we first create a
 dataset consisting of Gaussian noises labeled with bounding boxes corresponding
 to the objects appearing in the generated images and **train a detector that id
entifies these patches from the initial noise.** To explain the formation of the
se patches, we reveal that they are **outliers** in Gaussian noise, and follow d
istinct distributions through two-sample tests. These outliers can take effect w
hen injected into different noises and generalize well across different settings
. Finally, we find the misalignment between prompts and the trigger patch patter
ns can result in unsuccessful image generations. To overcome it, we propose a re

ject-sampling strategy to obtain optimal noise, aiming to improve prompt adheren ce and positional diversity in image generation.
****

Bowei Tian,Ziyao Wang,Shwai He,Wanghao Ye,Guoheng Sun,Yucong Dai,Yongkai Wu,Ang Li

Towards counterfactual fairness through auxiliary variables

The challenge of balancing fairness and predictive accuracy in machine learning models, especially when sensitive attributes such as race, gender, or age are co nsidered, has motivated substantial research in recent years. Counterfactual fai rness ensures that predictions remain consistent across counterfactual variation s of sensitive attributes, which is a crucial concept in addressing societal bia ses.

However, existing counterfactual fairness approaches usually overlook intrinsic information about sensitive features, limiting their ability to achieve fairness while simultaneously maintaining performance. To tackle this challenge, we intr oduce EXOgenous Causal reasoning (EXOC), a novel causal reasoning framework moti vated by exogenous variables. It leverages auxiliary variables to uncover intrin sic properties that give rise to sensitive attributes. Our framework explicitly defines an auxiliary node and a control node that contribute to counterfactual f airness and control the information flow within the model. Our evaluation, condu cted on synthetic and real-world datasets, validates EXOC's superiority, showing that it outperforms state-of-the-art approaches in achieving counterfactual fai rness without sacrificing accuracy. Our code is available at https://github.com/ CASE-Lab-UMD/counterfactual_fairness_2025.
****

Slava Elizarov,Ciara Rowles,Simon Donné

Geometry Image Diffusion: Fast and Data-Efficient Text-to-3D with Image-Based Su rface Representation

Generating high-quality 3D objects from textual descriptions remains a challengi ng problem due to high computational costs, the scarcity of 3D data, and the com plexity of 3D representations. We introduce Geometry Image Diffusion (GIMDiffusi on), a novel Text-to-3D model that utilizes geometry images to efficiently repre sent 3D shapes using 2D images, thereby avoiding the need for complex 3D-aware a rchitectures. By integrating a Collaborative Control mechanism, we exploit the r ich 2D priors of existing Text-to-Image models, such as Stable Diffusion, to ach ieve strong generalization despite limited 3D training data. This allows us to u se only high-quality training data while retaining compatibility with guidance t echniques such as IPAdapter. GIMDiffusion enables the generation of 3D assets at speeds comparable to current Text-to-Image models, without being restricted to manifold meshes during either training or inference. We simultaneously generate a UV unwrapping for the objects, consisting of semantically meaningful parts as well as internal structures, enhancing both usability and versatility.
****

Thomas Norrenbrock,Timo Kaiser,Sovan Biswas,Ramesh Manuvinakurike,Bodo Rosenhahn

QPM: Discrete Optimization for Globally Interpretable Image Classification

Understanding the classifications of deep neural networks, e.g. used in safety-c ritical situations, is becoming increasingly important. While recent models can locally explain a single decision, to provide a faithful global explanation abou t an accurate model’s general behavior is a more challenging open task. Towards that goal, we introduce the Quadratic Programming Enhanced Model (QPM), which le arns globally interpretable class representations. QPM represents every class wi th a binary assignment of very few, typically 5, features, that are also assigne d to other classes, ensuring easily comparable contrastive class representations . This compact binary assignment is found using discrete optimization based on p redefined similarity measures and interpretability constraints. The resulting op timal assignment is used to fine-tune the diverse features, so that each of them becomes the shared general concept between the assigned classes. Extensive eval uations show that QPM delivers unprecedented global interpretability across smal l and large-scale datasets while setting the state of the art for the accuracy o f interpretable models.

****
Alexander Meulemans,Seijin Kobayashi,Johannes Von Oswald,Nino Scherrer,Eric Elmo znino,Blake Aaron Richards,Guillaume Lajoie,Blaise Aguera y Arcas,Joao Sacrament o
Multi-agent cooperation through learning-aware policy gradients
Self-interested individuals often fail to cooperate, posing a fundamental challe nge for multi-agent learning. How can we achieve cooperation among self-interest ed, independent learning agents? Promising recent work has shown that in certain tasks cooperation can be established between ``learning-aware" agents who model the learning dynamics of each other. Here, we present the first unbiased, highe r-derivative-free policy gradient algorithm for learning-aware reinforcement lea rning, which takes into account that other agents are themselves learning throug h trial and error based on multiple noisy trials. We then leverage efficient seq uence models to condition behavior on long observation histories that contain tr aces of the learning dynamics of other agents. Training long-context policies wi th our algorithm leads to cooperative behavior and high returns on standard soci al dilemmas, including a challenging environment where temporally-extended actio n coordination is required. Finally, we derive from the iterated prisoner's dile mma a novel explanation for how and when cooperation arises among self-intereste d learning-aware agents.
****
Shufan Shen,Zhaobo Qi,Junshu Sun,Qingming Huang,Qi Tian,Shuhui Wang
Enhancing Pre-trained Representation Classifiability can Boost its Interpretabil ity
The visual representation of a pre-trained model prioritizes the classifiability on downstream tasks, while the widespread applications for pre-trained visual m odels have posed new requirements for representation interpretability. However, it remains unclear whether the pre-trained representations can achieve high inte rpretability and classifiability simultaneously. To answer this question, we qua ntify the representation interpretability by leveraging its correlation with the ratio of interpretable semantics within the representations. Given the pre-trai ned representations, only the interpretable semantics can be captured by interpr etations, whereas the uninterpretable part leads to information loss. Based on t his fact, we propose the Inherent Interpretability Score (IIS) that evaluates th e information loss, measures the ratio of interpretable semantics, and quantifie s the representation interpretability. In the evaluation of the representation i nterpretability with different classifiability, we surprisingly discover that th e interpretability and classifiability are positively correlated, i.e., represen tations with higher classifiability provide more interpretable semantics that ca n be captured in the interpretations. This observation further supports two bene fits to the pre-trained representations. First, the classifiability of represent ations can be further improved by fine-tuning with interpretability maximization . Second, with the classifiability improvement for the representations, we obtai n predictions based on their interpretations with less accuracy degradation. The discovered positive correlation and corresponding applications show that practi tioners can unify the improvements in interpretability and classifiability for p re-trained vision models. Codes are available at https://github.com/ssfgunner/II S.
****
Wenxuan Zhang,Philip Torr,Mohamed Elhoseiny,Adel Bibi
Bi-Factorial Preference Optimization: Balancing Safety-Helpfulness in Language M odels
Fine-tuning large language models (LLMs) on human preferences, typically throug h reinforcement learning from human feedback (RLHF), has proven successful in en hancing their capabilities. However, ensuring the safety of LLMs during fine-tu ning remains a critical concern, and mitigating the potential conflicts in safe ty and helpfulness is costly in RLHF. To address this issue, we propose a supe rvised learning framework called Bi-Factorial Preference Optimization (BFPO), wh ich re-parameterizes a joint RLHF objective of both safety and helpfulness into a single supervised learning objective. In the supervised optimization, a labeli

ng function is used to capture global preferences ranking to balance both safety and helpfulness. To evaluate BFPO, we develop a benchmark including comprehensive discriminative and generative tasks for helpfulness and harmlessness. The results indicate that our method significantly outperforms existing approaches in both safety and helpfulness. Moreover, BFPO eliminates the need for human prompting and annotation in LLM fine-tuning while achieving the same level of safety as methods that heavily rely on human labor, with less than 10\% of the computational resources. The training recipes and models will be released.
****

Zachary Novack,Ge Zhu,Jonah Casebeer,Julian McAuley,Taylor Berg-Kirkpatrick,Nicholas J. Bryan
Presto! Distilling Steps and Layers for Accelerating Music Generation
Despite advances in diffusion-based text-to-music (TTM) methods, efficient, high-quality generation remains a challenge. We introduce Presto!, an approach to inference acceleration for score-based diffusion transformers via reducing both sampling steps and cost per step. To reduce steps, we develop a new score-based distribution matching distillation (DMD) method for the EDM-family of diffusion models, the first GAN-based distillation method for TTM. To reduce the cost per step, we develop a simple, but powerful improvement to a recent layer distillation method that improves learning via better preserving hidden state variance. Finally, we combine our step and layer distillation methods together for a dual-faceted approach. We evaluate our step and layer distillation methods independently and show each yield best-in-class performance. Our combined distillation method can generate high-quality outputs with improved diversity, accelerating our base model by 10-18x (230/435ms latency for 32 second mono/stereo 44.1kHz, 15x faster than the comparable SOTA model) — the fastest TTM to our knowledge.
****

Zhuoqun Li,Xuanang Chen,Haiyang Yu,Hongyu Lin,Yaojie Lu,Qiaoyu Tang,Fei Huang,Xianpei Han,Le Sun,Yongbin Li
StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization
Retrieval-augmented generation (RAG) is a key means to effectively enhance large language models (LLMs) in many knowledge-based tasks.
However, existing RAG methods struggle with knowledge-intensive reasoning tasks, because useful information required to these tasks are badly scattered.
This characteristic makes it difficult for existing RAG methods to accurately identify key information and perform global reasoning with such noisy augmentation.
In this paper, motivated by the cognitive theories that humans convert raw information into various structured knowledge when tackling knowledge-intensive reasoning, we proposes a new framework, StructRAG, which can identify the optimal structure type for the task at hand, reconstruct original documents into this structured format, and infer answers based on the resulting structure.
Extensive experiments across various knowledge-intensive tasks show that StructRAG achieves state-of-the-art performance, particularly excelling in challenging scenarios, demonstrating its potential as an effective solution for enhancing LLMs in complex real-world applications.
****

Tal Herman,Guy N. Rothblum
How to Verify Any (Reasonable) Distribution Property: Computationally Sound Argument Systems for Distributions
As statistical analyses become more central to science, industry and society, there is a growing need to ensure correctness of their results. Approximate correctness can be verified by replicating the entire analysis, but can we verify without replication? We focus on distribution testing problems: verifying that an unknown distribution is close to having a claimed property. Our main contribution is an interactive protocol between a verifier and an untrusted prover, which can be used to verify any distribution property that can be decided in polynomial time given a full and explicit description of the distribution. If the distribution is at statistical distance ε from having the property, then the

verifier rejects with high probability. This soundness property holds against any polynomial-time strategy that a cheating prover might follow, assuming the existence of collision-resistant hash functions (a standard assumption in cryptography). For distributions over a domain of size N, the protocol consists of 4 messages and the communication complexity and verifier runtime are roughly $\widetilde{O}\left(\sqrt{N} / \varepsilon^2 \right)$. The verifier's sample complexity is $\widetilde{O}\left(\sqrt{N} / \varepsilon^2 \right)$, and this is optimal up to $\text{polylog}(N)$ factors (for any protocol, regardless of its communication complexity). Even for simple properties, approximately deciding whether an unknown distribution has the property can require quasi-linear sample complexity and running time. For any such property, our protocol provides a quadratic speedup over replicating the analysis.
****

En Yu,Kangheng Lin,Liang Zhao,Yana Wei,Zining Zhu,Haoran Wei,Jianjian Sun,Zheng Ge,Xiangyu Zhang,Jingyu Wang,Wenbing Tao

Unhackable Temporal Reward for Scalable Video MLLMs

In the pursuit of superior video-processing MLLMs, we have encountered a perplexing paradox: the “anti-scaling law”, where more data and larger models lead to worse performance. This study unmasks the culprit: “temporal hacking”, a phenomenon where models shortcut by fixating on select frames, missing the full video narrative. In this work, we systematically establish a comprehensive theory of temporal hacking, defining it from a reinforcement learning perspective, introducing the Temporal Perplexity (TPL) score to assess this misalignment, and proposing the Unhackable Temporal Rewarding (UTR) framework to mitigate the temporal hacking. Both theoretically and empirically, TPL proves to be a reliable indicator of temporal modeling quality, correlating strongly with frame activation patterns. Extensive experiments reveal that UTR not only counters temporal hacking but significantly elevates video comprehension capabilities. This work not only advances video-AI systems but also illuminates the critical importance of aligning proxy rewards with true objectives in MLLM development.
****

Xianbiao Qi,Yelin He,Jiaquan Ye,Chun-Guang Li,Bojia Zi,Xili Dai,Qin Zou,Rong Xiao

Taming Transformer Without Using Learning Rate Warmup

Scaling Transformer to a large scale without using some technical tricks such as learning rate warump and an obviously lower learning rate, is an extremely challenging task, and is increasingly gaining more attention. In this paper, we provide a theoretical analysis for training Transformer and reveal a key problem behind the model crash phenomenon in the training, \ie, the spectral energy concentration of $W_q^{\top} W_k$ (where W_q and W_k are the projection matrices for query and key in Transformer), which is the reason for a malignant entropy collapse. To remedy this problem, motivated by Weyl's Inequality, we present a novel optimization strategy---making weight updating in successive steps smooth, that is, if the ratio $\frac{\sigma_{1}(\nabla W_t)}{\sigma_{1}(W_{t-1})}$ is larger than a threshold, where ∇W_t is the updating quantity in step t, we will automatically bound the learning rate to a weighted multiply of $\frac{\sigma_{1}(W_{t-1})}{\sigma_{1}(\nabla W_t)}$. Our optimization strategy is able to prevent the rapid spectral energy concentration to only a few directions, and thus is able to avoid the malignant entropy collapse that will trigger the model crash. We conduct extensive experiments using ViT, Swin-Transformer and GPT, showing that our optimization strategy can effectively and stably train these (Transformer) models without using learning rate warmup.
****

Samuel Audia,Soheil Feizi,Matthias Zwicker,Dinesh Manocha

How Learnable Grids Recover Fine Detail in Low Dimensions: A Neural Tangent Kernel Analysis of Multigrid Parametric Encodings

Neural networks that map between low dimensional spaces are ubiquitous in computer graphics and scientific computing; however, in their naive implementation, they are unable to learn high frequency information. We present a comprehensive analysis comparing the two most common techniques for mitigating

this spectral bias: Fourier feature encodings (FFE) and multigrid parametric encodings (MPE). FFEs are seen as the standard for low dimensional mappings, but MPEs often outperform them and learn representations with higher resolution and finer detail. FFE's roots in the Fourier transform, make it susceptible to aliasing if pushed too far, while MPEs, which use a learned grid structure, have no such limitation. To understand the difference in performance, we use the neural tangent kernel (NTK) to evaluate these encodings through the lens of an analogous kernel regression. By finding a lower bound on the smallest eigenvalue of the NTK, we prove that MPEs improve a network's performance through the structure of their grid and not their learnable embedding. This mechanism is fundamentally different from FFEs, which rely solely on their embedding space to improve performance. Results are empirically validated on a 2D image regression task using images taken from 100 synonym sets of ImageNet and 3D implicit surface regression on objects from the Stanford graphics dataset. Using peak signal-to-noise ratio (PSNR) and multiscale structural similarity (MS-SSIM) to evaluate how well fine details are learned, we show that the MPE increases the minimum eigenvalue by 8 orders of magnitude over the baseline and 2 orders of magnitude over the FFE. The increase in spectrum corresponds to a 15 dB (PSNR) / 0.65 (MS-SSIM) increase over baseline and a 12 dB (PSNR) / 0.33 (MS-SSIM) increa se over the
FFE.
****
Tue Minh Cao,Nhat Hoang-Xuan,Hieu Pham,Phi Le Nguyen,My T. Thai
NeurFlow: Interpreting Neural Networks through Neuron Groups and Functional Inte ractions
Understanding the inner workings of neural networks is essential for enhancing m odel performance and interpretability. Current research predominantly focuses on examining the connection between individual neurons and the model's final predi ctions, which suffers from challenges in interpreting the internal workings of t he model, particularly when neurons encode multiple unrelated features. In this paper, we propose a novel framework that transitions the focus from analyzing in dividual neurons to investigating groups of neurons, shifting the emphasis from neuron-output relationships to the functional interactions between neurons. Our automated framework, NeurFlow, first identifies core neurons and clusters them i nto groups based on shared functional relationships, enabling a more coherent an d interpretable view of the network’s internal processes. This approach facilita tes the construction of a hierarchical circuit representing neuron interactions across layers, thus improving interpretability while reducing computational cost s. Our extensive empirical studies validate the fidelity of our proposed NeurFlo w. Additionally, we showcase its utility in practical applications such as image debugging and automatic concept labeling, thereby highlighting its potential to advance the field of neural network explainability.
****
Zhiwei He,Zhaopeng Tu,Xing Wang,Xingyu Chen,Zhijie Wang,Jiahao Xu,Tian Liang,Wen xiang Jiao,Zhuosheng Zhang,Rui Wang
RaSA: Rank-Sharing Low-Rank Adaptation
Low-rank adaptation (LoRA) has been prominently employed for parameter-efficient fine-tuning of large language models (LLMs). However, the limited expressive ca pacity of LoRA, stemming from the low-rank constraint, has been recognized as a bottleneck, particularly in rigorous tasks like code generation and mathematical reasoning. To address this limitation, we introduce Rank-Sharing Low-Rank Adapt ation (RaSA), an innovative extension that enhances the expressive capacity of L oRA by leveraging partial rank sharing across layers. By forming a shared rank p ool and applying layer-specific weighting, RaSA effectively increases the number of ranks without augmenting parameter overhead. Our theoretically grounded and empirically validated approach demonstrates that RaSA not only maintains the cor e advantages of LoRA but also significantly boosts performance in challenging co de and math tasks. Code, data and scripts are available at: https://github.com/z whe99/RaSA.
****

Tongda Xu,Xiyan Cai,Xinjie Zhang,Xingtong Ge,Dailan He,Ming Sun,Jingjing Liu,Ya-Qin Zhang,Jian Li,Yan Wang
Rethinking Diffusion Posterior Sampling: From Conditional Score Estimator to Maximizing a Posterior
Recent advancements in diffusion models have been leveraged to address inverse problems without additional training, and Diffusion Posterior Sampling (DPS) (Chung et al., 2022a) is among the most popular approaches. Previous analyses suggest that DPS accomplishes posterior sampling by approximating the conditional score. While in this paper, we demonstrate that the conditional score approximation employed by DPS is not as effective as previously assumed, but rather aligns more closely with the principle of maximizing a posterior (MAP). This assertion is substantiated through an examination of DPS on 512$\times$512 ImageNet images, revealing that: 1) DPS’s conditional score estimation significantly diverges from the score of a well-trained conditional diffusion model and is even inferior to the unconditional score; 2) The mean of DPS’s conditional score estimation deviates significantly from zero, rendering it an invalid score estimation; 3) DPS generates high-quality samples with significantly lower diversity. In light of the above findings, we posit that DPS more closely resembles MAP than a conditional score estimator, and accordingly propose the following enhancements to DPS: 1) we explicitly maximize the posterior through multi-step gradient ascent and projection; 2) we utilize a light-weighted conditional score estimator trained with only 100 images and 8 GPU hours. Extensive experimental results indicate that these proposed improvements significantly enhance DPS's performance. The source code for these improvements is provided in https://github.com/tongdaxu/Rethinking-Diffusion-Posterior-Sampling-From-Conditional-Score-Estimator-to-Maximizing-a-Posterior.
****
Zhuoyi PENG,Yixuan Tang,Yi Yang
Adversarial Mixup Unlearning
Machine unlearning is a critical area of research aimed at safeguarding data privacy by enabling the removal of sensitive information from machine learning models. One unique challenge in this field is catastrophic unlearning, where erasing specific data from a well-trained model unintentionally removes essential knowledge, causing the model to deviate significantly from a retrained one. To address this, we introduce a novel approach that regularizes the unlearning process by utilizing synthesized mixup samples, which simulate the data susceptible to catastrophic effects. At the core of our approach is a generator-unlearner framework, MixUnlearn, where a generator adversarially produces challenging mixup examples, and the unlearner effectively forgets target information based on these synthesized data. Specifically, we first introduce a novel contrastive objective to train the generator in an adversarial direction: generating examples that prompt the unlearner to reveal information that should be forgotten, while losing essential knowledge. Then the unlearner, guided by two other contrastive loss terms, processes the synthesized and real data jointly to ensure accurate unlearning without losing critical knowledge, overcoming catastrophic effects. Extensive evaluations across benchmark datasets demonstrate that our method significantly outperforms state-of-the-art approaches, offering a robust solution to machine unlearning. This work not only deepens understanding of unlearning mechanisms but also lays the foundation for effective machine unlearning with mixup augmentation.
****
Haoze He,Juncheng B Li,Xuan Jiang,Heather Miller
SMT: Fine-Tuning Large Language Models with Sparse Matrices
Various parameter-efficient fine-tuning (PEFT) methods, including LoRA and its variants, have gained popularity for reducing computational costs. However, there is often an accuracy gap between PEFT approaches and full fine-tuning (FT), and this discrepancy has not yet been systematically explored. In this work, we introduce a method for selecting sparse sub-matrices that aims to minimize the performance gap between PEFT vs. full fine-tuning (FT) while also reducing both fine-tuning computational costs and memory costs. We explored both gradient-based and activation-based parameter selection methods to identify the most significant

sub-matrices for downstream tasks, updating only these blocks during fine-tuning. In our experiments, we demonstrated that SMT consistently surpasses other PEFT baselines (e.g., LoRA and DoRA) in fine-tuning popular large language models such as LLaMA across a broad spectrum of tasks, while reducing the GPU memory footprint by 67% compared to FT. We also examine how the performance of LoRA and DoRA tends to plateau and decline as the number of trainable parameters increases, in contrast, our SMT method does not suffer from such issues.
****
Hengshuo Chu,Xiang Deng,Qi Lv,Xiaoyang Chen,Yinchuan Li,Jianye HAO,Liqiang Nie
3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds
3D Affordance detection is a challenging problem with broad applications on various robotic tasks.
Existing methods typically formulate the detection paradigm as a label-based semantic segmentation task.
This paradigm relies on predefined labels and lacks the ability to comprehend complex natural language, resulting in limited generalization in open-world scene.
To address these limitations, we reformulate the traditional affordance detection paradigm into \textit{Instruction Reasoning Affordance Segmentation} (IRAS) task.
This task is designed to output a affordance mask region given a query reasoning text, which avoids fixed categories of input labels.
We accordingly propose the \textit{3D-AffordanceLLM} (3D-ADLLM), a framework designed for reasoning affordance detection in 3D open-scene.
Specifically, 3D-ADLLM introduces large language models (LLMs) to 3D affordance perception with a custom-designed decoder for generating affordance masks, thus achieving open-world reasoning affordance detection.
In addition, given the scarcity of 3D affordance datasets for training large models, we seek to extract knowledge from general segmentation data and transfer it to affordance detection.
Thus, we propose a multi-stage training strategy that begins with a novel pre-training task, i.e., \textit{Referring Object Part Segmentation}~(ROPS).
This stage is designed to equip the model with general recognition and segmentation capabilities at the object-part level.
Then followed by fine-tuning with the IRAS task, 3D-ADLLM obtains the reasoning ability for affordance detection.
In summary, 3D-ADLLM leverages the rich world knowledge and human-object interaction reasoning ability of LLMs, achieving approximately an 8\% improvement in mIoU on open-vocabulary affordance detection tasks.
****
Sheng Li,Qitao Tan,Yue Dai,Zhenglun Kong,Tianyu Wang,Jun Liu,Ao Li,Ninghao Liu,Yufei Ding,Xulong Tang,Geng Yuan
Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning
Self-supervised learning (SSL) offers a compelling solution to the challenge of extensive labeled data requirements in traditional supervised learning.
With the proven success of Vision Transformers (ViTs) in supervised tasks, there is increasing interest in adapting them for SSL frameworks. However, the high computational demands of SSL pose substantial challenges, particularly on resource-limited platforms like edge devices, despite its ability to achieve high accuracy without labeled data.
Recent studies in supervised learning have shown that token pruning can reduce training costs by removing less informative tokens without compromising accuracy. However, SSL’s dual-branch encoders make traditional single-branch pruning strategies less effective, as they fail to account for the critical cross-branch similarity information, leading to reduced accuracy in SSL.
To this end, we introduce SimPrune, a novel token pruning strategy designed for ViTs in SSL. SimPrune leverages cross-branch similarity information to efficiently prune tokens, retaining essential semantic information across dual branches.
Additionally, we incorporate a difficulty-aware pruning strategy to further enha

nce SimPrune's effectiveness.
Experimental results show that our proposed approach effectively reduces trainin
g computation while maintaining accuracy. Specifically, our approach offers 24\%
 savings in training costs compared to SSL baseline, without sacrificing accurac
y.
****
Chia-Hao Kao,Cheng Chien,Yu-Jen Tseng,Yi-Hsin Chen,Alessandro Gnutti,Shao-Yuan L
o,Wen-Hsiao Peng,Riccardo Leonardi
Bridging Compressed Image Latents and Multimodal Large Language Models
This paper presents the first-ever study of adapting compressed image latents to
 suit the needs of downstream vision tasks that adopt Multimodal Large Language
Models (MLLMs). MLLMs have extended the success of large language models to moda
lities (e.g. images) beyond text, but their billion scale hinders deployment on
resource-constrained end devices. While cloud-hosted MLLMs could be available, t
ransmitting raw, uncompressed images captured by end devices to the cloud requir
es an efficient image compression system. To address this, we focus on emerging
neural image compression and propose a novel framework with a lightweight transf
orm-neck and a surrogate loss to adapt compressed image latents for MLLM-based v
ision tasks.
Given the huge scale of MLLMs, our framework excludes the entire downstream MLLM
 except part of its visual encoder from training our system. This stands out fro
m most existing coding for machine approaches that involve downstream networks i
n training and thus could be impractical when the networks are MLLMs. The propos
ed framework is general in that it is applicable to various MLLMs, neural image
codecs, and multiple application scenarios, where the neural image codec can be
(1) pre-trained for human perception without updating, (2) fully updated for joi
nt human and machine perception, or (3) fully updated for only machine perceptio
n.
Extensive experiments on different neural image codecs and various MLLMs show th
at our method achieves great rate-accuracy performance with much less complexity
.
****
T. Konstantin Rusch,Daniela Rus
Oscillatory State-Space Models
We propose Linear Oscillatory State-Space models (LinOSS) for efficiently learni
ng on long sequences. Inspired by cortical dynamics of biological neural network
s, we base our proposed LinOSS model on a system of forced harmonic oscillators.
 A stable discretization, integrated over time using fast associative parallel s
cans, yields the proposed state-space model. We prove that LinOSS produces stabl
e dynamics only requiring nonnegative diagonal state matrix. This is in stark co
ntrast to many previous state-space models relying heavily on restrictive parame
terizations. Moreover, we rigorously show that LinOSS is universal, i.e., it can
 approximate any continuous and causal operator mapping between time-varying fun
ctions, to desired accuracy. In addition, we show that an implicit-explicit disc
retization of LinOSS perfectly conserves the symmetry of time reversibility of t
he underlying dynamics. Together, these properties enable efficient modeling of
long-range interactions, while ensuring stable and accurate long-horizon forecas
ting. Finally, our empirical results, spanning a wide range of time-series tasks
 from mid-range to very long-range classification and regression, as well as lon
g-horizon forecasting, demonstrate that our proposed LinOSS model consistently o
utperforms state-of-the-art sequence models. Notably, LinOSS outperforms Mamba a
nd LRU by nearly 2x on a sequence modeling task with sequences of length 50k.
****
Erle Zhu,Yadi Liu,Zhe Zhang,Xujun Li,JinZhou,Xinjie Yu,Minlie Huang,Hongning Wan
g
MAPS: Advancing Multi-Modal Reasoning in Expert-Level Physical Science
Pre-trained on extensive text and image corpora, current Multi-Modal Large Langu
age Models (MLLM) have shown strong capabilities in general visual reasoning tas
ks.
However, their performance is still lacking in physical domains that require und

erstanding diagrams with complex physical structures and quantitative analysis b
ased on multi-modal information.
To address this, we develop a new framework, named **M**ulti-Modal Scientific Re
Asoning with **P**hysics Perception and **S**imulation (**MAPS**) based on a
n MLLM.
MAPS decomposes expert-level multi-modal reasoning task into physical diagram un
derstanding via a Physical Perception Model (PPM) and reasoning with physical kn
owledge via a simulator.
The PPM module is obtained by fine-tuning a visual language model using carefull
y designed synthetic data with paired physical diagrams and corresponding simula
tion language descriptions.
At the inference stage, MAPS integrates the simulation language description of t
he input diagram provided by PPM and results obtained through a Chain-of-Simulat
ion process with MLLM to derive the underlying rationale and the final answer.
Validated using our collected college-level circuit analysis problems, MAPS sign
ificantly improves reasoning accuracy of MLLM and outperforms all existing model
s.
The results confirm MAPS offers a promising direction for enhancing multi-modal
scientific reasoning ability of MLLMs.
We will release our code, model and dataset used for our experiments upon publis
hing of this paper.
****
Kibum Kim,Kanghoon Yoon,Yeonjun In,Jaehyeong Jeon,Jinyoung Moon,Donghyun Kim,Cha
nyoung Park
Weakly Supervised Video Scene Graph Generation via Natural Language Supervision
Existing Video Scene Graph Generation (VidSGG) studies are trained in a fully su
pervised manner, which requires all frames in a video to be annotated, thereby i
ncurring high annotation cost compared to Image Scene Graph Generation (ImgSGG).
 Although the annotation cost of VidSGG can be alleviated by adopting a weakly s
upervised approach commonly used for ImgSGG (WS-ImgSGG) that uses image captions
, there are two key reasons that hinder such a naive adoption: 1) Temporality wi
thin video captions, i.e., unlike image captions, video captions include tempora
l markers (e.g., before, while, then, after) that indicate time-related details,
 and 2) Variability in action duration, i.e., unlike human actions in image capt
ions, human actions in video captions unfold over varying duration. To address t
hese issues, we propose a Natural Language-based Video Scene Graph Generation (N
L-VSGG) framework that only utilizes the readily available video captions for tr
aining a VidSGG model. NL-VSGG consists of two key modules: Temporality-aware Ca
ption Segmentation (TCS) module and Action Duration Variability-aware caption-fr
ame alignment (ADV) module. Specifically, TCS segments the video captions into m
ultiple sentences in a temporal order based on a Large Language Model (LLM), and
 ADV aligns each segmented sentence with appropriate frames considering the vari
ability in action duration. Our approach leads to a significant enhancement in p
erformance compared to simply applying the WS-ImgSGG pipeline to VidSGG on the A
ction Genome dataset. As a further benefit of utilizing the video captions as we
ak supervision, we show that the VidSGG model trained by NL-VSGG is able to pred
ict a broader range of action classes that are not included in the training data
, which makes our framework practical in reality.
****
Daniil Vankov,Anton Rodomanov,Angelia Nedich,Lalitha Sankar,Sebastian U Stich
Optimizing (L_0, L_1)-Smooth Functions by Gradient Methods
We study gradient methods for optimizing (L_0, L_1)-smooth functions, a
class that generalizes Lipschitz-smooth functions and has gained attention for
its relevance in machine learning.
We provide new insights into the structure of this function class and develop
a principled framework for analyzing optimization methods in this setting.
While our convergence rate estimates recover existing results for minimizing
the gradient norm in nonconvex problems, our approach significantly improves
the best-known complexity bounds for convex objectives.
Moreover, we show that the gradient method with Polyak stepsizes and the

normalized gradient method achieve nearly the same complexity guarantees as methods that rely on explicit knowledge of (L_0, L_1).
Finally, we demonstrate that a carefully designed accelerated gradient method can be applied to (L_0, L_1)-smooth functions, further improving all previous results.
****
Yossi Gandelsman,Alexei A Efros,Jacob Steinhardt
Interpreting the Second-Order Effects of Neurons in CLIP
We interpret the function of individual neurons in CLIP by automatically describing them using text. Analyzing the direct effects (i.e. the flow from a neuron through the residual stream to the output) or the indirect effects (overall contribution) fails to capture the neurons' function in CLIP. Therefore, we present the "second-order lens", analyzing the effect flowing from a neuron through the later attention heads, directly to the output. We find that these effects are highly selective: for each neuron, the effect is significant for <2% of the images. Moreover, each effect can be approximated by a single direction in the text-image space of CLIP. We describe neurons by decomposing these directions into sparse sets of text representations. The sets reveal polysemantic behavior - each neuron corresponds to multiple, often unrelated, concepts (e.g. ships and cars). Exploiting this neuron polysemy, we mass-produce "semantic" adversarial examples by generating images with concepts spuriously correlated to the incorrect class.
Additionally, we use the second-order effects for zero-shot segmentation, outperforming previous methods. Our results indicate that a automated interpretation of neurons can be used for model deception and for introducing new model capabilities
****
Mingjie Li,Wai Man Si,Michael Backes,Yang Zhang,Yisen Wang
SaLoRA: Safety-Alignment Preserved Low-Rank Adaptation
As advancements in large language models (LLMs) continue and the demand for personalized models increases, parameter-efficient fine-tuning (PEFT) methods (e.g., LoRA) become essential due to their efficiency in reducing computation costs.
However, recent studies have raised alarming concerns that LoRA fine-tuning could potentially compromise the safety alignment in LLMs, posing significant risks for the model owner.
In this paper, we first investigate the underlying mechanism by analyzing the changes in safety alignment related features before and after fine-tuning.
Then, we propose a fixed safety module calculated by safety data and a task-specific initialization for trainable parameters in low-rank adaptations, termed Safety-alignment preserved Low-Rank Adaptation (SaLoRA).
Unlike previous LoRA methods and their variants, SaLoRA enables targeted modifications to LLMs without disrupting their original alignments.
Our experiments show that SaLoRA outperforms various adapters-based approaches across various evaluation metrics in different fine-tuning tasks.
****
Christopher Fifty,Ronald Guenther Junkins,Dennis Duan,Aniketh Iyengar,Jerry Weihong Liu,Ehsan Amid,Sebastian Thrun,Christopher Re
Restructuring Vector Quantization with the Rotation Trick
Vector Quantized Variational AutoEncoders (VQ-VAEs) are designed to compress a continuous input to a discrete latent space and reconstruct it with minimal distortion.
They operate by maintaining a set of vectors---often referred to as the codebook---and quantizing each encoder output to the nearest vector in the codebook.
However, as vector quantization is non-differentiable, the gradient to the encoder flows _around_ the vector quantization layer rather than _through_ it in a straight-through approximation.
This approximation may be undesirable as all information from the vector quantization operation is lost.
In this work, we propose a way to propagate gradients through the vector quantization layer of VQ-VAEs.
We smoothly transform each encoder output into its corresponding codebook vector

via a rotation and rescaling linear transformation that is treated as a constant during backpropagation.
As a result, the relative magnitude and angle between encoder output and codebook vector becomes encoded into the gradient as it propagates through the vector quantization layer and back to the encoder.
Across 11 different VQ-VAE training paradigms, we find this restructuring improves reconstruction metrics, codebook utilization, and quantization error.
****
Jinbiao Chen,Zhiguang Cao,Jiahai Wang,Yaoxin Wu,Hanzhang Qin,Zizhen Zhang,Yue-Jiao Gong
Rethinking Neural Multi-Objective Combinatorial Optimization via Neat Weight Embedding
Recent decomposition-based neural multi-objective combinatorial optimization (MOCO) methods struggle to achieve desirable performance. Even equipped with complex learning techniques, they often suffer from significant optimality gaps in weight-specific subproblems. To address this challenge, we propose a neat weight embedding method to learn weight-specific representations, which captures weight-instance interaction for the subproblems and was overlooked by most current methods. We demonstrate the potentials of our method in two instantiations. First, we introduce a succinct addition model to learn weight-specific node embeddings, which surpassed most existing neural methods. Second, we design an enhanced conditional attention model to simultaneously learn the weight embedding and node embeddings, which yielded new state-of-the-art performance. Experimental results on classic MOCO problems verified the superiority of our method. Remarkably, our method also exhibits favorable generalization performance across problem sizes, even outperforming the neural method specialized for boosting size generalization.
****
Tenglong Liu,Jianxiong Li,Yinan Zheng,Haoyi Niu,Yixing Lan,Xin Xu,Xianyuan Zhan
Skill Expansion and Composition in Parameter Space
Humans excel at reusing prior knowledge to address new challenges and developing skills while solving problems. This paradigm becomes increasingly popular in the development of autonomous agents, as it develops systems that can self-evolve in response to new challenges like human beings. However, previous methods suffer from limited training efficiency when expanding new skills and fail to fully leverage prior knowledge to facilitate new task learning. We propose Parametric Skill Expansion and Composition (PSEC), a new framework designed to iteratively evolve the agents' capabilities and efficiently address new challenges by maintaining a manageable skill library. This library can progressively integrate skill primitives as plug-and-play Low-Rank Adaptation (LoRA) modules in parameter-efficient finetuning, facilitating efficient and flexible skill expansion. This structure also enables the direct skill compositions in parameter space by merging LoRA modules that encode different skills, leveraging shared information across skills to effectively program new skills. Based on this, we propose a context-aware modular to dynamically activate different skills to collaboratively handle new tasks. Empowering diverse applications including multi-objective composition, dynamics shift, and continual policy shift, the results on D4RL, DSRL benchmarks, and the DeepMind Control Suite show that PSEC exhibits superior capacity to leverage prior knowledge to efficiently tackle new challenges, as well as expand its skill libraries to evolve the capabilities. Project website: https://ltlhuuu.github.io/PSEC/.
****
Ganzhao Yuan
ADMM for Nonconvex Optimization under Minimal Continuity Assumption
This paper introduces a novel approach to solving multi-block nonconvex composite optimization problems through a proximal linearized Alternating Direction Method of Multipliers (ADMM). This method incorporates an Increasing Penalization and Decreasing Smoothing (IPDS) strategy. Distinguishing itself from existing ADMM-style algorithms, our approach (denoted IPDS-ADMM) imposes a less stringent condition, specifically requiring continuity in just one block of the objective fun

ction. IPDS-ADMM requires that the penalty increases and the smoothing parameter decreases, both at a controlled pace. When the associated linear operator is bijective, IPDS-ADMM uses an over-relaxation stepsize for faster convergence; however, when the linear operator is surjective, IPDS-ADMM uses an under-relaxation stepsize for global convergence. We devise a novel potential function to facilitate our convergence analysis and prove an oracle complexity $\mathcal{O}(\epsilon^{-3})$ to achieve an ϵ-approximate critical point. To the best of our knowledge, this is the first complexity result for using ADMM to solve this class of nonsmooth nonconvex problems. Finally, some experiments on the sparse PCA problem are conducted to demonstrate the effectiveness of our approach.
****
Haowei Lin,Shanda Li,Haotian Ye,Yiming Yang,Stefano Ermon,Yitao Liang,Jianzhu Ma
TFG-Flow: Training-free Guidance in Multimodal Generative Flow
Given an unconditional generative model and a predictor for a target property (e.g., a classifier), the goal of training-free guidance is to generate samples with desirable target properties without additional training. As a highly efficient technique for steering generative models toward flexible outcomes, training-free guidance has gained increasing attention in diffusion models. However, existing methods only handle data in continuous spaces, while many scientific applications involve both continuous and discrete data (referred to as multimodality). Another emerging trend is the growing use of the simple and general flow matching framework in building generative foundation models, where guided generation remains under-explored. To address this, we introduce TFG-Flow, a novel training-free guidance method for multimodal generative flow. TFG-Flow addresses the curse-of-dimensionality while maintaining the property of unbiased sampling in guiding discrete variables. We validate TFG-Flow on four molecular design tasks and show that TFG-Flow has great potential in drug design by generating molecules with desired properties.
****
Jinbin Bai,Tian Ye,Wei Chow,Enxin Song,Qing-Guo Chen,Xiangtai Li,Zhen Dong,Lei Zhu,Shuicheng YAN
Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis
We present Meissonic, which elevates non-autoregressive text-to-image Masked Image Modeling (MIM) to a level comparable with state-of-the-art diffusion models like SDXL. By incorporating a comprehensive suite of architectural innovations, advanced positional encoding strategies, and optimized sampling conditions, Meissonic substantially improves MIM's performance and efficiency. Additionally, we leverage high-quality training data, integrate micro-conditions informed by human preference scores, and employ feature compression layers to further enhance image fidelity and resolution. Our model not only matches but often exceeds the performance of existing methods in generating high-quality, high-resolution images. Extensive experiments validate Meissonic’s capabilities, demonstrating its potential as a new standard in text-to-image synthesis.
****
Jian Wu,Linyi Yang,Dongyuan Li,Yuliang Ji,Manabu Okumura,Yue Zhang
MMQA: Evaluating LLMs with Multi-Table Multi-Hop Complex Questions
While large language models (LLMs) have made strides in understanding tabular data, current tabular evaluation benchmarks, such as WikiTableQuestions and WikiSQL, are focus on single-table scenarios, which cannot necessarily reflect the complexity of real-world applications. To bridge this gap, we present a \textbf{M}ulti-table and
Multi-hop Question Answering (MMQA) dataset to assess LLMs' understanding and reasoning capabilities in handling multi-table tasks. The MMQA dataset demands that models perform multiple inferences by drawing evidence from various tables, which are designed to be connected with each other and require models to identify and utilize relationships such as foreign and primary keys. Then, we introduce a comprehensive evaluation framework that tailors to assess LLMs' capabilities in several aspects including Multi-Table Retrieval, Text-to-SQL Generation, Multi-Table QA, Primary Key Selection, and Foreign Key Selection.

Finally, we propose a novel multi-table retrieval method that achieves state-of-the-art (SOTA) performance on the MMQA dataset compared to several strong baselines.
Our experiment results reveal that, compared with human performance, both open-source and commercial LLMs leave significant performance room for improvements in multi-table understanding and reasoning tasks. We believe that the MMQA benchmark will enhance and facilitate LLMs' multi-table capabilities in real-world scenarios.
****
Emilien Biré,Anthony Kobanda,Ludovic Denoyer,Rémy Portelas
Efficient Active Imitation Learning with Random Network Distillation
Developing agents for complex and underspecified tasks, where no clear objective exists, remains challenging but offers many opportunities. This is especially true in video games, where simulated players (bots) need to play realistically, and there is no clear reward to evaluate them. While imitation learning has shown promise in such domains, these methods often fail when agents encounter out-of-distribution scenarios during deployment. Expanding the training dataset is a common solution, but it becomes impractical or costly when relying on human demonstrations. This article addresses active imitation learning, aiming to trigger expert intervention only when necessary, reducing the need for constant expert input along training. We introduce Random Network Distillation DAgger (RND-DAgger), a new active imitation learning method that limits expert querying by using a learned state-based out-of-distribution measure to trigger interventions. This approach avoids frequent expert-agent action comparisons, thus making the expert intervene only when it is useful. We evaluate RND-DAgger against traditional imitation learning and other active approaches in 3D video games (racing and third-person navigation) and in a robotic locomotion task and show that RND-DAgger surpasses previous methods by reducing expert queries.
https://sites.google.com/view/rnd-dagger
****
Yining Ye,Xin Cong,Shizuo Tian,Yujia Qin,Chong Liu,Yankai Lin,Zhiyuan Liu,Maosong Sun
Rational Decision-Making Agent with Learning Internal Utility Judgment
With remarkable advancements, large language models (LLMs) have attracted significant efforts to develop LLM-based agents capable of executing intricate multi-step decision-making tasks. Existing approaches predominantly build upon the external performance measure to guide the decision-making process but the reliance on the external performance measure as prior is problematic in real-world scenarios, where such prior may be unavailable, flawed, or even erroneous. For genuine autonomous decision-making for LLM-based agents, it is imperative to develop rationality from their posterior experiences to judge the utility of each decision independently. In this work, we propose RaDAgent (Rational Decision-Making Agent), which fosters the development of its rationality through an iterative framework involving Experience Exploration and Utility Learning. Within this framework, Elo-based Utility Learning is devised to assign Elo scores to individual decision steps to judge their utilities via pairwise comparisons. Consequently, these Elo scores guide the decision-making process to derive optimal outcomes. Experimental results on the Game of 24, WebShop, ToolBench and RestBench datasets demonstrate RaDAgent’s superiority over baselines, achieving about 7.8% improvement on average. Besides, RaDAgent also can reduce costs (ChatGPT API calls), highlighting its effectiveness and efficiency.
****
Jinyang Li,En Yu,Sijia Chen,Wenbing Tao
OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer
Open-vocabulary multiple object tracking aims to generalize trackers to unseen categories during training, enabling their application across a variety of real-world scenarios. However, the existing open-vocabulary tracker is constrained by its framework structure, isolated frame-level perception, and insufficient modal interactions, which hinder its performance in open-vocabulary classification and tracking. In this paper, we propose OVTR (End-to-End Open-Vocabulary Multiple

Object Tracking with TRansformer), the first end-to-end open-vocabulary tracker that models motion, appearance, and category simultaneously. To achieve stable classification and continuous tracking, we design the CIP (Category Information Propagation) strategy, which establishes multiple high-level category information priors for subsequent frames. Additionally, we introduce a dual-branch structure for generalization capability and deep multimodal interaction, and incorporate protective strategies in the decoder to enhance performance. Experimental results show that our method surpasses previous trackers on the open-vocabulary MOT benchmark while also achieving faster inference speeds and significantly reducing preprocessing requirements. Moreover, the experiment transferring the model to another dataset demonstrates its strong adaptability.
****
Woojin Cho,Minju Jo,Kookjin Lee,Noseong Park
Neural Functions for Learning Periodic Signal
As function approximators, deep neural networks have served as an effective tool to represent various signal types. Recent approaches utilize multi-layer perceptrons (MLPs) to learn a nonlinear mapping from a coordinate to its corresponding signal, facilitating the learning of continuous neural representations from discrete data points. Despite notable successes in learning diverse signal types, coordinate-based MLPs often face issues of overfitting and limited generalizability beyond the training region, resulting in subpar extrapolation performance. This study addresses scenarios where the underlying true signals exhibit periodic properties, either spatially or temporally. We propose a novel network architecture, which extracts periodic patterns from measurements and leverages this information to represent the signal, thereby enhancing generalization and improving extrapolation performance. We demonstrate the efficacy of the proposed method through comprehensive experiments, including the learning of the periodic solutions for differential equations, and time series imputation (interpolation) and forecasting (extrapolation) on real-world datasets.
****
Zhangheng LI,Keen You,Haotian Zhang,Di Feng,Harsh Agrawal,Xiujun Li,Mohana Prasad Sathya Moorthy,Jeffrey Nichols,Yinfei Yang,Zhe Gan
Ferret-UI 2: Mastering Universal User Interface Understanding Across Platforms
Building a generalist model for user interface (UI) understanding is challenging due to various foundational issues, such as platform diversity, resolution variation, and data limitation. In this paper, we introduce Ferret-UI 2, a multimodal large language model (MLLM) designed for universal UI understanding across a wide range of platforms, including iPhone, Android, iPad, Webpage, and AppleTV. Building on the foundation of Ferret-UI, Ferret-UI 2 introduces three key innovations: support for multiple platform types, high-resolution perception through adaptive scaling, and advanced task training data generation powered by GPT-4o with set-of-mark visual prompting. These advancements enable Ferret-UI 2 to perform complex, user-centered interactions, making it highly versatile and adaptable for the expanding diversity of platform ecosystems. Extensive empirical experiments on referring, grounding, user-centric advanced tasks (comprising 9 subtasks $\times$ 5 platforms), GUIDE next-action prediction dataset, and GUI-World multi-platform benchmark demonstrate that Ferret-UI 2 significantly outperforms Ferret-UI, and also shows strong cross-platform transfer capabilities.
****
Xiao Yu,Baolin Peng,Vineeth Vajipey,Hao Cheng,Michel Galley,Jianfeng Gao,Zhou Yu
ExACT: Teaching AI Agents to Explore with Reflective-MCTS and Exploratory Learning
Autonomous agents have demonstrated significant potential in automating complex multistep decision-making tasks. However, even state-of-the-art vision-language models (VLMs), such as GPT-4o, still fall short of human-level performance, particularly in intricate web environments and long-horizon planning tasks. To address these limitations, we introduce Reflective Monte Carlo Tree Search (R-MCTS), a novel test-time algorithm designed to enhance the ability of AI agents, e.g., powered by GPT-4o, to explore decision space on the fly.
R-MCTS extends traditional MCTS by 1) incorporating contrastive reflection, allo

wing agents to learn from past interactions and dynamically improve their search efficiency; and 2) using multi-agent debate to provide reliable state evaluation. Moreover, we improve the agent's performance by fine-tuning GPT-4o through self-learning, using R-MCTS generated tree traversals without any human-provided labels. On the challenging VisualWebArena benchmark, our GPT-4o-based R-MCTS agent achieves a 6% to 30% relative improvement across various tasks compared to the previous state-of-the-art. Additionally, we show that the knowledge gained from test-time search can be effectively transferred back to GPT-4o via fine-tuning. The fine-tuned GPT-4o matches 97\% of R-MCTS's performance while reducing compute usage by a factor of four at test time. Furthermore, qualitative results reveal that the fine-tuned GPT-4o model demonstrates the ability to explore the environment, evaluate a state, and backtrack to viable ones when it detects that the current state cannot lead to success. Moreover, our work demonstrates the compute scaling properties in both training - data collection with R-MCTS - and testing time. These results suggest a promising research direction to enhance VLMs' reasoning and planning capabilities for agentic applications via test-time search and self-learning.
****

Peiyao Lai,Oren Mangoubi
Singular Subspace Perturbation Bounds via Rectangular Random Matrix Diffusions
Given a matrix $A \in \mathbb{R}^{m\times d}$ with singular values $\sigma_1\geq \cdots \geq \sigma_d$, and a random matrix $G \in \mathbb{R}^{m\times d}$ with iid $N(0,T)$ entries for some $T>0$, we derive new bounds on the Frobenius distance between subspaces spanned by the top-k (right) singular vectors of A and $A+G$. This problem arises in numerous applications in statistics where a data matrix may be corrupted by Gaussian noise, and in the analysis of the Gaussian mechanism in differential privacy, where Gaussian noise is added to data to preserve private information. We show that, for matrices A where the gaps in the top-k singular values are roughly $\Omega(\sigma_k-\sigma_{k+1})$ the expected Frobenius distance between the subspaces is $\tilde{O}(\frac{\sqrt{d}}{\sigma_k-\sigma_{k+1}} \times \sqrt{T})$, improving on previous bounds by a factor of $\frac{\sqrt{m}}{\sqrt{d}}$. To obtain our bounds we view the perturbation to the singular vectors as a diffusion process-- the Dyson-Bessel process-- and use tools from stochastic calculus to track the evolution of the subspace spanned by the top-k singular vectors, which may be of independent interest.
****

Vivek Farias,Adam Daniel Jozefiak
Self-Normalized Resets for Plasticity in Continual Learning
Plasticity Loss is an increasingly important phenomenon that refers to the empirical observation that as a neural network is continually trained on a sequence of changing tasks, its ability to adapt to a new task diminishes over time. We introduce Self-Normalized Resets (SNR), a simple adaptive algorithm that mitigates plasticity loss by resetting a neuron’s weights when evidence suggests its firing rate has effectively dropped to zero. Across a battery of continual learning problems and network architectures, we demonstrate that SNR consistently attains superior performance compared to its competitor algorithms. We also demonstrate that SNR is robust to its sole hyperparameter, its rejection percentile threshold, while competitor algorithms show significant sensitivity. SNR’s threshold-based reset mechanism is motivated by a simple hypothesis test we derive. Seen through the lens of this hypothesis test, competing reset proposals yield suboptimal error rates in correctly detecting inactive neurons, potentially explaining our experimental observations. We also conduct a theoretical investigation of the optimization landscape for the problem of learning a single ReLU. We show that even when initialized adversarially, an idealized version of SNR learns the target ReLU, while regularization based approaches can fail to learn.
****

Zhuowei Li,Zihao Xu,Ligong Han,Yunhe Gao,Song Wen,Di Liu,Hao Wang,Dimitris N. Metaxas
Implicit In-context Learning
In-context Learning (ICL) empowers large language models (LLMs) to swiftly adapt

to unseen tasks at inference-time by prefixing a few demonstration examples before queries. Despite its versatility, ICL incurs substantial computational and memory overheads compared to zero-shot learning and is sensitive to the selection and order of demonstration examples. In this work, we introduce \textbf{Implicit In-context Learning} (I2CL), an innovative paradigm that reduces the inference cost of ICL to that of zero-shot learning with minimal information loss. I2CL operates by first generating a condensed vector representation, namely a context vector, extracted from the demonstration examples. It then conducts an inference-time intervention through injecting a linear combination of the context vector and query activations back into the model’s residual streams. Empirical evaluation on nine real-world tasks across three model architectures demonstrates that I2CL achieves few-shot level performance at zero-shot inference cost, and it exhibits robustness against variations in demonstration examples. Furthermore, I2CL facilitates a novel representation of ``task-ids'', enhancing task similarity detection and fostering effective transfer learning. We also perform a comprehensive analysis and ablation study on I2CL, offering deeper insights into its internal mechanisms. Code is available at https://github.com/LzVv123456/I2CL.
****

Antonis Antoniades,Albert Örwall,Kexun Zhang,Yuxi Xie,Anirudh Goyal,William Yang Wang

SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement

Software engineers operating in complex and dynamic environments must continuously adapt to evolving requirements, learn iteratively from experience, and reconsider their approaches based on new insights. However, current large language model (LLM)-based software agents often follow linear, sequential processes that prevent backtracking and exploration of alternative solutions, limiting their ability to rethink their strategies when initial approaches prove ineffective. To address these challenges, we propose SWE-Search, a multi-agent framework that integrates Monte Carlo Tree Search (MCTS) with a self-improvement mechanism to enhance software agents' performance on repository-level software tasks. SWE-Search extends traditional MCTS by incorporating a hybrid value function that leverages LLMs for both numerical value estimation and qualitative evaluation. This enables self-feedback loops where agents iteratively refine their strategies based on both quantitative numerical evaluations and qualitative natural language assessments of pursued trajectories. The framework includes a SWE-Agent for adaptive exploration, a Value Agent for iterative feedback, and a Discriminator Agent that facilitates multi-agent debate for collaborative decision-making. Applied to the SWE-bench benchmark, our approach demonstrates a 23% relative improvement in performance across five models compared to standard open-source agents without MCTS. Our analysis reveals how performance scales with increased inference-time compute through deeper search, providing a pathway to improve software agents without requiring larger models or additional training data. This highlights the potential of self-evaluation driven search techniques in complex software engineering environments.
****

Yunhao Luo,Yilun Du

Grounding Video Models to Actions through Goal Conditioned Exploration

Large video models, pretrained on massive quantities of amount of Internet video, provide a rich source of physical knowledge about the dynamics and motions of objects and tasks.

However, video models are not grounded in the embodiment of an agent, and do not describe how to actuate the world to reach the visual states depicted in a video.

To tackle this problem, current methods use a separate vision-based inverse dynamic model trained on embodiment-specific data to map image states to actions.

Gathering data to train such a model is often expensive and challenging, and this model is limited to visual settings similar to the ones in which data is available.

In this paper, we investigate how to directly ground video models to continuous

actions through self-exploration in the embodied environment -- using generated video states as visual goals for exploration.
We propose a framework that uses trajectory level action generation in combination with video guidance to
enable an agent to solve complex tasks without any external supervision, e.g., rewards, action labels, or segmentation masks.
We validate the proposed approach on 8 tasks in Libero, 6 tasks in MetaWorld, 4 tasks in Calvin, and 12 tasks in iThor Visual Navigation.
We show how our approach is on par with or even surpasses multiple behavior cloning baselines trained on expert demonstrations while without requiring any action annotations.
****
Jingwei Xu,Yikai Wang,Yiqun Zhao,Yanwei Fu,Shenghua Gao
3D StreetUnveiler with Semantic-aware 2DGS - a simple baseline
Unveiling an empty street from crowded observations captured by in-car cameras is crucial for autonomous driving. However, removing all temporarily static objects, such as stopped vehicles and standing pedestrians, presents a significant challenge. Unlike object-centric 3D inpainting, which relies on thorough observation in a small scene, street scene cases involve long trajectories that differ from previous 3D inpainting tasks. The camera-centric moving environment of captured videos further complicates the task due to the limited degree and time duration of object observation. To address these obstacles, we introduce StreetUnveiler to reconstruct an empty street. StreetUnveiler learns a 3D representation of the empty street from crowded observations. Our representation is based on the hard-label semantic 2D Gaussian Splatting (2DGS) for its scalability and ability to identify Gaussians to be removed. We inpaint rendered image after removing unwanted Gaussians to provide pseudo-labels and subsequently re-optimize the 2DGS. Given its temporal continuous movement, we divide the empty street scene into observed, partial-observed, and unobserved regions, which we propose to locate through a rendered alpha map. This decomposition helps us to minimize the regions that need to be inpainted. To enhance the temporal consistency of the inpainting, we introduce a novel time-reversal framework to inpaint frames in reverse order and use later frames as references for earlier frames to fully utilize the long-trajectory observations. Our experiments conducted on the street scene dataset successfully reconstructed a 3D representation of the empty street. The mesh representation of the empty street can be extracted for further applications.
****
Toshinori Kitamura,Tadashi Kozuno,Wataru Kumagai,Kenta Hoshino,Yohei Hosoe,Kazumi Kasaura,Masashi Hamaya,Paavo Parmas,Yutaka Matsuo
Near-Optimal Policy Identification in Robust Constrained Markov Decision Processes via Epigraph Form
Designing a safe policy for uncertain environments is crucial in real-world control systems. However, this challenge remains inadequately addressed within the Markov decision process (MDP) framework. This paper presents the first algorithm guaranteed to identify a near-optimal policy in a robust constrained MDP (RCMDP), where an optimal policy minimizes cumulative cost while satisfying constraints in the worst-case scenario across a set of environments. We first prove that the conventional policy gradient approach to the Lagrangian max-min formulation can become trapped in suboptimal solutions. This occurs when its inner minimization encounters a sum of conflicting gradients from the objective and constraint functions. To address this, we leverage the epigraph form of the RCMDP problem, which resolves the conflict by selecting a single gradient from either the objective or the constraints. Building on the epigraph form, we propose a bisection search algorithm with a policy gradient subroutine and prove that it identifies an ε-optimal policy in an RCMDP with $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ robust policy evaluations.
****
Shayne Longpre,Nikhil Singh,Manuel Cherep,Kushagra Tiwary,Joanna Materzynska,William Brannon,Robert Mahari,Naana Obeng-Marnu,Manan Dey,Mohammed Hamdy,Nayan Saxena,Ahmad Mustafa Anis,Emad A. Alghamdi,Vu Minh Chien,Da Yin,Kun Qian,Yizhi LI,Mi

nnie Liang,An Dinh,Shrestha Mohanty,Deividas Mataciunas,Tobin South,Jianguo Zhan
g,Ariel N. Lee,Campbell S. Lund,Christopher Klamm,Damien Sileo,Diganta Misra,Enr
ico Shippole,Kevin Klyman,Lester James Validad Miranda,Niklas Muennighoff,Seongh
yeon Ye,Seungone Kim,Vipul Gupta,Vivek Sharma,Xuhui Zhou,Caiming Xiong,Luis Vill
a,Stella Biderman,Alex Pentland,Sara Hooker,Jad Kabbara
Bridging the Data Provenance Gap Across Text, Speech, and Video
Progress in AI is driven largely by the scale and quality of training data. Desp
ite this, there is a deficit of empirical analysis examining the attributes of w
ell-established datasets beyond text. In this work we conduct the largest and fi
rst-of-its-kind longitudinal audit across modalities --- popular text, speech, a
nd video datasets --- from their detailed sourcing trends and use restrictions t
o their geographical and linguistic representation. Our manual analysis covers n
early 4000 public datasets between 1990-2024, spanning 608 languages, 798 source
s, 659 organizations, and 67 countries. We find that multimodal machine learning
 applications have overwhelmingly turned to web-crawled, synthetic, and social m
edia platforms, such as YouTube, for their training sets, eclipsing all other so
urces since 2019. Secondly, tracing the chain of dataset derivations we find tha
t while less than 33% of datasets are restrictively licensed, over 80% of the so
urce content in widely-used text, speech, and video datasets, carry non-commerci
al restrictions. Finally, counter to the rising number of languages and geograph
ies represented in public AI training datasets, our audit demonstrates measures
of relative geographical and multilingual representation have failed to signific
antly improve their coverage since 2013. We believe the breadth of our audit ena
bles us to empirically examine trends in data sourcing, restrictions, and Wester
n-centricity at an ecosystem-level, and that visibility into these questions are
 essential to progress in responsible AI. As a contribution to ongoing improveme
nts in dataset transparency and responsible use, we release our entire multimoda
l audit, allowing practitioners to trace data provenance across text, speech, an
d video.
****
Wujiang Xu,Qitian Wu,Zujie Liang,Jiaojiao Han,Xuying Ning,Yunxiao Shi,Wenfang Li
n,Yongfeng Zhang
SLMRec: Distilling Large Language Models into Small for Sequential Recommendatio
n
Sequential Recommendation (SR) task involves predicting the next item a user is
likely to interact with, given their past interactions.
The SR models examine the sequence of a user's actions to discern more complex b
ehavioral patterns and temporal dynamics.
Recent research demonstrates the great impact of LLMs on sequential recommendati
on systems, either viewing sequential recommendation as language modeling or ser
ving as the backbone for user representation. Although these methods deliver out
standing performance, there is scant evidence of the necessity of a large langua
ge model and how large the language model is needed, especially in the sequentia
l recommendation scene. Meanwhile, due to the huge size of LLMs, it is inefficie
nt and impractical to apply a LLM-based model in real-world platforms that often
 need to process billions of traffic logs daily. In this paper, we explore the i
nfluence of LLMs' depth by conducting extensive experiments on large-scale indus
try datasets. Surprisingly, our motivational experiments reveal that most interm
ediate layers of LLMs are redundant, indicating that pruning the remaining layer
s can still maintain strong performance.
Motivated by this insight, we empower small language models for SR, namely SLMRe
c, which adopt a simple yet effective knowledge distillation method. Moreover, S
LMRec is orthogonal to other post-training efficiency techniques, such as quanti
zation and pruning, so that they can be leveraged in combination. Comprehensive
experimental results illustrate that the proposed SLMRec model attains the best
performance using only 13\% of the parameters found in LLM-based recommendation
models while simultaneously achieving up to 6.6x and 8.0x speedups in training a
nd inference time costs, respectively. Besides, we provide a theoretical justifi
cation for why small language models can perform comparably to large language mo
dels in SR.

****
Paola Cascante-Bonilla,Yu Hou,Yang Trista Cao,Hal Daumé III,Rachel Rudinger
Natural Language Inference Improves Compositionality in Vision-Language Models
Compositional reasoning in Vision-Language Models (VLMs) remains challenging as these models often struggle to relate objects, attributes, and spatial relationships. Recent methods aim to address these limitations by relying on the semantics of the textual description, using Large Language Models (LLMs) to break them down into subsets of questions and answers. However, these methods primarily operate on the surface level, failing to incorporate deeper lexical understanding while introducing incorrect assumptions generated by the LLM. In response to these issues, we present Caption Expansion with Contradictions and Entailments (CECE), a principled approach that leverages Natural Language Inference (NLI) to generate entailments and contradictions from a given premise. CECE produces lexically diverse sentences while maintaining their core meaning. Through extensive experiments, we show that CECE enhances interpretability and reduces overreliance on biased or superficial features. By balancing CECE along the original premise, we achieve significant improvements over previous methods without requiring additional fine-tuning, producing state-of-the-art results on benchmarks that score agreement with human judgments for image-text alignment, and achieving an increase in performance on Winoground of $+19.2\%$ (group score) and $+12.9\%$ on EqBen (group score) over the best prior work (finetuned with targeted data).
****
Mingquan Feng,Yixin Huang,Yizhou Liu,Bofang Jiang,Junchi Yan
PhysPDE: Rethinking PDE Discovery and a Physical HYpothesis Selection Benchmark
Despite extensive research, recovering PDE expressions from experimental observations often involves symbolic regression. This method generally lacks the incorporation of meaningful physical insights, resulting in outcomes lacking clear physical interpretations. Recognizing that the primary interest of Machine Learning for Science (ML4Sci) often lies in understanding the underlying physical mechanisms or even discovering new physical laws rather than simply obtaining mathematical expressions, this paper introduces a novel ML4Sci task paradigm. This paradigm focuses on interpreting experimental data within the framework of prior physical hypotheses and theories, thereby guiding and constraining the discovery of PDE expressions. We have formulated this approach as a nonlinear mixed-integer programming (MIP) problem, addressed through an efficient search scheme developed for this purpose. Our experiments on newly designed Fluid Mechanics and Laser Fusion datasets demonstrate the interpretability and feasibility of this method.
****
Harel Mendelman,Ronen Talmon
Supervised and Semi-Supervised Diffusion Maps with Label-Driven Diffusion
In this paper, we introduce Supervised Diffusion Maps (SDM) and Semi-Supervised Diffusion Maps (SSDM), which transform the well-known unsupervised dimensionality reduction algorithm, Diffusion Maps, into supervised and semi-supervised learning tools. The proposed methods, SDM and SSDM, are based on our new approach that treats the labels as a second view of the data. This unique framework allows us to incorporate ideas from multi-view learning. Specifically, we propose constructing two affinity kernels corresponding to the data and the labels. We then propose a multiplicative interpolation scheme of the two kernels, whose purpose is twofold. First, our scheme extracts the common structure underlying the data and the labels by defining a diffusion process driven by the data and the labels. This label-driven diffusion produces an embedding that emphasizes the properties relevant to the label-related task. Second, the proposed interpolation scheme balances the influence of the two kernels. We show on multiple benchmark datasets that the embedding learned by SDM and SSDM is more effective in downstream regression and classification tasks than existing unsupervised, supervised, and semi-supervised nonlinear dimension reduction methods.
****
Chenyu Wang,Masatoshi Uehara,Yichun He,Amy Wang,Avantika Lal,Tommi Jaakkola,Sergey Levine,Aviv Regev,Hanchen,Tommaso Biancalani
Fine-Tuning Discrete Diffusion Models via Reward Optimization with Applications

to DNA and Protein Design
Recent studies have demonstrated the strong empirical performance of diffusion m
odels on discrete sequences (i.e., discrete diffusion models) across domains suc
h as natural language and biological sequence generation. For example, in the pr
otein inverse folding task, where the goal is to generate a protein sequence fro
m a given backbone structure, conditional diffusion models have achieved impress
ive results in generating "natural" sequences that fold back into the original s
tructure. However, practical design tasks often require not only modeling a cond
itional distribution but also optimizing specific task objectives. For instance,
 in the inverse folding task, we may prefer proteins with high stability. To add
ress this, we consider the scenario where we have pre-trained discrete diffusion
 models that can generate "natural" sequences, as well as reward models that map
 sequences to task objectives. We then formulate the reward maximization problem
 within discrete diffusion models, analogous to reinforcement learning (RL), whi
le minimizing the KL divergence against pre-trained diffusion models to preserve
 naturalness. To solve this RL problem, we propose a novel algorithm that enable
s direct backpropagation of rewards through entire trajectories generated by dif
fusion models, by making the originally non-differentiable trajectories differen
tiable using the Gumbel-Softmax trick. Our theoretical analysis indicates that o
ur approach can generate sequences that are both "natural" (i.e., have a high pr
obability under a pre-trained model) and yield high rewards. While similar tasks
 have been recently explored in diffusion models for continuous domains, our wor
k addresses unique algorithmic and theoretical challenges specific to discrete d
iffusion models, which arise from their foundation in continuous-time Markov cha
ins rather than Brownian motion. Finally, we demonstrate the effectiveness of ou
r algorithm in generating DNA and protein sequences that optimize enhancer activ
ity and protein stability, respectively, important tasks for gene therapies and
protein-based therapeutics. The code is available at https://github.com/ChenyuWa
ng-Monica/DRAKES.
****
Zijian Li,Shunxing Fan,Yujia Zheng,Ignavier Ng,Shaoan Xie,Guangyi Chen,Xinshuai
Dong,Ruichu Cai,Kun Zhang
Synergy Between Sufficient Changes and Sparse Mixing Procedure for Disentangled
Representation Learning
Disentangled representation learning aims to uncover the latent variables underl
ying observed data, yet identifying these variables under mild assumptions remai
ns challenging. Some methods rely on sufficient changes in the distribution of l
atent variables indicated by auxiliary variables, such as domain indices, but ac
quiring enough domains is often impractical. Alternative approaches exploit the
structural sparsity assumption on mixing processes, but this constraint may not
hold in practice. Interestingly, we find that these two seemingly unrelated assu
mptions can actually complement each other. Specifically, when conditioned on au
xiliary variables, the sparse mixing process induces independence between latent
 and observed variables, which simplifies the mapping from estimated to true lat
ent variables and hence compensates for deficiencies of auxiliary variables. Bui
lding on this insight, we propose an identifiability theory with less restrictiv
e constraints regarding the auxiliary variables and the sparse mixing process, e
nhancing applicability to real-world scenarios. Additionally, we develop a gener
ative model framework incorporating a domain encoding network and a sparse mixin
g constraint and provide two implementations based on variational autoencoders a
nd generative adversarial networks. Experiment results on synthetic and real-wor
ld datasets support our theoretical results.
****
Weihan Xu,Paul Pu Liang,Haven Kim,Julian McAuley,Taylor Berg-Kirkpatrick,Hao-Wen
 Dong
TeaserGen: Generating Teasers for Long Documentaries
Teasers are an effective tool for promoting content in entertainment, commercial
 and educational fields. However, creating an effective teaser for long videos i
s challenging for it requires long-range multimodal modeling capability for the
input videos, while necessitating maintaining audiovisual alignments, managing s

cene transitions and preserving factual accuracy for the output teasers. Due to the lack of a publicly-available dataset, progress along this research direction has been hindered. In this work, we present DocumentaryNet, a collection of 1,269 documentaries paired with their teasers, featuring multimodal data streams of video, speech, music, sound effects and narrations. With DocumentaryNet, we propose a new two-stage system for generating teasers from long documentaries. The proposed TeaserGen system first generates the teaser narration from the transcribed narration from the documentary using a pretrained large language model, and then selects the most relevant visual content to accompany the generated narration through language-vision models. For narration-video matching, we explore two approaches: a pretraining-based model using pretrained contrastive language-vision models and a deep sequential model that learns the mapping between the narrations and visuals. Our experimental results show that the pretraining-based approach is more effective at identifying relevant visual content than directly trained deep autoregressive models.
****
Yixuan Zhu,Haolin Wang,Ao Li,Wenliang Zhao,Yansong Tang,Jingxuan Niu,Lei Chen,Jie Zhou,Jiwen Lu
InstaRevive: One-Step Image Enhancement via Dynamic Score Matching
Image enhancement finds wide-ranging applications in real-world scenarios due to complex environments and the inherent limitations of imaging devices. Recent diffusion-based methods yield promising outcomes but necessitate prolonged and computationally intensive iterative sampling. In response, we propose InstaRevive, a straightforward yet powerful image enhancement framework that employs score-based diffusion distillation to harness potent generative capability and minimize the sampling steps. To fully exploit the potential of the pre-trained diffusion model, we devise a practical and effective diffusion distillation pipeline using dynamic noise control to address inaccuracies in updating direction during score matching. Our noise control strategy enables a dynamic diffusing scope, facilitating precise learning of denoising trajectories within the diffusion model and ensuring accurate distribution matching gradients during training. Additionally, to enrich guidance for the generative power, we incorporate textual prompts via image captioning as auxiliary conditions, fostering further exploration of the diffusion model. Extensive experiments substantiate the efficacy of our framework across a diverse array of challenging tasks and datasets, unveiling the compelling efficacy and efficiency of InstaRevive in delivering high-quality and visually appealing results.
****
Sijun Tan,Siyuan Zhuang,Kyle Montgomery,William Yuan Tang,Alejandro Cuadron,Chenguang Wang,Raluca Popa,Ion Stoica
JudgeBench: A Benchmark for Evaluating LLM-Based Judges
LLM-based judges have emerged as a scalable alternative to human evaluation and are increasingly used to assess, compare, and improve models. However, the reliability of LLM-based judges themselves is rarely scrutinized. As LLMs become more advanced, their responses grow more sophisticated, requiring stronger judges to evaluate them. Existing benchmarks primarily focus on a judge’s alignment with human preferences, but often fail to account for more challenging tasks where crowdsourced human preference is a poor indicator of factual and logical correctness. To address this, we propose a novel evaluation framework to objectively evaluate LLM-based judges. Based on this framework, we propose JudgeBench, a benchmark for evaluating LLM-based judges on challenging response pairs spanning knowledge, reasoning, math, and coding. JudgeBench leverages a novel pipeline for converting existing difficult datasets into challenging response pairs with preference labels reflecting objective correctness. Our comprehensive evaluation on a collection of prompted judges, fine-tuned judges, multi-agent judges, and reward models shows that JudgeBench poses a significantly greater challenge than previous benchmarks, with many strong models (e.g. GPT-4o) performing just slightly better than random guessing. Overall, JudgeBench offers a reliable platform for assessing increasingly advanced LLM-based judges. Data and code are available at \url{https://github.com/ScalerLab/JudgeBench}.

****
Cheol Jun Cho,Nicholas Lee,Akshat Gupta,Dhruv Agarwal,Ethan Chen,Alan Black,Gopala Anumanchipalli
Sylber: Syllabic Embedding Representation of Speech from Raw Audio
Syllables are compositional units of spoken language that efficiently structure human speech perception and production. However, current neural speech representations lack such structure, resulting in dense token sequences that are costly to process. To bridge this gap, we propose a new model, Sylber, that produces speech representations with clean and robust syllabic structure. Specifically, we propose a self-supervised learning (SSL) framework that bootstraps syllabic embeddings by distilling from its own initial unsupervised syllabic segmentation. This results in a highly structured representation of speech features, offering three key benefits: 1) a fast, linear-time syllable segmentation algorithm, 2) efficient syllabic tokenization with an average of 4.27 tokens per second, and 3) novel phonological units suited for efficient spoken language modeling. Our proposed segmentation method is highly robust and generalizes to out-of-domain data and unseen languages without any tuning. By training token-to-speech generative models, fully intelligible speech can be reconstructed from Sylber tokens with a significantly lower bitrate than baseline SSL tokens. This suggests that our model effectively compresses speech into a compact sequence of tokens with minimal information loss. Lastly, we demonstrate that categorical perception—a linguistic phenomenon in speech perception—emerges naturally in Sylber, making the embedding space more categorical and sparse than previous speech features and thus supporting the high efficiency of our tokenization. Together, we present a novel SSL approach for representing speech as syllables, with significant potential for efficient speech tokenization and spoken language modeling.
****
Zeyuan Allen-Zhu,Yuanzhi Li
Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws
Scaling laws describe the relationship between the size of language models and their capabilities. Unlike prior studies that evaluate a model's capability via loss or benchmarks, we estimate information-theoretically the number of knowledge \emph{bits} a model stores. We focus on factual knowledge represented as tuples, such as (USA, capital, Washington D.C.) from a Wikipedia page. Through multiple controlled datasets, we establish that language models can and only can store \emph{2 bits of knowledge per parameter, even when quantized to int8}, and such knowledge can be flexibly extracted for downstream applications.

More broadly, we present 12 results on how (1) training duration, (2) model architecture, (3) quantization, (4) sparsity constraints such as MoE, and (5) data signal-to-noise ratio affect a model's knowledge storage capacity.
****
Heyang Zhao,Xingrui Yu,David Mark Bossens,Ivor Tsang,Quanquan Gu
Beyond-Expert Performance with Limited Demonstrations: Efficient Imitation Learning with Double Exploration
Imitation learning is a central problem in reinforcement learning where the goal is to learn a policy that mimics the expert's behavior. In practice, it is often challenging to learn the expert policy from a limited number of demonstrations accurately due to the complexity of the state space. Moreover, it is essential to explore the environment and collect data to achieve beyond-expert performance. To overcome these challenges, we propose a novel imitation learning algorithm called Imitation Learning with Double Exploration (ILDE), which implements exploration in two aspects: (1) optimistic policy optimization via an exploration bonus that rewards state-action pairs with high uncertainty to potentially improve the convergence to the expert policy, and (2) curiosity-driven exploration of the states that deviate from the demonstration trajectories to potentially yield beyond-expert performance. Empirically, we demonstrate that ILDE outperforms the state-of-the-art imitation learning algorithms in terms of sample efficiency and achieves beyond-expert performance on Atari and MuJoCo tasks with fewer demonstrations than in previous work. We also provide a theoretical justification of IL

DE as an uncertainty-regularized policy optimization method with optimistic expl
oration, leading to a regret growing sublinearly in the number of episodes.
****

Haque Ishfaq,Guangyuan Wang,Sami Nur Islam,Doina Precup
Langevin Soft Actor-Critic: Efficient Exploration through Uncertainty-Driven Cri
tic Learning
Existing actor-critic algorithms, which are popular for continuous control reinf
orcement learning (RL) tasks, suffer from poor sample efficiency due to lack of
principled exploration mechanism within them. Motivated by the success of Thomps
on sampling for efficient exploration in RL, we propose a novel model-free RL al
gorithm, \emph{Langevin Soft Actor Critic} (LSAC), which prioritizes enhancing c
ritic learning through uncertainty estimation over policy optimization. LSAC emp
loys three key innovations: approximate Thompson sampling through distributional
 Langevin Monte Carlo (LMC) based Q updates, parallel tempering for exploring
multiple modes of the posterior of the Q function, and diffusion synthesized s
tate-action samples regularized with Q action gradients. Our extensive experim
ents demonstrate that LSAC outperforms or matches the performance of mainstream
model-free RL algorithms for continuous control tasks.
Notably, LSAC marks the first successful application of an LMC based Thompson sa
mpling in continuous control tasks with continuous action spaces.
****

Jianqi Chen,Panwen Hu,Xiaojun Chang,Zhenwei Shi,Michael Kampffmeyer,Xiaodan Lian
g
Sitcom-Crafter: A Plot-Driven Human Motion Generation System in 3D Scenes
Recent advancements in human motion synthesis have focused on specific types of
motions, such as human-scene interaction, locomotion or human-human interaction,
 however, there is a lack of a unified system capable of generating a diverse co
mbination of motion types. In response, we introduce *Sitcom-Crafter*, a compreh
ensive and extendable system for human motion generation in 3D space, which can
be guided by extensive plot contexts to enhance workflow efficiency for anime an
d game designers. The system is comprised of eight modules, three of which are d
edicated to motion generation, while the remaining five are augmentation modules
 that ensure consistent fusion of motion sequences and system functionality. Cen
tral to the generation modules is our novel 3D scene-aware human-human interacti
on module, which addresses collision issues by synthesizing implicit 3D Signed D
istance Function (SDF) points around motion spaces, thereby minimizing human-sce
ne collisions without additional data collection costs. Complementing this, our
locomotion and human-scene interaction modules leverage existing methods to enri
ch the system's motion generation capabilities. Augmentation modules encompass p
lot comprehension for command generation, motion synchronization for seamless in
tegration of different motion types, hand pose retrieval to enhance motion reali
sm, motion collision revision to prevent human collisions, and 3D retargeting to
 ensure visual fidelity. Experimental evaluations validate the system's ability
to generate high-quality, diverse, and physically realistic motions, underscorin
g its potential for advancing creative workflows. Code and demonstration videos
can be found in the supplementary files.
****

Honghui Wang,Shiji Song,Gao Huang
GridMix: Exploring Spatial Modulation for Neural Fields in PDE Modeling
Significant advancements have been achieved in PDE modeling using neural fields.
 Despite their effectiveness, existing methods rely on global modulation, limiti
ng their ability to reconstruct local details. While spatial modulation with van
illa grid-based representations offers a promising alternative, it struggles wit
h inadequate global information modeling and over-fitting to the training spatia
l domain. To address these challenges, we propose GridMix, a novel approach that
 models spatial modulation as a mixture of grid-based representations. GridMix e
ffectively explores global structures while preserving locality for fine-grained
 modulation. Furthermore, we introduce spatial domain augmentation to enhance th
e robustness of the modulated neural fields against spatial domain variations.
With all these innovations,

our comprehensive approach culminates in MARBLE, a framework that significantly advancing the capabilities of neural fields in PDE modeling. The effectiveness o f MARBLE is extensively
validated on diverse benchmarks encompassing dynamics modeling and geometric pre diction.
****
Yufan Chen,Ruiping Liu,Junwei Zheng,Di Wen,Kunyu Peng,Jiaming Zhang,Rainer Stief elhagen
Graph-based Document Structure Analysis
When reading a document, glancing at the spatial layout of a document is an init ial step to understand it roughly. Traditional document layout analysis (DLA) me thods, however, offer only a superficial parsing of documents, focusing on basic instance detection and often failing to capture the nuanced spatial and logical relationships between instances. These limitations hinder DLA-based models from achieving a gradually deeper comprehension akin to human reading. In this work, we propose a novel graph-based Document Structure Analysis (gDSA) task. This ta sk requires that model not only detects document elements but also generates spa tial and logical relations in form of a graph structure, allowing to understand documents in a holistic and intuitive manner. For this new task, we construct a relation graph-based document structure analysis dataset(GraphDoc) with 80K docu ment images and 4.13M relation annotations, enabling training models to complete multiple tasks like reading order, hierarchical structures analysis, and comple x inter-element relationship inference. Furthermore, a document relation graph g enerator (DRGG) is proposed to address the gDSA task, which achieves performance with 57.6% at mAP_g@0.5 for a strong benchmark baseline on this novel task and dataset. We hope this graphical representation of document structure can mar k an innovative advancement in document structure analysis and understanding. Th e new dataset and code will be made publicly available.
****
Ishika Agarwal,Krishnateja Killamsetty,Lucian Popa,Marina Danilevsky
DELIFT: Data Efficient Language model Instruction Fine-Tuning
Fine-tuning large language models (LLMs) is crucial for task specialization but often becomes resource-intensive due to redundant or uninformative data. Existin g data selection methods typically rely either on computationally expensive grad ient-based metrics or static embeddings that fail to adapt dynamically to the mo del’s evolving state, thus limiting their practical effectiveness. To address th is,
we propose DELIFT (Data Efficient Language model Instruction Fine-Tuning), lever aging a novel, computationally efficient utility metric inspired by In-Context L earning (ICL). Our ICL-based metric measures the informational value of each dat a sample by quantifying its effectiveness as an in-context example in improving model predictions for other samples, reflecting its actual contribution relative to the model’s current state. Integrated with tailored submodular optimization methods, DELIFT systematically selects diverse, informative subsets optimized sp ecifically for each fine-tuning stage: instruction tuning, task-specific adaptat ion, and continual fine-tuning. Experimental results across multiple datasets an d model scales show DELIFT reduces fine-tuning data requirements by up to 70% wi thout compromising performance, consistently outperforming existing methods by u p to 26% in effectiveness and efficiency.
****
Sichen Zhu,Yuchen Zhu,Molei Tao,Peng Qiu
Diffusion Generative Modeling for Spatially Resolved Gene Expression Inference f rom Histology Images
Spatial Transcriptomics (ST) allows a high-resolution measurement of RNA sequenc e abundance by systematically connecting cell morphology depicted in Hematoxylin and eosin (H\&E) stained histology images to spatially resolved gene expression s. ST is a time-consuming, expensive yet powerful experimental technique that pr ovides new opportunities to understand cancer mechanisms at a fine-grained molec ular level, which is critical for uncovering new approaches for disease diagnosi s and treatments. Here, we present $\textbf{Stem}$ ($\underline{\textbf{S}}$pa$\

underline{\textbf{T}}$ially resolved gene $\underline{\textbf{E}}$xpression infe
rence with diffusion $\underline{\textbf{M}}$odel), a novel computational tool t
hat leverages a conditional diffusion generative model to enable in silico gene
expression inference from H&E stained images. Through better capturing the inher
ent stochasticity and heterogeneity in ST data, $\textbf{Stem}$ achieves state-o
f-the-art performance on spatial gene expression prediction and generates biolog
ically meaningful gene profiles for new H&E stained images at test time. We eval
uate the proposed algorithm on datasets with various tissue sources and sequenci
ng platforms, where it demonstrates clear improvement over existing approaches.
$\textbf{Stem}$ generates high-fidelity gene expression predictions that share s
imilar gene variation levels as ground truth data, suggesting that our method pr
eserves the underlying biological heterogeneity. Our proposed pipeline opens up
the possibility of analyzing existing, easily accessible H&E stained histology i
mages from a genomics point of view without physically performing gene expressio
n profiling and empowers potential biological discovery from H&E stained histolo
gy images. Code is available at: https://github.com/SichenZhu/Stem.
****

Wei Wang,Dong-Dong Wu,Jindong Wang,Gang Niu,Min-Ling Zhang,Masashi Sugiyama
Realistic Evaluation of Deep Partial-Label Learning Algorithms
Partial-label learning (PLL) is a weakly supervised learning problem in which
each example is associated with multiple candidate labels and only one is the
true label. In recent years, many deep PLL algorithms have been developed to
improve model performance. However, we find that some early developed
algorithms are often underestimated and can outperform many later algorithms
with complicated designs. In this paper, we delve into the empirical
perspective of PLL and identify several critical but previously overlooked
issues. First, model selection for PLL is non-trivial, but has never been
systematically studied. Second, the experimental settings are highly
inconsistent, making it difficult to evaluate the effectiveness of the
algorithms. Third, there is a lack of real-world image datasets that can be
compatible with modern network architectures. Based on these findings, we
propose PLENCH, the first Partial-Label learning bENCHmark to systematically
compare state-of-the-art deep PLL algorithms. We investigate the model
selection problem for PLL for the first time, and propose novel model selection
criteria with theoretical guarantees. We also create Partial-Label CIFAR-10
(PLCIFAR10), an image dataset of human-annotated partial labels collected from
Amazon Mechanical Turk, to provide a testbed for evaluating the performance of
PLL algorithms in more realistic scenarios. Researchers can quickly and
conveniently perform a comprehensive and fair evaluation and verify the
effectiveness of newly developed algorithms based on PLENCH. We hope that
PLENCH will facilitate standardized, fair, and practical evaluation of PLL
algorithms in the future.
****

Richard Zhuang,Tianhao Wu,Zhaojin Wen,Andrew Li,Jiantao Jiao,Kannan Ramchandran
EmbedLLM: Learning Compact Representations of Large Language Models
With hundreds of thousands of language models available on Huggingface today, ef
ficiently evaluating and utilizing these models across various downstream tasks
has become increasingly critical. Many existing methods repeatedly learn task-sp
ecific representations of Large Language Models (LLMs), which leads to inefficie
ncies in both time and computational resources. To address this, we propose Embe
dLLM, a framework designed to learn compact vector representations of LLMs that
facilitate downstream applications involving many models, such as model routing.
We introduce an encoder-decoder approach for learning such embedding, along wit
h a systematic framework to evaluate their effectiveness. Empirical results show
that EmbedLLM outperforms prior methods in model routing. Additionally, we demo
nstrate that our method can forecast a model's performance on multiple benchmark
s, without incurring additional inference cost. Extensive probing experiments va
lidate that the learned embeddings capture key model characteristics, e.g. wheth
er the model is specialized for coding tasks, even without being explicitly trai
ned on them. We open source our dataset, code and embedder to facilitate further

research and application.
****
Tvrtko Tadi■,Cassiano O Becker,Jennifer Neville
Node Similarities under Random Projections: Limits and Pathological Cases
Random Projections have been widely used to generate embeddings for various graph learning tasks due to their computational efficiency. The majority of applications have been justified through the Johnson-Lindenstrauss Lemma. In this paper, we take a step further and investigate how well dot product and cosine similarity are preserved by random projections when these are applied over the rows of the graph matrix. Our analysis provides new asymptotic and finite-sample results, identifies pathological cases, and tests them with numerical experiments. We specialize our fundamental results to a ranking application by computing the probability of random projections flipping the node ordering induced by their embeddings. We find that, depending on the degree distribution, the method produces especially unreliable embeddings for the dot product, regardless of whether the adjacency or the normalized transition matrix is used. With respect to the statistical noise introduced by random projections, we show that cosine similarity produces remarkably more precise approximations.
****
Zhaofeng Wu,Xinyan Velocity Yu,Dani Yogatama,Jiasen Lu,Yoon Kim
The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities
Modern language models can process inputs across diverse languages and modalities. We hypothesize that models acquire this capability through learning a _shared representation space_ across heterogeneous data types (e.g., different languages and modalities), which places semantically similar inputs near one another, even if they are from different modalities/languages. We term this the _semantic hub hypothesis_, following the hub-and-spoke model from neuroscience (Patterson et al., 2007) which posits that semantic knowledge in the human brain is organized through a transmodal semantic "hub" which integrates information from various modality-specific ``spokes'' regions. We first show that model representations for semantically equivalent inputs in different languages are similar in the intermediate layers, and that this space can be interpreted using the model's dominant pretraining language via the logit lens. This tendency extends to other data types, including arithmetic expressions, code, and visual/audio inputs. Interventions in the shared representation space in one data type also predictably affect model outputs in other data types, suggesting that this shared representations space is not simply a vestigial byproduct of large-scale training on broad data, but something that is actively utilized by the model during input processing.
****
Aaron Jiaxun Li,Satyapriya Krishna,Himabindu Lakkaraju
More RLHF, More Trust? On The Impact of Preference Alignment On Trustworthiness
The trustworthiness of Large Language Models (LLMs) refers to the extent to which their outputs are reliable, safe, and ethically aligned, and it has become a crucial consideration alongside their cognitive performance. In practice, Reinforcement Learning From Human Feedback (RLHF) has been widely used to align LLMs with labeled human preferences, but its assumed effect on model trustworthiness hasn't been rigorously evaluated. To bridge this knowledge gap, this study investigates how models aligned with general-purpose preference data perform across five trustworthiness verticals: toxicity, stereotypical bias, machine ethics, truthfulness, and privacy. Our results demonstrate that RLHF on human preferences doesn't automatically guarantee trustworthiness, and reverse effects are often observed. Furthermore, we propose to adapt efficient influence function based data attribution methods to the RLHF setting to better understand the influence of fine-tuning data on individual trustworthiness benchmarks, and show its feasibility by providing our estimated attribution scores. Together, our results underscore the need for more nuanced approaches for model alignment from both the data and framework perspectives, and we hope this research will guide the community towards developing language models that are increasingly capable without sacrificing trustworthiness.

****
Zhong Zheng,Haochen Zhang,Lingzhou Xue
Federated Q-Learning with Reference-Advantage Decomposition: Almost Optimal Regret and Logarithmic Communication Cost
In this paper, we consider model-free federated reinforcement learning for tabular episodic Markov decision processes. Under the coordination of a central server, multiple agents collaboratively explore the environment and learn an optimal policy without sharing their raw data. Despite recent advances in federated Q-learning algorithms achieving near-linear regret speedup with low communication cost, existing algorithms only attain suboptimal regrets compared to the information bound. We propose a novel model-free federated Q-Learning algorithm, termed FedQ-Advantage. Our algorithm leverages reference-advantage decomposition for variance reduction and adopts three novel designs: separate event-triggered communication and policy switching, heterogeneous communication triggering conditions, and optional forced synchronization. We prove that our algorithm not only requires a lower logarithmic communication cost but also achieves an almost optimal regret, reaching the information bound up to a logarithmic factor and near-linear regret speedup compared to its single-agent counterpart when the time horizon is sufficiently large.
****
Xinyi Shang,Peng Sun,Tao Lin
GIFT: Unlocking Full Potential of Labels in Distilled Dataset at Near-zero Cost
Recent advancements in dataset distillation have demonstrated the significant benefits of employing soft labels generated by pre-trained teacher models.

In this paper, we introduce a novel perspective by emphasizing the full utilization of labels.

We first conduct a comprehensive comparison of various loss functions for soft label utilization in dataset distillation, revealing that the model trained on the synthetic dataset exhibits high sensitivity to the choice of loss function for soft label utilization.

This finding highlights the necessity of a universal loss function for training models on synthetic datasets.

Building on these insights, we introduce an extremely simple yet surprisingly effective plug-and-play approach, GIFT, which encompasses soft label refinement and a cosine similarity-based loss function to efficiently leverage full label information.

Extensive experiments indicate that GIFT consistently enhances state-of-the-art dataset distillation methods across various dataset scales without incurring additional computational costs.

Importantly, GIFT significantly enhances cross-optimizer generalization, an area previously overlooked.

For instance, on ImageNet-1K with IPC = 10, GIFT enhances the state-of-the-art method RDED by 30.8% in cross-optimizer generalization. Our code is available at https://github.com/LINs-lab/GIFT.
****
Prasanna Mayilvahanan,Roland S. Zimmermann,Thaddäus Wiedemer,Evgenia Rusak,Attila Juhos,Matthias Bethge,Wieland Brendel
In Search of Forgotten Domain Generalization
Out-of-Domain (OOD) generalization is the ability of a model trained on one or more domains to generalize to unseen domains. In the ImageNet era of computer vision, evaluation sets for measuring a model's OOD performance were designed to be strictly OOD with respect to style. However, the emergence of foundation models and expansive web-scale datasets has obfuscated this evaluation process, as datasets cover a broad range of domains and risk test domain contamination. In search of the forgotten domain generalization, we create large-scale datasets subsampled from LAION---LAION-Natural and LAION-Rendition---that are strictly OOD to corresponding ImageNet and DomainNet test sets in terms of style. Training CLIP models on these datasets reveals that a significant portion of their performance is explained by in-domain examples. This indicates that the OOD generalization challenges from the ImageNet era still prevail and that training on web-scale dat

a merely creates the illusion of OOD generalization. Furthermore, through a syst
ematic exploration of combining natural and rendition datasets in varying propor
tions, we identify optimal mixing ratios for model generalization across these d
omains. Our datasets and results re-enable meaningful assessment of OOD robustne
ss at scale---a crucial prerequisite for improving model robustness.
****
Wong Hiu Tung,Darrick Lee,Hong Yan
Towards Scalable Topological Regularizers
Latent space matching, which consists of matching distributions of features in l
atent space, is a crucial component for tasks such as adversarial attacks and de
fenses, domain adaptation, and generative modelling.
    Metrics for probability measures, such as Wasserstein and maximum mean discr
epancy, are commonly used to quantify the differences between such distributions
.
    However, these are often costly to compute, or do not appropriately take the
 geometric and topological features of the distributions into consideration.
    Persistent homology is a tool from topological data analysis which quantifie
s the multi-scale topological structure of point clouds, and has recently been u
sed as a topological regularizer in learning tasks.
    However, computation costs preclude larger scale computations, and discontin
uities in the gradient lead to unstable training behavior such as in adversarial
 tasks.
    We propose the use of principal persistence measures, based on computing the
 persistent homology of a large number of small subsamples, as a topological reg
ularizer.
    We provide a parallelized GPU implementation of this regularizer, and prove
that gradients are continuous for smooth densities.
    Furthermore, we demonstrate the efficacy of this regularizer on shape matchi
ng, image generation, and semi-supervised learning tasks, opening the door towar
ds a scalable regularizer for topological features.
****
Xiaofeng Wang,Yiming Wang,Wenhong Zhu,Rui Wang
Do Large Language Models Truly Understand Geometric Structures?
Geometric ability is a significant challenge for large language models (LLMs) du
e to the need for advanced spatial comprehension and abstract thinking. Existing
 datasets primarily evaluate LLMs on their final answers, but they cannot truly
measure their true understanding of geometric structures, as LLMs can arrive at
correct answers by coincidence. To fill this gap, we introduce the GeomRel datas
et, designed to evaluate LLMs' understanding of geometric structures by isolatin
g the core step of geometric relationship identification in problem-solving. Usi
ng this benchmark, we conduct thorough evaluations of diverse LLMs and identify
key limitations in understanding geometric structures. We further propose the Ge
ometry Chain-of-Thought (GeoCoT) method, which enhances LLMs' ability to identif
y geometric relationships, resulting in significant performance improvements.
****
Zenan Li,Zhaoyu Li,Wen Tang,Xian Zhang,Yuan Yao,Xujie Si,Fan Yang,Kaiyu Yang,Xia
oxing Ma
Proving Olympiad Inequalities by Synergizing LLMs and Symbolic Reasoning
Large language models (LLMs) can prove mathematical theorems formally by generat
ing proof steps (\textit{a.k.a.} tactics) within a proof system. However, the sp
ace of possible tactics is vast and complex, while the available training data f
or formal proofs is limited, posing a significant challenge to LLM-based tactic
generation. To address this, we introduce a neuro-symbolic tactic generator that
 synergizes the mathematical intuition learned by LLMs with domain-specific insi
ghts encoded by symbolic methods. The key aspect of this integration is identify
ing which parts of mathematical reasoning are best suited to LLMs and which to s
ymbolic methods. While the high-level idea of neuro-symbolic integration is broa
dly applicable to various mathematical problems, in this paper, we focus specifi
cally on Olympiad inequalities (Figure~1). We analyze how humans solve these pro
blems and distill the techniques into two types of tactics: (1) scaling, handled

by symbolic methods, and (2) rewriting, handled by LLMs. In addition, we combine symbolic tools with LLMs to prune and rank the proof goals for efficient proof search. We evaluate our framework on 161 challenging inequalities from multiple mathematics competitions, achieving state-of-the-art performance and significantly outperforming existing LLM and symbolic approaches without requiring additional training data.
****
Michael Noukhovitch,Shengyi Huang,Sophie Xhonneux,Arian Hosseini,Rishabh Agarwal,Aaron Courville
Asynchronous RLHF: Faster and More Efficient Off-Policy RL for Language Models
The dominant paradigm for RLHF is *online* and *on-policy* RL: synchronously generating from the large language model (LLM) policy, labelling with a reward model, and learning using feedback on the LLM's own outputs. While performant, this paradigm is computationally inefficient. Inspired by classical deep RL literature, we propose separating generation and learning in RLHF. This enables asynchronous generation of new samples while simultaneously training on old samples, leading to faster training and more compute-optimal scaling. However, asynchronous training relies on an underexplored regime, online but *off-policy* RLHF: learning on samples from previous iterations of our model which give a worse training signal. We tackle the fundamental challenge in this regime: how much off-policyness can we tolerate for asynchronous training to speed up learning but maintain performance? Among several RLHF algorithms we test, online DPO is found to be most robust to off-policy data, and robustness increases with the scale of the policy model. We study further compute optimizations for asynchronous RLHF but find that they come at a performance cost, giving rise to a trade-off. We verify the scalability of asynchronous RLHF by training a general-purpose chatbot from LLaMA 3.1 8B on an instruction-following task $\sim$40\% faster than a synchronous run while matching final performance. Finally, we extend our results to math and reasoning to demonstrate asynchronous RL can finetune Rho 1B on GSM8k $\sim$70\% faster while matching synchronous accuracy.
****
Nicholas Gao,Eike Eberhard,Stephan Günnemann
Learning Equivariant Non-Local Electron Density Functionals
The accuracy of density functional theory hinges on the approximation of non-local contributions to the exchange-correlation (XC) functional. To date, machine-learned and human-designed approximations suffer from insufficient accuracy, limited scalability, or dependence on costly reference data. To address these issues, we introduce Equivariant Graph Exchange Correlation (EG-XC), a novel non-local XC functional based on equivariant graph neural networks (GNNs). Where previous works relied on semi-local functionals or fixed-size descriptors of the density, we compress the electron density into an SO(3)-equivariant nuclei-centered point cloud for efficient non-local atomic-range interactions. By applying an equivariant GNN on this point cloud, we capture molecular-range interactions in a scalable and accurate manner. To train EG-XC, we differentiate through a self-consistent field solver requiring only energy targets. In our empirical evaluation, we find EG-XC to accurately reconstruct `gold-standard' CCSD(T) energies on MD17. On out-of-distribution conformations of 3BPA, EG-XC reduces the relative MAE by 35% to 50%. Remarkably, EG-XC excels in data efficiency and molecular size extrapolation on QM9, matching force fields trained on 5 times more and larger molecules. On identical training sets, EG-XC yields on average 51% lower MAEs.
****
Ji Qi,Ming Ding,Weihan Wang,Yushi Bai,Qingsong Lv,Wenyi Hong,Bin Xu,Lei Hou,Juanzi Li,Yuxiao Dong,Jie Tang
CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning
Vision-Language Models (VLMs) have shown broad effectiveness due to extensive training that aligns visual inputs with corresponding language responses. However, this conclusive alignment training causes models to overlook essential visual reasoning, leading to failures in handling detailed visual tasks and producing unfaithful responses. Drawing inspiration from human cognition in solving visual problems (e.g., marking, zoom in), this paper introduces Chain of Manipulations,

a mechanism that enables VLMs to tackle problems step-by-step with evidence. Aft
er training, models can solve various visual problems by eliciting intrinsic man
ipulations (e.g., grounding, zoom in) with results (e.g., boxes, image) actively
without relying external tools, while also allowing users to trace error causes
. In this paper, we study the comprehensive methodology that includes: (1) a fle
xible design of manipulations based on extensive analysis, (2) an efficient auto
mated data generation pipeline, (3) a compatible VLM architecture capable of mul
ti-turn, multi-image, and (4) a model training process for versatile capabilitie
s. With the design, we also manually annotate **6K** high-quality samples for ch
allenging graphical mathematical problems. Our trained model, CogCoM, equipped w
ith this mechanism and 17B parameters, achieves SOTA performance across **9** be
nchmarks in **4** categories, demonstrating its effectiveness while maintaining
interpretability. Code, model, and data are available at https://github.com/THUD
M/CogCoM.
****
Yixuan Zhou,Long Bai,Sijia Cai,Bing Deng,Xing Xu,Heng Tao Shen
TAU-106K: A New Dataset for Comprehensive Understanding of Traffic Accident
Multimodal Large Language Models (MLLMs) have demonstrated impressive performanc
e in general visual understanding tasks. However, their potential for high-level
, fine-grained comprehension, such as anomaly understanding, remains unexplored.
Focusing on traffic accidents, a critical and practical scenario within anomaly
understanding, we investigate the advanced capabilities of MLLMs and propose TA
Bot, a multimodal MLLM specialized for accident-related tasks. To facilitate thi
s, we first construct TAU-106K, a large-scale multimodal dataset containing 106K
traffic accident videos and images collected from academic benchmarks and publi
c platforms. The dataset is meticulously annotated through a video-to-image anno
tation pipeline to ensure comprehensive and high-quality labels. Building upon T
AU-106K, we train TABot using a two-step approach designed to integrate multi-gr
anularity tasks, including accident recognition, spatial-temporal grounding, and
an auxiliary description task to enhance the model's understanding of accident
elements. Extensive experiments demonstrate TABot's superior performance in traf
fic accident understanding, highlighting not only its capabilities in high-level
anomaly comprehension but also the robustness of the TAU-106K benchmark. Our co
de and data will be available at https://github.com/cool-xuan/TABot.
****
Mingfei Han,Linjie Yang,Xiaojun Chang,Lina Yao,Heng Wang
Shot2Story: A New Benchmark for Comprehensive Understanding of Multi-shot Videos
A short clip of video may contain progression of multiple events and an interest
ing story line. A human need to capture both the event in every shot and associa
te them together to understand the story behind it. In this work, we present a n
ew multi-shot video understanding benchmark \dataset with detailed shot-level ca
ptions, comprehensive video summaries and question-answering pairs. To facilitat
e better semantic understanding of videos, we provide captions for both visual s
ignals and human narrations. We design several distinct tasks including single-s
hot video captioning, multi-shot video summarization, and multi-shot video quest
ion answering. Preliminary experiments show some challenges to generate a long a
nd comprehensive video summary for multi-shot videos. Nevertheless, the generate
d imperfect summaries can already achieve competitive performance on existing vi
deo understanding tasks such as video question-answering, promoting an under-exp
lored setting of video understanding with detailed summaries.
****
Anji Liu,Oliver Broadrick,Mathias Niepert,Guy Van den Broeck
Discrete Copula Diffusion
Discrete diffusion models have recently shown significant progress in modeling c
omplex data, such as natural languages and DNA sequences. However, unlike diffus
ion models for continuous data, which can generate high-quality samples in just
a few denoising steps, modern discrete diffusion models still require hundreds o
r even thousands of denoising steps to perform well. In this paper, we identify
a fundamental limitation that prevents discrete diffusion models from achieving
strong performance with fewer steps -- they fail to capture dependencies between

output variables at each denoising step. To address this issue, we provide a formal explanation and introduce a general approach to supplement the missing dependency information by incorporating another deep generative model, termed the copula model. Our method does not require fine-tuning either the diffusion model or the copula model, yet it enables high-quality sample generation with significantly fewer denoising steps. When we apply this approach to autoregressive copula models, the combined model outperforms both models individually in unconditional and conditional text generation. Specifically, the hybrid model achieves better (un)conditional text generation using 8 to 32 times fewer denoising steps than the diffusion model alone. In addition to presenting an effective discrete diffusion generation algorithm, this paper emphasizes the importance of modeling inter-variable dependencies in discrete diffusion.
****
Geeling Chau,Christopher Wang,Sabera J Talukder,Vighnesh Subramaniam,Saraswati Soedarmadji,Yisong Yue,Boris Katz,Andrei Barbu
Population Transformer: Learning Population-level Representations of Neural Activity
We present a self-supervised framework that learns population-level codes for arbitrary ensembles of neural recordings at scale. We address key challenges in scaling models with neural time-series data, namely, sparse and variable electrode distribution across subjects and datasets. The Population Transformer (PopT) stacks on top of pretrained temporal embeddings and enhances downstream decoding by enabling learned aggregation of multiple spatially-sparse data channels. The pretrained PopT lowers the amount of data required for downstream decoding experiments, while increasing accuracy, even on held-out subjects and tasks. Compared to end-to-end methods, this approach is computationally lightweight, while achieving similar or better decoding performance. We further show how our framework is generalizable to multiple time-series embeddings and neural data modalities. Beyond decoding, we interpret the pretrained and fine-tuned PopT models to show how they can be used to extract neuroscience insights from large amounts of data. We release our code as well as a pretrained PopT to enable off-the-shelf improvements in multi-channel intracranial data decoding and interpretability. Code is available at https://github.com/czlwang/PopulationTransformer.
****
Ameer Qaqish,Didong Li
Identifiability for Gaussian Processes with Holomorphic Kernels
Gaussian processes (GPs) are widely recognized for their robustness and flexibility across various domains, including machine learning, time series, spatial statistics, and biomedicine. In addition to their common usage in regression tasks, GP kernel parameters are frequently interpreted in various applications. For example, in spatial transcriptomics, estimated kernel parameters are used to identify spatial variable genes, which exhibit significant expression patterns across different tissue locations. However, before these parameters can be meaningfully interpreted, it is essential to establish their identifiability. Existing studies of GP parameter identifiability have focused primarily on Mat\'ern-type kernels, as their spectral densities allow for more established mathematical tools. In many real-world applications, particuarly in time series analysis, other kernels such as the squared exponential, periodic, and rational quadratic kernels, as well as their combinations, are also widely used. These kernels share the property of being holomorphic around zero, and their parameter identifiability remains underexplored.
In this paper, we bridge this gap by developing a novel theoretical framework for determining kernel parameter identifiability for kernels holomorphic near zero. Our findings enable practitioners to determine which parameters are identifiable in both existing and newly constructed kernels, supporting application-specific interpretation of the identifiable parameters, and highlighting non-identifiable parameters that require careful interpretation.
****
Zhenrui Yue,Honglei Zhuang,Aijun Bai,Kai Hui,Rolf Jagerman,Hansi Zeng,Zhen Qin,Dong Wang,Xuanhui Wang,Michael Bendersky

Inference Scaling for Long-Context Retrieval Augmented Generation
The scaling of inference computation has unlocked the potential of long-context large language models (LLMs) across diverse settings. For knowledge-intensive tasks, the increased compute is often allocated to incorporate more external knowledge. However, without effectively utilizing such knowledge, solely expanding context does not always enhance performance. In this work, we investigate inference scaling for retrieval augmented generation (RAG), exploring the combination of multiple strategies beyond simply increasing the quantity of knowledge, including in-context learning and iterative prompting. These strategies provide additional flexibility to scale test-time computation (e.g., by increasing retrieved documents or generation steps), thereby enhancing LLMs' ability to effectively acquire and utilize contextual information. We address two key questions: (1) How does RAG performance benefit from the scaling of inference computation when optimally configured? (2) Can we predict the optimal test-time compute allocation for a given budget by modeling the relationship between RAG performance and inference parameters? Our observations reveal that increasing inference computation leads to nearly linear gains in RAG performance when optimally allocated, a relationship we describe as the inference scaling laws for RAG. Building on this, we further develop the computation allocation model to estimate RAG performance across different inference configurations. The model predicts optimal inference parameters under various computation constraints, which align closely with the experimental results. By applying these optimal configurations, we demonstrate that scaling inference compute on long-context LLMs achieves up to 58.9% gains on benchmark datasets compared to standard RAG.
****
Younwoo Choi,Muhammad Adil Asif,Ziwen Han,John Willes,Rahul Krishnan
Teaching LLMs How to Learn with Contextual Fine-Tuning
Prompting Large Language Models (LLMs), or providing context on the expected model of operation, is an effective way to steer the outputs of such models to satisfy human desiderata after they have been trained. But in rapidly evolving domains, there is often need to fine-tune LLMs to improve either the kind of knowledge in their memory or their abilities to perform open ended reasoning in new domains. When human's learn new concepts, we often do so by linking the new material that we are studying to concepts we have already learned before. To that end, we ask, "can prompting help us teach LLMs how to learn". In this work, we study a novel generalization of instruction tuning, called contextual fine-tuning, to fine-tune LLMs. Our method leverages instructional prompts designed to mimic human cognitive strategies in learning and problem-solving to guide the learning process during training, aiming to improve the model's interpretation and understanding of domain-specific knowledge. We empirically demonstrate that this simple yet effective modification improves the ability of LLMs to be fine-tuned rapidly on new datasets both within the medical and financial domains.
****
Wei Chen,Yuxuan Liang
Expand and Compress: Exploring Tuning Principles for Continual Spatio-Temporal Graph Forecasting
The widespread deployment of sensing devices leads to a surge in data for spatio-temporal forecasting applications such as traffic flow, air quality, and wind energy. Although spatio-temporal graph neural networks (STGNNs) have achieved success in modeling various static spatio-temporal forecasting scenarios, real-world spatio-temporal data are typically received in a streaming manner, and the network continuously expands with the installation of new sensors. Thus, spatio-temporal forecasting in streaming scenarios faces dual challenges: the inefficiency of retraining models over newly-arrived data and the detrimental effects of catastrophic forgetting over long-term history. To address these challenges, we propose a novel prompt tuning-based continuous forecasting method, **_EAC_**, following two fundamental tuning principles guided by empirical and theoretical analysis: _**e**xpand **a**nd **c**ompress_, which effectively resolve the aforementioned problems with lightweight tuning parameters. Specifically, we integrate the base STGNN with a continuous prompt pool, utilizing stored prompts (\ie, few le

arnable parameters) in memory, and jointly optimize them with the base STGNN. Th
is method ensures that the model sequentially learns from the spatio-temporal da
ta stream to accomplish tasks for corresponding periods. Extensive experimental
results on multiple real-world datasets demonstrate the multi-faceted superiorit
y of **_EAC_** over the state-of-the-art baselines, including effectiveness, eff
iciency, universality, etc.
****

Peiyan Hu,Rui Wang,Xiang Zheng,Tao Zhang,Haodong Feng,Ruiqi Feng,Long Wei,Yue Wa
ng,Zhi-Ming Ma,Tailin Wu
Wavelet Diffusion Neural Operator
Simulating and controlling physical systems described by partial differential eq
uations (PDEs) are crucial tasks across science and engineering. Recently, diffu
sion generative models have emerged as a competitive class of methods for these
tasks due to their ability to capture long-term dependencies and model high-dime
nsional states. However, diffusion models typically struggle with handling syste
m states with abrupt changes and generalizing to higher resolutions. In this wor
k, we propose Wavelet Diffusion Neural Operator (WDNO), a novel PDE simulation a
nd control framework that enhances the handling of these complexities. WDNO comp
rises two key innovations. Firstly, WDNO performs diffusion-based generative mod
eling in the wavelet domain for the entire trajectory to handle abrupt changes a
nd long-term dependencies effectively. Secondly, to address the issue of poor ge
neralization across different resolutions, which is one of the fundamental tasks
 in modeling physical systems, we introduce multi-resolution training. We valida
te WDNO on five physical systems, including 1D advection equation, three challen
ging physical systems with abrupt changes (1D Burgers' equation, 1D compressible
 Navier-Stokes equation and 2D incompressible fluid), and a real-world dataset E
RA5, which demonstrates superior performance on both simulation and control task
s over state-of-the-art methods, with significant improvements in long-term and
detail prediction accuracy. Remarkably, in the challenging context of the 2D hig
h-dimensional and indirect control task aimed at reducing smoke leakage, WDNO re
duces the leakage by 33.2% compared to the second-best baseline.
****

Zijie Geng,Jie Wang,Xijun Li,Fangzhou Zhu,Jianye HAO,Bin Li,Feng Wu
Differentiable Integer Linear Programming
Machine learning (ML) techniques have shown great potential in generating high-q
uality solutions for integer linear programs (ILPs).
However, existing methods typically rely on a *supervised learning* paradigm, le
ading to (1) *expensive training cost* due to repeated invocations of traditiona
l solvers to generate training labels, and (2) *plausible yet infeasible solutio
ns* due to the misalignment between the training objective (minimizing predictio
n loss) and the inference objective (generating high-quality solutions).
To tackle this challenge, we propose **DiffILO** (**Diff**erentiable **I**nteger
 Linear Programming **O**ptimization), an *unsupervised learning paradigm fo
r learning to solve ILPs*.
Specifically, through a novel probabilistic modeling, DiffILO reformulates ILPs-
--discrete and constrained optimization problems---into continuous, differentiab
le (almost everywhere), and unconstrained optimization problems.
This reformulation enables DiffILO to simultaneously solve ILPs and train the mo
del via straightforward gradient descent, providing two major advantages.
First, it significantly reduces the training cost, as the training process does
not need the aid of traditional solvers at all.
Second, it facilitates the generation of feasible and high-quality solutions, as
 the model *learns to solve ILPs* in an end-to-end manner, thus aligning the tra
ining and inference objectives.
Experiments on commonly used ILP datasets demonstrate that DiffILO not only achi
eves an average training speedup of 13.2 times compared to supervised methods,
 but also outperforms them by generating heuristic solutions with significantly
higher feasibility ratios and much better solution qualities.
****

Hoang V. Tran,Thanh Chu,Minh-Khoi Nguyen-Nhat,Huyen Trang Pham,Tam Le,Tan Minh N

guyen
Spherical Tree-Sliced Wasserstein Distance
Sliced Optimal Transport (OT) simplifies the OT problem in high-dimensional spac
es by projecting supports of input measures onto one-dimensional lines, then exp
loiting the closed-form expression of the univariate OT to reduce the computatio
nal burden of OT. Recently, the Tree-Sliced method has been introduced to replac
e these lines with more intricate structures, known as tree systems. This approa
ch enhances the ability to capture topological information of integration domain
s in Sliced OT while maintaining low computational cost. Inspired by this approa
ch, in this paper, we present an adaptation of tree systems on OT problem for me
asures supported on a sphere. As counterpart to the Radon transform variant on t
ree systems, we propose a novel spherical Radon transform, with a new integratio
n domain called spherical trees. By leveraging this transform and exploiting the
 spherical tree structures, we derive closed-form expressions for OT problems on
 the sphere. Consequently, we obtain an efficient metric for measures on the sph
ere, named Spherical Tree-Sliced Wasserstein (STSW) distance. We provide an exte
nsive theoretical analysis to demonstrate the topology of spherical trees, the w
ell-definedness and injectivity of our Radon transform variant, which leads to a
n orthogonally invariant distance between spherical measures. Finally, we conduc
t a wide range of numerical experiments, including gradient flows and self-super
vised learning, to assess the performance of our proposed metric, comparing it t
o recent benchmarks.
****
Samuel Lippl,Kim Stachenfeld
When does compositional structure yield compositional generalization? A kernel t
heory.
Compositional generalization (the ability to respond correctly to novel combinat
ions of familiar components) is thought to be a cornerstone of intelligent behav
ior. Compositionally structured (e.g. disentangled) representations support this
 ability; however, the conditions under which they are sufficient for the emerge
nce of compositional generalization remain unclear. To address this gap, we pres
ent a theory of compositional generalization in kernel models with fixed, compos
itionally structured representations. This provides a tractable framework for ch
aracterizing the impact of training data statistics on generalization. We find t
hat these models are limited to functions that assign values to each combination
 of components seen during training, and then sum up these values ("conjunction-
wise additivity"). This imposes fundamental restrictions on the set of tasks com
positionally structured kernel models can learn, in particular preventing them f
rom transitively generalizing equivalence relations. Even for compositional task
s that they can learn in principle, we identify novel failure modes in compositi
onal generalization (memorization leak and shortcut bias) that arise from biases
 in the training data. Finally, we empirically validate our theory, showing that
 it captures the behavior of deep neural networks (convolutional networks, resid
ual networks, and Vision Transformers) trained on a set of compositional tasks w
ith similarly structured data. Ultimately, this work examines how statistical st
ructure in the training data can affect compositional generalization, with impli
cations for how to identify and remedy failure modes in deep learning models.
****
Wenyu Jiang,Zhenlong Liu,Zejian Xie,Songxin Zhang,Bingyi Jing,Hongxin Wei
Exploring Learning Complexity for Efficient Downstream Dataset Pruning
The ever-increasing fine-tuning cost of large-scale pre-trained models gives ris
e to the importance of dataset pruning, which aims to reduce dataset size while
maintaining task performance.
However, existing dataset pruning methods require training on the entire dataset
, which is impractical for large-scale pre-trained models.
In this paper, we propose a straightforward, novel, and training-free hardness s
core named Distorting-based Learning Complexity (DLC), to identify informative i
mages and instructions from the downstream dataset efficiently.
Our method is motivated by the observation that easy samples learned faster can
also be learned with fewer parameters.

Specifically, we define the Learning Complexity to quantify sample hardness and utilize a lightweight weights masking process for fast estimation, instead of th e costly SGD optimization.
Based on DLC, we further design a flexible under-sampling strategy with randomne ss (dubbed FlexRand), replacing the top-K strategy, to alleviate the severe subs et distribution shift.
Extensive experiments with downstream image and instructions dataset pruning ben chmarks demonstrate the effectiveness and efficiency of the proposed approach.
In the images pruning benchmark, DLC significantly reduces the pruning time by 3 5$\times$ while establishing state-of-the-art performance with FlexRand.
****

Matthew X. Burns,Qingyuan Hou,Michael Huang
Provable Convergence Bounds for Hybrid Dynamical Sampling and Optimization
Analog dynamical accelerators (DXs) are a growing sub-field in computer architec ture research, offering order-of-magnitude gains in power efficiency and latency over traditional digital methods in several machine learning, optimization, and sampling tasks. However, limited-capacity accelerators require hybrid analog/di gital algorithms to solve real-world problems, commonly using large-neighborhood local search (LNLS) frameworks. Unlike fully digital algorithms, hybrid LNLS ha s no non-asymptotic convergence guarantees and no principled hyperparameter sele ction schemes, particularly limiting cross-device training and inference.

In this work, we provide non-asymptotic convergence guarantees for hybrid LNLS b y reducing to block Langevin Diffusion (BLD) algorithms.
Adapting tools from classical sampling theory, we prove exponential KL-divergenc e convergence for randomized and cyclic block selection strategies using ideal D Xs. With finite device variation, we provide explicit bounds on the 2-Wasserstei n bias in terms of step duration, noise strength, and function parameters. Our B LD model provides a key link between established theory and novel computing plat forms, and our theoretical results provide a closed-form expression linking devi ce variation, algorithm hyperparameters, and performance.
****

Yu Fu,Zefan Cai,Abedelkadir Asi,Wayne Xiong,Yue Dong,Wen Xiao
Not All Heads Matter: A Head-Level KV Cache Compression Method with Integrated R etrieval and Reasoning
Key-Value (KV) caching is a common technique to enhance the computational effici ency of Large Language Models (LLMs), but its memory overhead grows rapidly with input length. Prior work has shown that not all tokens are equally important fo r text generation, proposing layer-level KV cache compression to selectively ret ain key information. Recognizing the distinct roles of attention heads in genera tion, we propose HeadKV, a head-level KV cache compression method, and HeadKV-R2 , which leverages a novel contextual reasoning ability estimation for compressio n. Our approach operates at the level of individual heads, estimating their impo rtance for contextual QA tasks that require both retrieval and reasoning capabil ities. Extensive experiments across diverse benchmarks (LongBench, LooGLE), mode l architectures (e.g., Llama-3-8B-Instruct, Mistral-7B-Instruct), and long-conte xt abilities tests demonstrate that our head-level KV cache compression signific antly outperforms strong baselines, particularly in low-resource settings (KV si ze = 64 & 128). Notably, our method retains just 1.5% of the KV cache while achi eving 97% of the performance of the full KV cache on the contextual question ans wering benchmark.
****

Song Tang,Wenxin Su,Yan Gan,Mao Ye,Jianwei Dr. Zhang,Xiatian Zhu
Proxy Denoising for Source-Free Domain Adaptation
Source-Free Domain Adaptation (SFDA) aims to adapt a pre-trained source model to an unlabeled target domain with no access to the source data. Inspired by the s uccess of large Vision-Language (ViL) models in many applications, the latest re search has validated ViL's benefit for SFDA by using their predictions as pseudo supervision. However, we observe that ViL's supervision could be noisy and inac

curate at an unknown rate, potentially introducing additional negative effects during adaption. To address this thus-far ignored challenge, we introduce a novel Proxy Denoising (__ProDe__) approach. The key idea is to leverage the ViL model as a proxy to facilitate the adaptation process towards the latent domain-invariant space. Concretely, we design a proxy denoising mechanism to correct ViL's predictions. This is grounded on a proxy confidence theory that models the dynamic effect of proxy's divergence against the domain-invariant space during adaptation. To capitalize the corrected proxy, we further derive a mutual knowledge distilling regularization. Extensive experiments show that ProDe significantly outperforms the current state-of-the-art alternatives under both conventional closed-set setting and the more challenging open-set, partial-set, generalized SFDA, multi-target, multi-source, and test-time settings. Our code and data are available at https://github.com/tntek/source-free-domain-adaptation.
****

Yanqi Dai,Huanran Hu,Lei Wang,Shengjie Jin,Xu Chen,Zhiwu Lu
MMRole: A Comprehensive Framework for Developing and Evaluating Multimodal Role-Playing Agents
Recently, Role-Playing Agents (RPAs) have garnered increasing attention for their potential to deliver emotional value and facilitate sociological research.
However, existing studies are primarily confined to the textual modality, unable to simulate humans' multimodal perceptual capabilities.
To bridge this gap, we introduce the concept of Multimodal Role-Playing Agents (MRPAs), and propose a comprehensive framework, MMRole, for their development and evaluation, which comprises a personalized multimodal dataset and a robust evaluation approach.
Specifically, we construct a large-scale, high-quality dataset, MMRole-Data, consisting of 85 characters, 11K images, and 14K single or multi-turn dialogues.
Additionally, we present a robust evaluation approach, MMRole-Eval, encompassing eight metrics across three dimensions, where a reward model is designed to score MRPAs with the constructed ground-truth data for comparison.
Moreover, we develop the first specialized MRPA, MMRole-Agent.
Extensive evaluation results demonstrate the improved performance of MMRole-Agent and highlight the primary challenges in developing MRPAs, emphasizing the need for enhanced multimodal understanding and role-playing consistency.
The data, code, and models are all available at https://github.com/YanqiDai/MMRole.
****

Yue Zhang,Zhiyang Xu,Ying Shen,Parisa Kordjamshidi,Lifu Huang
SPARTUN3D: Situated Spatial Understanding of 3D World in Large Language Model
Integrating the 3D world into large language models (3D-based LLMs) has been a promising research direction for 3D scene understanding. However, current 3D-based LLMs fall short in situated understanding due to two key limitations: 1) existing 3D datasets are constructed from a global perspective of the 3D scenes and lack situated context.
2) the architectures of the current 3D-based LLMs lack an explicit mechanism for aligning situated spatial information between 3D representations and natural language, limiting their performance in tasks requiring precise spatial reasoning.

In this work, we address these issues by introducing a scalable situated 3D dataset, named Spartun3D, that incorporates various situated spatial information.
In addition, we propose a situated spatial alignment module to enhance the learning between 3D visual representations and their corresponding textual descriptions. Our experimental results demonstrate that both our dataset and alignment module enhance situated spatial understanding ability.
****

BOLUN SUN,Yifan Zhou,Haiyun Jiang
Empowering Users in Digital Privacy Management through Interactive LLM-Based Agents
This paper presents a novel application of large language models (LLMs) to enhance user comprehension of privacy policies through an interactive dialogue agent.

We demonstrate that LLMs significantly outperform traditional models in tasks like Data Practice Identification, Choice Identification, Policy Summarization, and Privacy Question Answering, setting new benchmarks in privacy policy analysis. Building on these findings, we introduce an innovative LLM-based agent that functions as an expert system for processing website privacy policies, guiding users through complex legal language without requiring them to pose specific questions. A user study with 100 participants showed that users assisted by the agent had higher comprehension levels (mean score of 2.6 out of 3 vs. 1.8 in the control group), reduced cognitive load (task difficulty ratings of 3.2 out of 10 vs. 7.8), increased confidence in managing privacy, and completed tasks in less time (5.5 minutes vs. 15.8 minutes). This work highlights the potential of LLM-based agents to transform user interaction with privacy policies, leading to more informed consent and empowering users in the digital services landscape.
****

Matthew Eric Lee,Jaesik Kim,Matei Ionita,Jonghyun Lee,Michelle L. McKeague,YONGHYUN NAM,Irene Khavin,Yidi Huang,Victoria Fang,Sokratis Apostolidis,Divij Mathew,Shwetank,Ajinkya Pattekar,Zahabia Rangwala,Amit Bar-Or,Benjamin A Fensterheim,Benjamin A. Abramoff,Rennie L. Rhee,Damian Maseda,Allison R Greenplate,John Wherry,Dokyoon Kim

MAESTRO: Masked Encoding Set Transformer with Self-Distillation

The interrogation of cellular states and interactions in immunology research is an ever-evolving task, requiring adaptation to the current levels of high dimensionality. Cytometry enables high-dimensional profiling of immune cells, but its analysis is hindered by the complexity and variability of the data. We present MAESTRO, a self-supervised set representation learning model that generates vector representations of set-structured data, which we apply to learn immune profiles from cytometry data. Unlike previous studies only learn cell-level representations, whereas MAESTRO uses all of a sample's cells to learn a set representation. MAESTRO leverages specialized attention mechanisms to handle sets of variable number of cells and ensure permutation invariance, coupled with an online tokenizer by self-distillation framework. We benchmarked our model against existing cytometry approaches and other existing machine learning methods that have never been applied in cytometry. Our model outperforms existing approaches in retrieving cell-type proportions and capturing clinically relevant features for downstream tasks such as disease diagnosis and immune cell profiling.
****

Yuxuan Yao,Han Wu,Mingyang LIU,Sichun Luo,Xiongwei Han,Jie Liu,Zhijiang Guo,Linqi Song

Determine-Then-Ensemble: Necessity of Top-k Union for Large Language Model Ensembling

Large language models (LLMs) exhibit varying strengths and weaknesses across different tasks, prompting recent studies to explore the benefits of ensembling models to leverage their complementary advantages. However, existing LLM ensembling methods often overlook model compatibility and struggle with inefficient alignment of probabilities across the entire vocabulary. In this study, we empirically investigate the factors influencing ensemble performance, identifying model performance, vocabulary size, and response style as key determinants, revealing that compatibility among models is essential for effective ensembling. This analysis leads to the development of a simple yet effective model selection strategy that identifies compatible models. Additionally, we introduce the \textsc{Uni}on \textsc{T}op-k \textsc{E}nsembling (\textsc{UniTE}), a novel approach that efficiently combines models by focusing on the union of the top-k tokens from each model, thereby avoiding the need for full vocabulary alignment and reducing computational overhead. Extensive evaluations across multiple benchmarks demonstrate that \textsc{UniTE} significantly enhances performance compared to existing methods, offering a more efficient framework for LLM ensembling.
****

Dayuan Fu,Keqing He,Yejie Wang,Wentao Hong,Zhuoma GongQue,Weihao Zeng,Wei Wang,Jingang Wang,Xunliang Cai,Weiran Xu

AgentRefine: Enhancing Agent Generalization through Refinement Tuning

Large Language Model (LLM) based agents have proved their ability to perform complex tasks like humans. However, there is still a large gap between open-sourced LLMs and commercial models like the GPT series. In this paper, we focus on improving the agent generalization capabilities of LLMs via instruction tuning. We first observe that the existing agent training corpus exhibits satisfactory results on held-in evaluation sets but fails to generalize to held-out sets. These agent-tuning works face severe formatting errors and are frequently stuck in the same mistake for a long while. We analyze that the poor generalization ability comes from overfitting to several manual agent environments and a lack of adaptation to new situations. They struggle with the wrong action steps and can not learn from the experience but just memorize existing observation-action relations. Inspired by the insight, we propose a novel AgentRefine framework for agent-tuning. The core idea is to enable the model to learn to correct its mistakes via observation in the trajectory. Specifically, we propose an agent synthesis framework to encompass a diverse array of environments and tasks and prompt a strong LLM to refine its error action according to the environment feedback. AgentRefine significantly outperforms state-of-the-art agent-tuning work in terms of generalization ability on diverse agent tasks. It also has better robustness facing perturbation and can generate diversified thought in inference. Our findings establish the correlation between agent generalization and self-refinement and provide a new paradigm for future research.
****
Jian Chen,Ruiyi Zhang,Yufan Zhou,Tong Yu,Franck Dernoncourt,Jiuxiang Gu,Ryan A. Rossi,Changyou Chen,Tong Sun
SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding
Multimodal large language models (MLLMs) have recently shown great progress in text-rich image understanding, yet they still struggle with complex, multi-page visually-rich documents. Traditional methods using document parsers for retrieval-augmented generation suffer from performance and efficiency limitations, while directly presenting all pages to MLLMs leads to inefficiencies, especially with lengthy ones. In this work, we present a novel framework named **S**elf-**V**isual **R**etrieval-**A**ugmented **G**eneration (SV-RAG), which can broaden horizons of *any* MLLM to support long-document understanding. We demonstrate that **MLLMs themselves can be an effective multimodal retriever** to fetch relevant pages and then answer user questions based on these pages. SV-RAG is implemented with two specific MLLM adapters, one for evidence page retrieval and the other for question answering. Empirical results show state-of-the-art performance on public benchmarks, demonstrating the effectiveness of SV-RAG.
****
Hengquan Guo,Xin Liu
On Stochastic Contextual Bandits with Knapsacks in Small Budget Regime
This paper studies stochastic contextual bandits with knapsack constraints (CBwK), where a learner observes a context, takes an action, receives a reward, and incurs a vector of costs at every round. The learner aims to maximize the cumulative rewards across T rounds under the knapsack constraints with an initial budget of B. We study CBwK in the small budget regime where the budget $B = \Omega(\sqrt{T})$
and propose an Adaptive and Universal Primal--Dual algorithm (AUPD) that achieves strong regret performance:
i) AUPD achieves $\tilde{O}((1 + \frac{\nu^*}{\delta b})\sqrt{T})$ regret under the strict feasibility assumption without any prior information, matching the best-known bounds;
ii) AUPD achieves $\tilde{O}(\sqrt{T}+ \frac{\nu^*}{\sqrt{b}}T^{\frac{3}{4}})$ regret without strict feasibility assumption,
which, to the best of our knowledge, is the first result in the literature. Here, the parameter ν^* represents the optimal average reward; $b=B/T$ is the average budget and δb is the feasibility/safety margin.
We establish these strong results through the adaptive budget-aware design, which effectively balances reward maximization and budget consumption. We provide a

new perspective on analyzing budget consumption using the Lyapunov drift method, along with a refined analysis of its cumulative variance. Our theory is further supported by experiments conducted on a large-scale dataset.
****
Ganyu Wang,Boyu Wang,Bin Gu,Charles Ling
Event-Driven Online Vertical Federated Learning
Online learning is more adaptable to real-world scenarios in Vertical Federated Learning (VFL) compared to offline learning.
However, integrating online learning into VFL presents challenges due to the unique nature of VFL, where clients possess non-intersecting feature sets for the same sample.
In real-world scenarios, the clients may not receive data streaming for the disjoint features for the same entity synchronously. Instead, the data are typically generated by an *event* relevant to only a subset of clients.
We are the first to identify these challenges in online VFL, which have been overlooked by previous research. To address these challenges, we proposed an event-driven online VFL framework. In this framework, only a subset of clients were activated during each event, while the remaining clients passively collaborated in the learning process.
Furthermore, we incorporated *dynamic local regret (DLR)* into VFL to address the challenges posed by online learning problems with non-convex models within a non-stationary environment.
We conducted a comprehensive regret analysis of our proposed framework, specifically examining the DLR under non-convex conditions with event-driven online VFL.

Extensive experiments demonstrated that our proposed framework was more stable than the existing online VFL framework under non-stationary data conditions while also significantly reducing communication and computation costs.
****
Nguyen Nhat Minh,Andrew Baker,Clement Neo,Allen G Roush,Andreas Kirsch,Ravid Shwartz-Ziv
Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs
Large Language Models (LLMs) generate text by sampling the next token from a probability distribution over the vocabulary at each decoding step. Popular sampling methods like top-p (nucleus sampling) often struggle to balance quality and diversity, especially at higher temperatures which lead to incoherent or repetitive outputs. We propose min-p sampling, a dynamic truncation method that adjusts the sampling threshold based on the model's confidence by using the top token's probability as a scaling factor. Our experiments on benchmarks including GPQA, GSM8K, and AlpacaEval Creative Writing show that min-p sampling improves both the quality and diversity of generated text across different model families (Mistral and Llama 3) and model sizes (1B to 123B parameters), especially at higher temperatures. Human evaluations further show a clear preference for min-p sampling, in both text quality and creativity. Min-p sampling has been adopted by popular open-source LLM frameworks, including Hugging Face Transformers, VLLM, and many others, highlighting its significant impact on improving text generation quality.
****
Kira Michaela Düsterwald,Samo Hromadka,Makoto Yamada
Fast unsupervised ground metric learning with tree-Wasserstein distance
The performance of unsupervised methods such as clustering depends on the choice of distance metric between features, or ground metric. Commonly, ground metrics are decided with heuristics or learned via supervised algorithms. However, since many interesting datasets are unlabelled, unsupervised ground metric learning approaches have been introduced. One promising option employs Wasserstein singular vectors (WSVs), which emerge when computing optimal transport distances between features and samples simultaneously. WSVs are effective, but can be prohibitively computationally expensive in some applications: $\mathcal{O}(n^2m^2(n \log(n) + m \log(m))$ for n samples and m features. In this work, we propose to augment the WSV method by embedding samples and features on trees, on which we co

mpute the tree-Wasserstein distance (TWD). We demonstrate theoretically and empi
rically that the algorithm converges to a better approximation of the standard W
SV approach than the best known alternatives, and does so with $\mathcal{O}(n^3+
m^3+mn)$ complexity. In addition, we prove that the initial tree structure can b
e chosen flexibly, since tree geometry does not constrain the richness of the ap
proximation up to the number of edge weights. This proof suggests a fast and rec
ursive algorithm for computing the tree parameter basis set, which we find cruci
al to realising the efficiency gains at scale. Finally, we employ the tree-WSV a
lgorithm to several single-cell RNA sequencing genomics datasets, demonstrating
its scalability and utility for unsupervised cell-type clustering problems. Thes
e results poise unsupervised ground metric learning with TWD as a low-rank appro
ximation of WSV with the potential for widespread application.
****

Xiaoyu Zhang,Juan Zhai,Shiqing Ma,Chao Shen,Tianlin Li,Weipeng Jiang,Yang Liu
STAFF: Speculative Coreset Selection for Task-Specific Fine-tuning
Task-specific fine-tuning is essential for the deployment of large language mode
ls (LLMs), but it requires significant computational resources and time. Existin
g solutions have proposed coreset selection methods to improve data efficiency a
nd reduce model training overhead, but they still have limitations: ❶ Overlookin
g valuable samples at high pruning rates, which degrades the coreset’s performan
ce.
❷ Requiring high time overhead during coreset selection to fine-tune and evaluat
e the target LLM. In this paper, we introduce STAFF, a speculative coreset selec
tion method. STAFF leverages a small model from the same family as the target LL
M to efficiently estimate data scores and then verifies the scores on the target
 LLM to accurately identify and allocate more selection budget to important regi
ons while maintaining coverage of easy regions. We evaluate STAFF on three LLMs
and three downstream tasks and show that STAFF improves the performance of SOTA
methods by up to 54.3% and reduces selection overhead by up to 70.5% at differen
t pruning rates. Furthermore, we observe that the coreset selected by STAFF at l
ow pruning rates (i.e., 20%) can even obtain better fine-tuning performance than
 the full dataset.
****

Alberto Del Pia,Dekun Zhou,Yinglun Zhu
Efficient Sparse PCA via Block-Diagonalization
Sparse Principal Component Analysis (Sparse PCA) is a pivotal tool in data analy
sis and dimensionality reduction. However, Sparse PCA is a challenging problem i
n both theory and practice: it is known to be NP-hard and current exact methods
generally require exponential runtime. In this paper, we propose a novel framewo
rk to efficiently approximate Sparse PCA by (i) approximating the general input
 covariance matrix with a re-sorted block-diagonal matrix, (ii) solving the Spar
se PCA sub-problem in each block, and (iii) reconstructing the solution to the o
riginal problem. Our framework is simple and powerful: it can leverage any off-t
he-shelf Sparse PCA algorithm and achieve significant computational speedups, wi
th a minor additive error that is linear in the approximation error of the block
-diagonal matrix. Suppose $g(k, d)$ is the runtime of an algorithm (approximatel
y) solving Sparse PCA in dimension d and with sparsity constant k. Our frame
work, when integrated with this algorithm, reduces the runtime to $\mathcal{O}\l
eft(\frac{d}{d^\star} \cdot g(k, d^\star) + d^2\right)$, where $d^\star \leq d$
is the largest block size of the block-diagonal matrix. For instance, integratin
g our framework with the Branch-and-Bound algorithm reduces the complexity from
$g(k, d) = \mathcal{O}(k^3\cdot d^k)$ to $\mathcal{O}(k^3\cdot d \cdot (d^\star)
^{k-1})$, demonstrating exponential speedups if $d^\star$ is small. We perform l
arge-scale evaluations on many real-world datasets: for exact Sparse PCA algorit
hm, our method achieves an average speedup factor of 100.50, while maintaining a
n average approximation error of 0.61%; for approximate Sparse PCA algorithm, ou
r method achieves an average speedup factor of 6.00 and an average approximation
 error of -0.91%, meaning that our method oftentimes finds better solutions.
****

Jiwook Kim,Seonho Lee,Jaeyo Shin,Jiho Choi,Hyunjung Shim

DreamCatalyst: Fast and High-Quality 3D Editing via Controlling Editability and Identity Preservation

Score distillation sampling (SDS) has emerged as an effective framework in text-driven 3D editing tasks, leveraging diffusion models for 3D-consistent editing. However, existing SDS-based 3D editing methods suffer from long training times and produce low-quality results. We identify that the root cause of this performance degradation is their conflict with the sampling dynamics of diffusion models. Addressing this conflict allows us to treat SDS as a diffusion reverse process for 3D editing via sampling from data space. In contrast, existing methods naively distill the score function using diffusion models. From these insights, we propose DreamCatalyst, a novel framework that considers these sampling dynamics in the SDS framework. Specifically, we devise the optimization process of our DreamCatalyst to approximate the diffusion reverse process in editing tasks, thereby aligning with diffusion sampling dynamics. As a result, DreamCatalyst successfully reduces training time and improves editing quality. Our method offers two modes: (1) a fast mode that edits Neural Radiance Fields (NeRF) scenes approximately 23 times faster than current state-of-the-art NeRF editing methods, and (2) a high-quality mode that produces superior results about 8 times faster than these methods. Notably, our high-quality mode outperforms current state-of-the-art NeRF editing methods in terms of both speed and quality. DreamCatalyst also surpasses the state-of-the-art 3D Gaussian Splatting (3DGS) editing methods, establishing itself as an effective and model-agnostic 3D editing solution.

****

Ruizhe Shi,Runlong Zhou,Simon Shaolei Du

The Crucial Role of Samplers in Online Direct Preference Optimization

Direct Preference Optimization (DPO) has emerged as a stable, scalable, and efficient solution for language model alignment.

Despite its empirical success, the optimization properties, particularly the impact of samplers on its convergence rates, remain under-explored. In this paper, we provide a rigorous analysis of DPO's convergence rates with different sampling strategies under the exact gradient setting, revealing a surprising separation: uniform sampling achieves $\textbf{linear}$ convergence, while our proposed online sampler achieves $\textbf{quadratic}$ convergence. We further adapt the sampler to practical settings by incorporating posterior distributions and logit mixing, demonstrating improvements over previous methods. For example, it outperforms vanilla DPO by over 7.4% on Safe-RLHF dataset. Our results not only offer insights into the theoretical understanding of DPO but also pave the way for further algorithm designs.

****

Cornelius Emde,Alasdair Paren,Preetham Arvind,Maxime Guillaume Kayser,Tom Rainforth,Thomas Lukasiewicz,Philip Torr,Adel Bibi

Shh, don't say that! Domain Certification in LLMs

Large language models (LLMs) are often deployed to do constrained tasks, with narrow domains. For example, customer support bots can be built on top of LLMs, relying on their broad language understanding and capabilities to enhance performance. However, these LLMs are adversarially susceptible, potentially generating outputs outside the intended domain. To formalize, assess and mitigate this risk, we introduce domain certification; a guarantee that accurately characterizes the out-of-domain behavior of language models. We then propose a simple yet effective approach dubbed VALID that provides adversarial bounds as a certificate. Finally, we evaluate our method across a diverse set of datasets, demonstrating that it yields meaningful certificates.

****

Yue Huang,Siyuan Wu,Chujie Gao,Dongping Chen,Qihui Zhang,Yao Wan,Tianyi Zhou,Chaowei Xiao,Jianfeng Gao,Lichao Sun,Xiangliang Zhang

DataGen: Unified Synthetic Dataset Generation via Large Language Models

Large Language Models (LLMs) such as GPT-4 and Llama3 have significantly impacted various fields by enabling high-quality synthetic data generation and reducing dependence on expensive human-generated datasets.

Despite this, challenges remain in the areas of generalization, controllability,

diversity, and truthfulness within the existing generative frameworks. To addre
ss these challenges, this paper presents DataGen, a comprehensive LLM-powered fr
amework designed to produce diverse, accurate, and highly controllable datasets.
DataGen is adaptable, supporting all types of text datasets and enhancing the g
enerative process through innovative mechanisms. To augment data diversity, Data
Gen incorporates an attribute-guided generation module and a group checking feat
ure. For accuracy, it employs a code-based mathematical assessment for label ver
ification alongside a retrieval-augmented generation technique for factual valid
ation. The framework also allows for user-specified constraints, enabling custom
ization of the data generation process to suit particular requirements. Extensiv
e experiments demonstrate the superior quality of data generated by DataGen, and
each module within DataGen plays a critical role in this enhancement. Additiona
lly, DataGen is applied in two practical scenarios: benchmarking LLMs and data a
ugmentation. The results indicate that DataGen effectively supports dynamic and
evolving benchmarking and that data augmentation improves LLM capabilities in va
rious domains, including agent-oriented abilities and reasoning skills.
****

Xue Han,Yitong Wang,Junlan Feng,wenchun.gao,Qian Hu,Chao Deng
LOIRE: LifelOng learning on Incremental data via pre-trained language model gRow
th Efficiently
Large-scale pre-trained language models (PLMs) require significant computational
resources to train from scratch on large volumes of data. But in the real world
, emerging data from diverse sources may not be initially available for pre-trai
ning. Recent studies on lifelong learning have tried to solve this problem by ex
ploring the use of model growth techniques to effectively incorporate new knowle
dge without the need for complete re-training. However, model growth approaches
utilized have issues with growth operators that do not ensure strict function pr
eservation or growth schedules that only include a few growth dimensions, reduci
ng lifelong learning's effect. Furthermore, existing approaches often assume tha
t emerging data has the same distribution as pre-training data, causing catastro
phic forgetting of previously acquired knowledge. To address the aforementioned
issues, we introduce LOIRE, a framework for lifelong learning that enables PLMs
to effectively grow their capacity using incremental data. LOIRE employs growth
operators for all feasible dimensions and a growth schedule to generate the opti
mal expansion sequence in the field of lifelong learning. Specifically, we prese
nt a novel plug-in layer growth operator with residual connections that skip the
newly added layer during initial training while ensuring function preservation.
We additionally propose an iterative distillation strategy for LOIRE that allow
s an intermediate model in the growth stages to switch between being a student a
nd a teacher, reducing catastrophic forgetting during growth. Experiments show t
hat LOIRE can reduce computational expenses by an average of 29.22\% while retai
ning equivalent or better downstream performance.
****

Xu Wang,Fuyou Miao,Wenjie Liu,Yan Xiong
Efficient and Robust Neural Combinatorial Optimization via Wasserstein-Based Cor
esets
Combinatorial optimization (CO) is a fundamental tool in many fields.
Many neural combinatorial optimization (NCO) methods have been proposed to solve
CO problems.
However, existing NCO methods typically require significant computational and st
orage resources, and face challenges in maintaining robustness to distribution s
hifts between training and test data.
To address these issues, we model CO instances into probability measures, and in
troduce Wasserstein-based metrics to quantify the difference between CO instance
s.
We then leverage a popular data compression technique, \emph{coreset}, to constr
uct a small-size proxy for the original large dataset.
However, the time complexity of constructing a coreset is linearly dependent on
the size of the dataset. Consequently, it becomes challenging when datasets are
particularly large.

Further, we accelerate the coreset construction by adapting it to the merge-and-reduce framework, enabling parallel computing. Additionally, we prove that our coreset is a good representation in theory.
{Subsequently}, to speed up the training process for existing NCO methods, we propose an efficient training framework based on the coreset technique. We train the model on a small-size coreset rather than on the full dataset, and thus save substantial computational and storage resources. Inspired by hierarchical Gonzalez’s algorithm, our coreset method is designed to capture the diversity of the dataset, which consequently improves robustness to distribution shifts.
Finally, experimental results demonstrate that our training framework not only enhances robustness to distribution shifts but also achieves better performance with reduced resource requirements.
****

Weiwei Kong,Andres Munoz medina,Mónica Ribero
Differentially private optimization for non-decomposable objective functions
Unsupervised pre-training is a common step in developing computer vision models and large language models. In this setting, the absence of labels requires the use of similarity-based loss functions, such as the contrastive loss, that favor minimizing the distance between similar inputs and maximizing the distance between distinct inputs. As privacy concerns mount, training these models using differential privacy has become more important. However, due to how inputs are generated for these losses, one of their undesirable properties is that their L_2 sensitivity grows with the batch size. This property is particularly disadvantageous for differentially private training methods, such as DP-SGD. To overcome this issue, we develop a new DP-SGD variant for similarity based loss functions --- in particular, the commonly-used contrastive loss --- that manipulates gradients of the objective function in a novel way to obtain a sensitivity of the summed gradient that is $O(1)$ for batch size n. We test our DP-SGD variant on some CIFAR-10 pre-training and CIFAR-100 finetuning tasks and show that, in both tasks, our method's performance comes close to that of a non-private model and generally outperforms DP-SGD applied directly to the contrastive loss.
****

Zhao Yang,Bing Su,Chuan Cao,Ji-Rong Wen
Regulatory DNA Sequence Design with Reinforcement Learning
$\textit{Cis}$-regulatory elements (CREs), such as promoters and enhancers, are relatively short DNA sequences that directly regulate gene expression. The fitness of CREs, measured by their ability to modulate gene expression, highly depends on the nucleotide sequences, especially specific motifs known as transcription factor binding sites (TFBSs). Designing high-fitness CREs is crucial for therapeutic and bioengineering applications. Current CRE design methods are limited by two major drawbacks: (1) they typically rely on iterative optimization strategies that modify existing sequences and are prone to local optima, and (2) they lack the guidance of biological prior knowledge in sequence optimization. In this paper, we address these limitations by proposing a generative approach that leverages reinforcement learning (RL) to fine-tune a pre-trained autoregressive (AR) model. Our method incorporates data-driven biological priors by deriving computational inference-based rewards that simulate the addition of activator TFBSs and removal of repressor TFBSs, which are then integrated into the RL process. We evaluate our method on promoter design tasks in two yeast media conditions and enhancer design tasks for three human cell types, demonstrating its ability to generate high-fitness CREs while maintaining sequence diversity. The code is available at https://github.com/yangzhao1230/TACO.
****

Yutaka Shimizu,Masayoshi Tomizuka
Bisimulation Metric for Model Predictive Control
Model-based reinforcement learning (MBRL) has shown promise for improving sample efficiency and decision-making in complex environments. However, existing methods face challenges in training stability, robustness to noise, and computational efficiency. In this paper, we propose Bisimulation Metric for Model Predictive Control (BS-MPC), a novel approach that incorporates bisimulation metric loss in

its objective function to directly optimize the encoder. This optimization enab
les the learned encoder to extract intrinsic information from the original state
space while discarding irrelevant details. BS-MPC improves training stability,
robustness against input noise, and computational efficiency by reducing trainin
g time. We evaluate BS-MPC on both continuous control and image-based tasks from
the DeepMind Control Suite, demonstrating superior performance and robustness c
ompared to state-of-the-art baseline methods.
****
Divya Jyoti Bajpai,Manjesh Kumar Hanawal
BEEM: Boosting Performance of Early Exit DNNs using Multi-Exit Classifiers as Ex
perts
Early Exit (EE) techniques have emerged as a means to reduce inference latency i
n Deep Neural Networks (DNNs). The latency improvement and accuracy in these tec
hniques crucially depend on the criteria used to make exit decisions. We propose
a new decision criterion BEEM where exit classifiers are treated as experts and
aggregate their confidence scores. The confidence scores are aggregated only if
neighbouring experts are consistent in prediction as the samples pass through t
hem, thus capturing their ensemble effect. A sample exits when the aggregated co
nfidence value exceeds a threshold. The threshold is set using the error rates o
f the intermediate exits aiming to surpass the performance of conventional DNN i
nference. Experimental results on the COCO dataset for Image captioning and GLUE
datasets for various language tasks demonstrate that our method enhances the pe
rformance of state-of-the-art EE methods, achieving improvements in speed-up by
a factor $1.5\times$ to $2.1\times$. When compared to the final layer, its accur
acy is comparable in harder Image Captioning and improves in the easier language
tasks. The source code is available at https://github.com/Div290/BEEM1/tree/mai
n.
****
Yam Eitan,Yoav Gelberg,Guy Bar-Shalom,Fabrizio Frasca,Michael M. Bronstein,Hagga
i Maron
Topological Blindspots: Understanding and Extending Topological Deep Learning Th
rough the Lens of Expressivity
Topological deep learning (TDL) is a rapidly growing field that seeks to leverag
e topological structure in data and facilitate learning from data supported on t
opological objects, ranging from molecules to 3D shapes. Most TDL architectures
can be unified under the framework of higher-order message-passing (HOMP), which
generalizes graph message-passing to higher-order domains. In the first part of
the paper, we explore HOMP's expressive power from a topological perspective, d
emonstrating the framework's inability to capture fundamental topological and me
tric invariants such as diameter, orientability, planarity, and homology. In add
ition, we demonstrate HOMP's limitations in fully leveraging lifting and pooling
methods on graphs. To the best of our knowledge, this is the first work to stud
y the expressivity of TDL from a topological perspective. In the second part of
the paper, we develop two new classes of architectures -- multi-cellular network
s (MCN) and scalable MCN (SMCN) -- which draw inspiration from expressive GNNs.
MCN can reach full expressivity, but scaling it to large data objects can be com
putationally expansive. Designed as a more scalable alternative, SMCN still miti
gates many of HOMP's expressivity limitations. Finally, we design new benchmarks
for evaluating models based on their ability to learn topological properties of
complexes. We then evaluate SMCN on these benchmarks as well as on real-world g
raph datasets, demonstrating improvements over both HOMP baselines and expressiv
e graph methods, highlighting the value of expressively leveraging topological i
nformation.
****
Wenhao Wu,Yizhong Wang,Guangxuan Xiao,Hao Peng,Yao Fu
Retrieval Head Mechanistically Explains Long-Context Factuality
Despite the recent progress in long-context language models, it remains elusive
how transformer-based models exhibit the capability to retrieve relevant informa
tion from arbitrary locations within the long context. This paper aims to addres
s this question. Our systematic investigation across a wide spectrum of models r

eveals that a special type of attention heads are largely responsible for retrie
ving information, which we dub retrieval heads. We identify intriguing propertie
s of retrieval heads:(1) universal: all the explored models with long-context ca
pability have a set of retrieval heads; (2) sparse: only a small portion (less t
han 5\%) of the attention heads are retrieval. (3) intrinsic: retrieval heads al
ready exist in models pretrained with short context. When extending the context
length by continual pretraining, it is still the same set of heads that perform
information retrieval. (4) dynamically activated: take Llama-2 7B for example, 1
2 retrieval heads always attend to the required information no matter how the co
ntext is changed. The rest of the retrieval heads are activated in different con
texts. (5) causal: completely pruning retrieval heads leads to failure in retrie
ving relevant information and results in hallucination, while pruning random non
-retrieval heads does not affect the model's retrieval ability. We further show
that retrieval heads strongly influence chain-of-thought (CoT) reasoning, where
the model needs to frequently refer back the question and previously-generated c
ontext. Conversely, tasks where the model directly generates the answer using it
s intrinsic knowledge are less impacted by masking out retrieval heads. These ob
servations collectively explain which internal part of the model seeks informati
on from the input tokens. We believe our insights will foster future research on
 reducing hallucination, improving reasoning, and compressing the KV cache.
****

Itay Beit Halachmi,Ido Kaminer

The Ramanujan Library - Automated Discovery on the Hypergraph of Integer Relatio
ns

Fundamental mathematical constants appear in nearly every field of science, from
 physics to biology.
Formulas that connect different constants often bring great insight by hinting a
t connections between previously disparate fields.
Discoveries of such relations, however, have remained scarce events, relying on
sporadic strokes of creativity by human mathematicians.
Recent developments of algorithms for automated conjecture generation have accel
erated the discovery of formulas for specific constants.
Yet, the discovery of connections between constants has not been addressed.
In this paper, we present the first library dedicated to mathematical constants
and their interrelations. This library can serve as a central repository of know
ledge for scientists from different areas, and as a collaborative platform for d
evelopment of new algorithms.
The library is based on a new representation that we propose for organizing the
formulas of mathematical constants:
a hypergraph, with each node representing a constant and each edge representing
a formula.
Using this representation, we propose and demonstrate a systematic approach for
automatically enriching this library using PSLQ, an integer relation algorithm b
ased on QR decomposition and lattice construction. During its development and te
sting, our strategy led to the discovery of 75 previously unknown connections be
tween constants, including a new formula for the `first continued fraction' cons
tant C_1, novel formulas for natural logarithms, and new formulas connecting $
\pi$ and e.
The latter formulas generalize a century-old relation between π and e by R
amanujan, which until now was considered a singular formula and is now found to
be part of a broader mathematical structure.
The code supporting this library is a public, open-source API that can serve res
earchers in experimental mathematics and other fields of science.
****

Junting Chen,Checheng Yu,Xunzhe Zhou,Tianqi Xu,Yao Mu,Mengkang Hu,Wenqi Shao,Yik
ai Wang,Guohao Li,Lin Shao

EMOS: Embodiment-aware Heterogeneous Multi-robot Operating System with LLM Agent
s

Heterogeneous multi-robot systems (HMRS) have emerged as a powerful ap-
proach for tackling complex tasks that single robots cannot manage alone. Curren

t
large-language-model-based multi-agent systems (LLM-based MAS) have shown
success in areas like software development and operating systems, but applying
these systems to robot control presents unique challenges. In particular, the ca
-
pabilities of each agent in a multi-robot system are inherently tied to the phys
ical
composition of the robots, rather than predefined roles. To address this issue,
we introduce a novel multi-agent framework designed to enable effective collab-
oration among heterogeneous robots with varying embodiments and capabilities,
along with a new benchmark named Habitat-MAS. One of our key designs is
Robot Resume: Instead of adopting human-designed role play, we propose a self-
prompted approach, where agents comprehend robot URDF files and call robot
kinematics tools to generate descriptions of their physics capabilities to guide
their behavior in task planning and action execution. The Habitat-MAS bench-
mark is designed to assess how a multi-agent framework handles tasks that requir
e
embodiment-aware reasoning, which includes 1) manipulation, 2) perception, 3)
navigation, and 4) comprehensive multi-floor object rearrangement. The experi-
mental results indicate that the robot’s resume and the hierarchical design of o
ur
multi-agent system are essential for the effective operation of the heterogeneou
s
multi-robot system within this intricate problem context.
****
Yuancheng Wang,Haoyue Zhan,Liwei Liu,Ruihong Zeng,Haotian Guo,Jiachen Zheng,Qian
g Zhang,Xueyao Zhang,Shunsi Zhang,Zhizheng Wu
MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer
The recent large-scale text-to-speech (TTS) systems are usually grouped as autor
egressive and non-autoregressive systems. The autoregressive systems implicitly
model duration but exhibit certain deficiencies in robustness and lack of durati
on controllability. Non-autoregressive systems require explicit alignment inform
ation between text and speech during training and predict durations for linguist
ic units (e.g. phone), which may compromise their naturalness.
In this paper, we introduce $\textbf{Mask}$ed $\textbf{G}$enerative $\textbf{C}$
odec $\textbf{T}$ransformer (MaskGCT), a fully non-autoregressive TTS model that
 eliminates the need for explicit alignment information between text and speech
supervision, as well as phone-level duration prediction. MaskGCT is a two-stage
model: in the first stage, the model uses text to predict semantic tokens extrac
ted from a speech self-supervised learning (SSL) model, and in the second stage,
 the model predicts acoustic tokens conditioned on these semantic tokens. MaskGC
T follows the mask-and-predict learning paradigm. During training, MaskGCT learn
s to predict masked semantic or acoustic tokens based on given conditions and pr
ompts. During inference, the model generates tokens of a specified length in a p
arallel manner.
Experiments with 100K hours of in-the-wild speech demonstrate that MaskGCT outpe
rforms the current state-of-the-art zero-shot TTS systems in terms of quality, s
imilarity, and intelligibility. Audio samples are available at https://maskgct.g
ithub.io/. We release our code and model checkpoints at https://github.com/open-
mmlab/Amphion/blob/main/models/tts/maskgct.
****
Jiahao Lu,Yifan Zhang,Qiuhong Shen,Xinchao Wang,Shuicheng YAN
Poison-splat: Computation Cost Attack on 3D Gaussian Splatting
3D Gaussian splatting (3DGS), known for its groundbreaking performance and effic
iency, has become a dominant 3D representation and brought progress to many 3D v
ision tasks. However, in this work, we reveal a significant security vulnerabili
ty that has been largely overlooked in 3DGS: the computation cost of training 3D
GS could be maliciously tampered by poisoning the input data. By developing an a
ttack named Poison-splat, we reveal a novel attack surface where the adversary c
an poison the input images to drastically increase the computation memory and ti

me needed for 3DGS training, pushing the algorithm towards its worst computation complexity. In extreme cases, the attack can even consume all allocable memory, leading to a Denial-of-Service (DoS) that disrupts servers, resulting in practical damages to real-world 3DGS service vendors. Such a computation cost attack is achieved by addressing a bi-level optimization problem through three tailored strategies: attack objective approximation, proxy model rendering, and optional constrained optimization. These strategies not only ensure the effectiveness of our attack but also make it difficult to defend with simple defensive measures. We hope the revelation of this novel attack surface can spark attention to this crucial yet overlooked vulnerability of 3DGS systems. Our code is available at https://github.com/jiahaolu97/poison-splat .

****

Ningyuan Li,Haihong E,Tianyu Yao,Tianyi Hu,Yuhan Li,Haoran Luo,Meina Song,Yifan Zhu

INFER: A Neural-symbolic Model For Extrapolation Reasoning on Temporal Knowledge Graph

Temporal Knowledge Graph(TKG) serves as an efficacious way to store dynamic facts in real-world. Extrapolation reasoning on TKGs, which aims at predicting possible future events, has attracted consistent research interest. Recently, some rule-based methods have been proposed, which are considered more interpretable compared with embedding-based methods. Existing rule-based methods apply rules through path matching or subgraph extraction, which falls short in inference ability and suffers from missing facts in TKGs. Besides, during rule application period, these methods consider the standing of facts as a binary 0 or 1 problem and ignores the validity as well as frequency of historical facts under temporal settings.

In this paper, by designing a novel paradigm for rule application, we propose INFER, a neural-symbolic model for TKG extrapolation. With the introduction of Temporal Validity Function, INFER firstly considers the frequency and validity of historical facts and extends the truth value of facts into continuous real number to better adapt for temporal settings. INFER builds Temporal Weight Matrices with a pre-trained static KG embedding model to enhance its inference ability. Moreover to facilitates potential integration with existing embedding-based methods INFER adopts a rule projection module which enables it apply rules through conducting matrices operation on GPU. This feature also improves the efficiency of rule application.

Experimental results show that INFER achieves state-of-the-art performance on three datasets and significantly outperforms existing rule-based models on our modified, more sparse TKG datasets, which demonstrates the superiority of our model in inference ability.

****

Adyasha Maharana,Jaehong Yoon,Tianlong Chen,Mohit Bansal

Adapt-∞: Scalable Continual Multimodal Instruction Tuning via Dynamic Data Selection

Visual instruction datasets from various distributors are released at different times and often contain a significant number of semantically redundant text-image pairs, depending on their task compositions (i.e., skills) or reference sources. This redundancy greatly limits the efficient deployment of continually adaptable multimodal large language models, hindering their ability to refine existing skills and acquire new competencies over time. To address this, we reframe the problem of lifelong Instruction Tuning (LiIT) via data selection, where the model automatically selects beneficial samples to learn from earlier and new datasets based on the current state of acquired knowledge in the model. Based on empirical analyses that show that selecting the best data subset using a static importance measure is often ineffective for multi-task datasets with evolving distributions, we propose Adapt-∞, a new multi-way and adaptive data selection approach that dynamically balances sample efficiency and effectiveness during LiIT. We first construct pseudo-skill clusters by grouping gradient-based sample vectors. Next, we select the best-performing data selector for each skill cluster from a pool of selector experts, including our newly proposed scoring function, I

mage Grounding score. This data selector samples a subset of the most important samples from each skill cluster for training. To prevent the continuous increase in the size of the dataset pool during LIT, which would result in excessive computation, we further introduce a cluster-wise permanent data pruning strategy to remove the most semantically redundant samples from each cluster, keeping computational requirements manageable. We validate the effectiveness and efficiency of Adapt-∞ over a sequence of various multimodal instruction tuning datasets with various tasks, including (Knowledge) VQA, multilingual, grounding, reasoning, language-only, and multi-image comprehension tasks. Training with samples selected by Adapt-∞ alleviates catastrophic forgetting, especially for rare tasks, and promotes forward transfer across the continuum using only a fraction of the original datasets.
****

Fanhu Zeng,Zhen Cheng,Fei Zhu,Hongxin Wei,Xu-Yao Zhang

Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection

Out-of-Distribution (OOD) detection, aiming to distinguish outliers from known categories, has gained prominence in practical scenarios. Recently, the advent of vision-language models (VLM) has heightened interest in enhancing OOD detection for VLM through few-shot tuning. However, existing methods mainly focus on optimizing global prompts, ignoring refined utilization of local information with regard to outliers. Motivated by this, we freeze global prompts and introduce Local-Prompt, a novel coarse-to-fine tuning paradigm to emphasize regional enhancement with local prompts. Our method comprises two integral components: global prompt guided negative augmentation and local prompt enhanced regional regularization. The former utilizes frozen, coarse global prompts as guiding cues to incorporate negative augmentation, thereby leveraging local outlier knowledge. The latter employs trainable local prompts and a regional regularization to capture local information effectively, aiding in outlier identification. We also propose regional-related metric to empower the enrichment of OOD detection. Moreover, since our approach explores enhancing local prompts only, it can be seamlessly integrated with trained global prompts during inference to boost the performance. Comprehensive experiments demonstrate the effectiveness and potential of our method. Notably, our method reduces average FPR95 by 5.17% against state-of-the-art method in 4-shot tuning on challenging ImageNet-1k dataset, even outperforming 16-shot results of previous methods.
****

Yifei Xing,Xiangyuan Lan,Ruiping Wang,Dongmei Jiang,Wenjun Huang,Zheng Qingfang, Yaowei Wang

EMMA: Empowering Multi-modal Mamba with Structural and Hierarchical Alignment

Mamba-based architectures have shown to be a promising new direction for deep learning models owing to their competitive performance and sub-quadratic deployment speed. However, current Mamba multi-modal large language models (MLLM) are insufficient in extracting visual features, leading to imbalanced cross-modal alignment between visual and textural latents, negatively impacting performance on multi-modal tasks. In this work, we propose Empowering Multi-modal Mamba with Structural and Hierarchical Alignment (EMMA), which enables the MLLM to extract fine-grained visual information. Specifically, we propose a pixel-wise alignment module to autoregressively optimize the learning and processing of spatial image-level features along with textual tokens, enabling structural alignment at the image level. In addition, to prevent the degradation of visual information during the cross-model alignment process, we propose a multi-scale feature fusion (MFF) module to combine multi-scale visual features from intermediate layers, enabling hierarchical alignment at the feature level. Extensive experiments are conducted across a variety of multi-modal benchmarks. Our model shows lower latency than other Mamba-based MLLMs and is nearly four times faster than transformer-based MLLMs of similar scale during inference. Due to better cross-modal alignment, our model exhibits lower degrees of hallucination and enhanced sensitivity to visual details, which manifests in superior performance across diverse multi-modal benchmarks. Code provided at https://github.com/xingyifei2016/EMMA.

****
Chongyang Gao,Lixu Wang,Kaize Ding,Chenkai Weng,Xiao Wang,Qi Zhu
On Large Language Model Continual Unlearning
While large language models have demonstrated impressive performance across various domains and tasks, their security issues have become increasingly severe. Machine unlearning has emerged as a representative approach for model safety and security by removing the influence of undesired data on the target model.
However, these methods do not sufficiently consider that unlearning requests in real-world scenarios are continuously emerging, especially in the context of LLMs, which may lead to accumulated model utility loss that eventually becomes unacceptable.
Moreover, existing LLM unlearning methods often ignore previous data access limitations due to privacy concerns and copyright protection. Without previous data, the utility preservation during unlearning is much harder.
To overcome these challenges, we propose the \OOO{} framework that includes an \underline{\textit{O}}rthogonal low-rank adapter (LoRA) for continually unlearning requested data and an \underline{\textit{O}}ut-\underline{\textit{O}}f-Distribution (OOD) detector to measure the similarity between input and unlearning data. The orthogonal LoRA achieves parameter disentanglement among continual unlearning requests. The OOD detector is trained with a novel contrastive entropy loss and utilizes a glocal-aware scoring mechanism. During inference, our \OOO{} framework can decide whether and to what extent to load the unlearning LoRA based on the OOD detector's predicted similarity between the input and the unlearned knowledge. Notably, \OOO{}'s effectiveness does not rely on any retained data. We conducted extensive experiments on \OOO{} and state-of-the-art LLM unlearning methods across three tasks and seven datasets. The results indicate that \OOO{} consistently achieves the best unlearning effectiveness and utility preservation, especially when facing continuous unlearning requests. The source codes can be found at \url{https://github.com/GCYZSL/O3-LLM-UNLEARNING}.
****
Hyun woo Lee,Hayoung Choi,Hyunju Kim
Robust Weight Initialization for Tanh Neural Networks with Fixed Point Analysis
As a neural network's depth increases, it can improve generalization performance. However, training deep networks is challenging due to gradient and signal propagation issues. To address these challenges, extensive theoretical research and various methods have been introduced. Despite these advances, effective weight initialization methods for tanh neural networks remain insufficiently investigated. This paper presents a novel weight initialization method for neural networks with tanh activation function. Based on an analysis of the fixed points of the function $\tanh(ax)$, the proposed method aims to determine values of a that mitigate activation saturation. A series of experiments on various classification datasets and physics-informed neural networks demonstrates that the proposed method outperforms Xavier initialization methods (with or without normalization) in terms of robustness across different network sizes, data efficiency, and convergence speed. Code is available at https://github.com/1HyunwooLee/Tanh-Init.
****
Wei Dai,Jicong Fan
AutoUAD: Hyper-parameter Optimization for Unsupervised Anomaly Detection
Unsupervised anomaly detection (UAD) has important applications in diverse fields such as manufacturing industry and medical diagnosis. In the past decades, although numerous insightful and effective UAD methods have been proposed, it remains a huge challenge to tune the hyper-parameters of each method and select the most appropriate method among many candidates for a specific dataset, due to the absence of labeled anomalies in the training phase of UAD methods and the high diversity of real datasets. In this work, we aim to address this challenge, so as to make UAD more practical and reliable. We propose two internal evaluation metrics, relative-top-median and expected-anomaly-gap, and one semi-internal evaluation metric, normalized pseudo discrepancy (NPD), as surrogate functions of the expected model performance on unseen test data. For instance, NPD measures the discrepancy between the anomaly scores of a validation set drawn from the trainin

g data and a validation set drawn from an isotropic Gaussian. NPD is simple and hyper-parameter-free and is able to compare different UAD methods, and its effectiveness is theoretically analyzed. We integrate the three metrics with Bayesian optimization to effectively optimize the hyper-parameters of UAD models. Extensive experiments on 38 datasets show the effectiveness of our methods.
****
Linh Tran,Wei Sun,Stacy Patterson,Ana Milanova
Privacy-Preserving Personalized Federated Prompt Learning for Multimodal Large Language Models
Multimodal Large Language Models (LLMs) are pivotal in revolutionizing customer support and operations by integrating multiple modalities such as text, images, and audio. Federated Prompt Learning (FPL) is a recently proposed approach that combines pre-trained multimodal LLMs such as vision-language models with federated learning to create personalized, privacy-preserving AI systems. However, balancing the competing goals of personalization, generalization, and privacy remains a significant challenge. Over-personalization can lead to overfitting, reducing generalizability, while stringent privacy measures, such as differential privacy, can hinder both personalization and generalization. In this paper, we propose a Differentially Private Federated Prompt Learning (DP-FPL) approach to tackle this challenge by leveraging a low-rank factorization scheme to capture generalization while maintaining a residual term that preserves expressiveness for personalization. To ensure privacy, we introduce a novel method where we apply local differential privacy to the two low-rank components of the local prompt, and global differential privacy to the global prompt. Our approach mitigates the impact of privacy noise on the model performance while balancing the tradeoff between personalization and generalization. Extensive experiments demonstrate the effectiveness of our approach over other benchmarks.
****
Ao Li,Yuexiang Xie,Songze Li,Fugee Tsung,Bolin Ding,Yaliang Li
Agent-Oriented Planning in Multi-Agent Systems
Through the collaboration of multiple LLM-empowered agents possessing diverse expertise and tools, multi-agent systems achieve impressive progress in solving real-world problems. Given the user queries, the meta-agents, serving as the brain within multi-agent systems, are required to decompose the queries into multiple sub-tasks that can be allocated to suitable agents capable of solving them, so-called agent-oriented planning. In this study, we identify three critical design principles of agent-oriented planning, including solvability, completeness, and non-redundancy, to ensure that each sub-task can be effectively resolved, resulting in satisfactory responses to user queries. These principles further inspire us to propose AOP, a novel framework for agent-oriented planning in multi-agent systems, leveraging a fast task decomposition and allocation process followed by an effective and efficient evaluation via a reward model. According to the evaluation results, the meta-agent is also responsible for promptly making necessary adjustments to sub-tasks and scheduling. Besides, we integrate a feedback loop into AOP to further enhance the effectiveness and robustness of such a problem-solving process. Extensive experiments demonstrate the advancement of AOP in solving real-world problems compared to both single-agent systems and existing planning strategies for multi-agent systems. The source code is available at https://github.com/lalaliat/Agent-Oriented-Planning
****
Lei Yuan,Yuqi Bian,Lihe Li,Ziqian Zhang,Cong Guan,Yang Yu
Efficient Multi-agent Offline Coordination via Diffusion-based Trajectory Stitching
Learning from offline data without interacting with the environment is a promising way to fully leverage the intelligent decision-making capabilities of multi-agent reinforcement learning (MARL). Previous approaches have primarily focused on developing learning techniques, such as conservative methods tailored to MARL using limited offline data. However, these methods often overlook the temporal relationships across different timesteps and spatial relationships between teammates, resulting in low learning efficiency in imbalanced data scenarios. To compr

ehensively explore the data structure of MARL and enhance learning efficiency, w
e propose Multi-Agent offline coordination via Diffusion-based Trajectory Stitch
ing (MADiTS), a novel diffusion-based data augmentation pipeline that systematic
ally generates trajectories by stitching high-quality coordination segments toge
ther. MADiTS first generates trajectory segments using a trained diffusion model
, followed by applying a bidirectional dynamics constraint to ensure that the tr
ajectories align with environmental dynamics. Additionally, we develop an offlin
e credit assignment technique to identify and optimize the behavior of underperf
orming agents in the generated segments. This iterative procedure continues unti
l a satisfactory augmented episode trajectory is generated within the predefined
 limit or is discarded otherwise. Empirical results on imbalanced datasets of mu
ltiple benchmarks demonstrate that MADiTS significantly improves MARL performanc
e.
****

Wenxuan Bao,Zhichen Zeng,Zhining Liu,Hanghang Tong,Jingrui He
Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation
Powerful as they are, graph neural networks (GNNs) are known to be vulnerable to
 distribution shifts. Recently, test-time adaptation (TTA) has attracted attenti
on due to its ability to adapt a pre-trained model to a target domain, without r
e-accessing the source domain. However, existing TTA algorithms are primarily de
signed for attribute shifts in vision tasks, where samples are independent. Thes
e methods perform poorly on graph data that experience structure shifts, where n
ode connectivity differs between source and target graphs. We attribute this per
formance gap to the distinct impact of node attribute shifts versus graph struct
ure shifts: the latter significantly degrades the quality of node representation
s and blurs the boundaries between different node categories. To address structu
re shifts in graphs, we propose Matcha, an innovative framework designed for eff
ective and efficient adaptation to structure shifts by adjusting the htop-aggreg
ation parameters in GNNs. To enhance the representation quality, we design a pre
diction-informed clustering loss to encourage the formation of distinct clusters
 for different node categories. Additionally, Matcha seamlessly integrates with
existing TTA algorithms, allowing it to handle attribute shifts effectively whil
e improving overall performance under combined structure and attribute shifts. W
e validate the effectiveness of Matcha on both synthetic and real-world datasets
, demonstrating its robustness across various combinations of structure and attr
ibute shifts. Our code
is available at https://github.com/baowenxuan/Matcha.
****

Lijie Yang,Zhihao Zhang,Zhuofu Chen,Zikun Li,Zhihao Jia
TidalDecode: Fast and Accurate LLM Decoding with Position Persistent Sparse Atte
ntion
Large language models (LLMs) have driven significant advancements across diverse
 NLP tasks, with long-context models gaining prominence for handling extended in
puts. However, the expanding key-value (KV) cache size required by Transformer a
rchitectures intensifies the memory constraints, particularly during the decodin
g phase, creating a significant bottleneck. Existing sparse attention mechanisms
 designed to address this bottleneck have two limitations: (1) they often fail t
o reliably identify the most relevant tokens for attention, and (2) they overloo
k the spatial coherence of token selection across consecutive Transformer layers
, which can lead to performance degradation and substantial overhead in token se
lection. This paper introduces TidalDecode, a simple yet effective algorithm and
 system for fast and accurate LLM decoding through position persistent sparse at
tention. TidalDecode leverages the spatial coherence of tokens selected by exist
ing sparse attention methods and introduces a few token selection layers that pe
rform full attention to identify the tokens with the highest attention scores, w
hile all other layers perform sparse attention with the pre-selected tokens. Thi
s design enables TidalDecode to substantially reduce the overhead of token selec
tion for sparse attention without sacrificing the quality of the generated resul
ts. Evaluation on a diverse set of LLMs and tasks shows that TidalDecode closely
 matches the generative performance of full attention methods while reducing the

LLM decoding latency by up to $2.1\times$.
****
Mrinal Mathur,Barak A. Pearlmutter,Sergey M. Plis
MIND over Body: Adaptive Thinking using Dynamic Computation
While the human brain efficiently handles various computations with a limited number of neurons, traditional deep learning networks require a significant increase in parameters to improve performance.
Yet, these parameters are used inefficiently as the networks employ the same amount of computation for inputs of the same size, regardless of the input's complexity.
We address this inefficiency by introducing self-introspection capabilities to the network, enabling it to adjust the number of used parameters based on the internal representation of the task and adapt the computation time based on the task complexity.
This enables the network to adaptively reuse parameters across tasks, dynamically adjusting the computational effort to match the complexity of the input.
We demonstrate the effectiveness of this method on language modeling and computer vision tasks.
Notably, our model achieves 96.62\% accuracy on ImageNet with just a three-layer network, surpassing much larger ResNet-50 and EfficientNet. When applied to a transformer architecture, the approach achieves 95.8\%/88.7\% F1 scores on the SQuAD v1.1/v2.0 datasets at negligible parameter cost.
These results showcase the potential for dynamic and reflective computation, contributing to the creation of intelligent systems that efficiently manage resources based on input data complexity.
****
Dhananjaya Jayasundara,Heng Zhao,Demetrio Labate,Vishal M. Patel
PIN: Prolate Spheroidal Wave Function-based Implicit Neural Representations
Implicit Neural Representations (INRs) provide a continuous mapping between the coordinates of a signal and the corresponding values. As the performance of INRs heavily depends on the choice of nonlinear-activation functions, there has been a significant focus on encoding explicit signals within INRs using diverse activation functions. Despite recent advancements, existing INRs often encounter significant challenges, particularly at fine scales where they often introduce noise-like artifacts over smoother areas compromising the quality of the output. Moreover, they frequently struggle to generalize to unseen coordinates. These drawbacks highlight a critical area for further research and development to enhance the robustness and applicability of INRs across diverse scenarios. To address this challenge, we introduce the Prolate Spheroidal Wave Function-based Implicit Neural Representations (PIN), which exploits the optimal space-frequency domain concentration of Prolate Spheroidal Wave Functions (PSWFs) as the nonlinear mechanism in INRs. Our experimental results reveal that PIN excels not only in representing images and 3D shapes but also significantly outperforms existing methods in various vision tasks that require INR generalization, including image inpainting, novel view synthesis, edge detection, and image denoising.
****
Wenda Xu,Rujun Han,Zifeng Wang,Long Le,Dhruv Madeka,Lei Li,William Yang Wang,Rishabh Agarwal,Chen-Yu Lee,Tomas Pfister
Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling
Recent advances in knowledge distillation (KD) have enabled smaller student models to approach the performance of larger teacher models. However, popular methods such as supervised KD and on-policy KD, are adversely impacted by the knowledge gaps between teacher-student in practical scenarios. Supervised KD suffers from a distribution mismatch between training with a static dataset and inference over final student-generated outputs. Conversely, on-policy KD, which uses student-generated samples for training, can suffer from low-quality training examples with which teacher models are not familiar, resulting in inaccurate teacher feedback. To address these limitations, we introduce Speculative Knowledge Distillation (SKD), a novel approach that leverages cooperation between student and teach

er models to generate high-quality training data on-the-fly while aligning with the student's inference-time distribution. In SKD, the student proposes tokens, and the teacher replaces poorly ranked ones based on its own distribution, trans ferring high-quality knowledge adaptively. We evaluate SKD on various text gener ation tasks, including translation, summarization, math, and instruction followi ng, and show that SKD consistently outperforms existing KD methods across differ ent domains, data sizes, and model initialization strategies.
****

Huanjian Zhou,Baoxiang Wang,Masashi Sugiyama

The adaptive complexity of parallelized log-concave sampling

In large-data applications, such as the inference process of diffusion models, i t is desirable to design sampling algorithms with a high degree of parallelizati on. In this work, we study the adaptive complexity of sampling, which is the min imum number of sequential rounds required to achieve sampling given polynomially many queries executed in parallel at each round. For unconstrained sampling, we examine distributions that are log-smooth or log-Lipschitz and log strongly or non-strongly concave. We show that an almost linear iteration algorithm cannot r eturn a sample with a specific exponentially small error under total variation d istance. For box-constrained sampling, we show that an almost linear iteration a lgorithm cannot return a sample with sup-polynomially small error under total va riation distance for log-concave distributions. Our proof relies upon novel anal ysis with the characterization of the output for the hardness potentials based o n the chain-like structure with random partition and classical smoothing techniq ues.
****

Wooseong Jeong,Kuk-Jin Yoon

Selective Task Group Updates for Multi-Task Optimization

Multi-task learning enables the acquisition of task-generic knowledge by trainin g multiple tasks within a unified architecture. However, training all tasks toge ther in a single architecture can lead to performance degradation, known as nega tive transfer, which is a main concern in multi-task learning. Previous works ha ve addressed this issue by optimizing the multi-task network through gradient ma nipulation or weighted loss adjustments. However, their optimization strategy fo cuses on addressing task imbalance in shared parameters, neglecting the learning of task-specific parameters. As a result, they show limitations in mitigating n egative transfer, since the learning of shared space and task-specific informati on influences each other during optimization. To address this, we propose a diff erent approach to enhance multi-task performance by selectively grouping tasks a nd updating them for each batch during optimization. We introduce an algorithm t hat adaptively determines how to effectively group tasks and update them during the learning process. To track inter-task relations and optimize multi-task netw orks simultaneously, we propose proximal inter-task affinity, which can be measu red during the optimization process. We provide a theoretical analysis on how di viding tasks into multiple groups and updating them sequentially significantly a ffects multi-task performance by enhancing the learning of task-specific paramet ers. Our methods substantially outperform previous multi-task optimization appro aches and are scalable to different architectures and various numbers of tasks.
****

Yucheng Suo,Fan Ma,Kaixin Shen,Linchao Zhu,Yi Yang

Long-horizon Visual Instruction Generation with Logic and Attribute Self-reflect ion

Visual instructions for long-horizon tasks are crucial as they intuitively clari fy complex concepts and enhance retention across extended steps.

Directly generating a series of images using text-to-image models without consid ering the context of previous steps results in inconsistent images, increasing c ognitive load. Additionally, the generated images often miss objects or the att ributes such as color, shape, and state of the objects are inaccurate.

To address these challenges, we propose LIGER, the first training-free framework for Long-horizon Instruction GEneration with logic and attribute self-Reflectio n. LIGER first generates a draft image for each step with the historical prompt

and visual memory of previous steps. This step-by-step generation approach maint
ains consistency between images in long-horizon tasks. Moreover, LIGER utilizes
various image editing tools to rectify errors including wrong attributes, logic
errors, object redundancy, and identity inconsistency in the draft images. Throu
gh this self-reflection mechanism, LIGER improves the logic and object attribute
 correctness of the images.
To verify whether the generated images assist human understanding, we manually c
urated a new benchmark consisting of various long-horizon tasks. Human-annotated
 ground truth expressions reflect the human-defined criteria for how an image sh
ould appear to be illustrative.
Experiments demonstrate the visual instructions generated by LIGER are more comp
rehensive compared with baseline methods. The code and dataset will be available
 once accepted.
****
Floriano Tori,Vincent Holst,Vincent Ginis
The Effectiveness of Curvature-Based Rewiring and the Role of Hyperparameters in
 GNNs Revisited
Message passing is the dominant paradigm in Graph Neural Networks (GNNs). The ef
ficiency of message passing, however, can be limited by the topology of the grap
h. This happens when information is lost during propagation due to being oversqu
ashed when travelling through bottlenecks. To remedy this, recent efforts have f
ocused on graph rewiring techniques, which disconnect the input graph originatin
g from the data and the computational graph, on which message passing is perform
ed. A prominent approach for this is to use discrete graph curvature measures, o
f which several variants have been proposed, to identify and rewire around bottl
enecks, facilitating information propagation. While oversquashing has been demon
strated in synthetic datasets, in this work we reevaluate the performance gains
that curvature-based rewiring brings to real-world datasets. We show that in the
se datasets, edges selected during the rewiring process are not in line with the
oretical criteria identifying bottlenecks. This implies they do not necessarily
oversquash information during message passing. Subsequently, we demonstrate that
 SOTA accuracies on these datasets are outliers originating from sweeps of hyper
parameters—both the ones for training and dedicated ones related to the rewiring
 algorithm—instead of consistent performance gains. In conclusion, our analysis
nuances the effectiveness of curvature-based rewiring in real-world datasets and
 brings a new perspective on the methods to evaluate GNN accuracy improvements.
****
Biao Yi,Tiansheng Huang,Sishuo Chen,Tong Li,Zheli Liu,Zhixuan Chu,Yiming Li
Probe before You Talk: Towards Black-box Defense against Backdoor Unalignment fo
r Large Language Models
Backdoor unalignment attacks against Large Language Models (LLMs) enable the ste
althy compromise of safety alignment using a hidden trigger while evading normal
 safety auditing. These attacks pose significant threats to the applications of
LLMs in the real-world Large Language Model as a Service (LLMaaS) setting, where
 the deployed model is a fully black-box system that can only interact through t
ext. Furthermore, the sample-dependent nature of the attack target exacerbates t
he threat. Instead of outputting a fixed label, the backdoored LLM follows the s
emantics of any malicious command with the hidden trigger, significantly expandi
ng the target space. In this paper, we introduce BEAT, a black-box defense that
detects triggered samples during inference to deactivate the backdoor. It is mot
ivated by an intriguing observation (dubbed the **probe concatenate effect**), w
here concatenated triggered samples significantly reduce the refusal rate of the
 backdoored LLM towards a malicious probe, while non-triggered samples have litt
le effect. Specifically, BEAT identifies whether an input is triggered by measur
ing the degree of distortion in the output distribution of the probe before and
after concatenation with the input. Our method addresses the challenges of sampl
e-dependent targets from an opposite perspective. It captures the impact of the
trigger on the refusal signal (which is sample-independent) instead of sample-sp
ecific successful attack behaviors. It overcomes black-box access limitations by
 using multiple sampling to approximate the output distribution.

Extensive experiments are conducted on various backdoor attacks and LLMs (including the closed-source GPT-3.5-turbo), verifying the effectiveness and efficiency of our defense.
Besides, we also preliminarily verify that BEAT can effectively defend against popular jailbreak attacks, as they can be regarded as "natural backdoors". Our source code is available at https://github.com/clearloveclearlove/BEAT.
****

Peng Gao,Le Zhuo,Dongyang Liu,Ruoyi Du,Xu Luo,Longtian Qiu,Yuhang Zhang,Rongjie Huang,Shijie Geng,Renrui Zhang,Junlin Xie,Wenqi Shao,Zhengkai Jiang,Tianshuo Yang,Weicai Ye,Tong He,Jingwen He,Junjun He,Yu Qiao,Hongsheng Li
Lumina-T2X: Scalable Flow-based Large Diffusion Transformer for Flexible Resolution Generation
Sora unveils the potential of scaling Diffusion Transformer (DiT) for generating photorealistic images and videos at arbitrary resolutions, aspect ratios, and durations, yet it still lacks sufficient implementation details. In this paper, we introduce the Lumina-T2X family -- a series of Flow-based Large Diffusion Transformers (Flag-DiT) equipped with zero-initialized attention, as a simple and scalable generative framework that can be adapted to various modalities, e.g., transforming noise into images, videos, multi-view 3D objects, or audio clips conditioned on text instructions. By tokenizing the latent spatial-temporal space and incorporating learnable placeholders such as |[nextline]| and |[nextframe]| tokens, Lumina-T2X seamlessly unifies the representations of different modalities across various spatial-temporal resolutions. Advanced techniques like RoPE, KQ-Norm, and flow matching enhance the stability, flexibility, and scalability of Flag-DiT, enabling models of Lumina-T2X to scale up to 7 billion parameters and extend the context window to 128K tokens. This is particularly beneficial for creating ultra-high-definition images with our Lumina-T2I model and long 720p videos with our Lumina-T2V model. Remarkably, Lumina-T2I, powered by a 5-billion-parameter Flag-DiT, requires only 35% of the training computational costs of a 600-million-parameter naive DiT (PixArt-alpha), indicating that increasing the number of parameters significantly accelerates convergence of generative models without compromising visual quality. Our further comprehensive analysis underscores Lumina-T2X's preliminary capability in resolution extrapolation, high-resolution editing, generating consistent 3D views, and synthesizing videos with seamless transitions. All code and checkpoints of Lumina-T2X are released at https://github.com/Alpha-VLLM/Lumina-T2X to further foster creativity, transparency, and diversity in the generative AI community.
****

Phat Tan Nguyen,Tsun-Hsuan Wang,Zhang-Wei Hong,Erfan Aasi,Andrew Silva,Guy Rosman,Sertac Karaman,Daniela Rus
ReGen: Generative Robot Simulation via Inverse Design
Simulation plays a key role in scaling robot learning and validating policies, but constructing simulations remains labor-intensive. In this paper, we introduce ReGen, a generative simulation framework that automates this process using inverse design. Given an agent's behavior (such as a motion trajectory or objective function) and its textual description, we infer the underlying scenarios and environments that could have caused the behavior.
Our approach leverages large language models to construct and expand a graph that captures cause-and-effect relationships and relevant entities with properties in the environment, which is then processed to configure a robot simulation environment. Our approach supports (i) augmenting simulations based on ego-agent behaviors, (ii) controllable, counterfactual scenario generation, (iii) reasoning about agent cognition and mental states, and (iv) reasoning with distinct sensing modalities, such as braking due to faulty GPS signals.
We demonstrate our method in autonomous driving and robot manipulation tasks, generating more diverse, complex simulated environments compared to existing simulations with high success rates, and enabling controllable generation for corner cases. This approach enhances the validation of robot policies and supports data or simulation augmentation, advancing scalable robot learning for improved generalization and robustness.

****
Ahana Deb,Roberto Cipollone,Anders Jonsson,Alessandro Ronca,Mohammad Sadegh Tale bi
Offline RL in Regular Decision Processes: Sample Efficiency via Language Metrics
This work studies offline Reinforcement Learning (RL) in a class of non-Markovia n environments called Regular Decision Processes (RDPs). In RDPs, the unknown de pendency of future observations and rewards from the past interactions can be ca ptured by some hidden finite-state automaton. For this reason, many RDP algorith ms first reconstruct this unknown dependency using automata learning techniques. In this paper, we consider episodic RDPs and show that it is possible to overco me the limitations of existing offline RL algorithms for RDPs via
the introduction of two original techniques: a novel metric grounded in formal l anguage theory and an approach based on Count-Min-Sketch (CMS). Owing to the nov el language metric, our algorithm is proven to be more sample efficient than exi sting results, and in some problem instances admitting low complexity languages, the gain is showcased to be exponential in the episode length. The CMS-based ap proach removes the need for naïve counting and alleviates the memory requirement s for long planning horizons. We derive Probably Approximately Correct (PAC) sam ple complexity bounds associated to each of these techniques, and validate the a pproach experimentally.
****
Yin Xiangjun,Huihui Yue
Reveal Object in Lensless Photography via Region Gaze and Amplification
Detecting concealed objects, such as in vivo lesions or camouflage, requires cus tomized imaging systems. Lensless cameras, being compact and flexible, offer a p romising alternative to bulky lens systems. However, the absence of lenses leads to measurements lacking visual semantics, posing significant challenges for con cealed object detection (COD). To tackle this issue, we propose a region gaze-am plification network (RGANet) for progressively exploiting concealed objects from lensless imaging measurements. Specifically, a region gaze module (RGM) is prop osed to mine spatial-frequency cues informed by biological and psychological mec hanisms, and a region amplifier (RA) is designed to amplify the details of objec t regions to enhance COD performance. Furthermore, we contribute the first relev ant dataset as a benchmark to prosper the lensless imaging community. Extensive experiments demonstrate the exciting performance of our method. Our codes will b e released in \url{https://github.com/YXJ-NTU/Lensless-COD}.
****
Zhiqiang Wu,Yingjie Liu,Hanlin Dong,Xuan Tang,Jian Yang,Bo Jin,Mingsong Chen,Xia n Wei
R2Det: Exploring Relaxed Rotation Equivariance in 2D Object Detection
Group Equivariant Convolution (GConv) empowers models to explore underlying symm etry in data, improving performance. However, real-world scenarios often deviate from ideal symmetric systems caused by physical permutation, characterized by n on-trivial actions of a symmetry group, resulting in asymmetries that affect the outputs, a phenomenon known as Symmetry Breaking. Traditional GConv-based metho ds are constrained by rigid operational rules within group space, assuming data remains strictly symmetry after limited group transformations. This limitation m akes it difficult to adapt to Symmetry-Breaking and non-rigid transformations. M otivated by this, we mainly focus on a common scenario: Rotational Symmetry-Brea king. By relaxing strict group transformations within Strict Rotation-Equivarian t group $\mathbf{C}_n$, we redefine a Relaxed Rotation-Equivariant group $\mathb f{R}_n$ and introduce a novel Relaxed Rotation-Equivariant GConv (R2GConv) with only a minimal increase of $4n$ parameters compared to GConv. Based on R2GConv, we propose a Relaxed Rotation-Equivariant Network (R2Net) as the backbone and de velop a Relaxed Rotation-Equivariant Object Detector (R2Det) for 2D object detec tion. Experimental results demonstrate the effectiveness of the proposed R2GConv in natural image classification, and R2Det achieves excellent performance in 2D object detection with improved generalization capabilities and robustness. The code is available in \texttt{https://github.com/wuer5/r2det}.
****

Yao Tong,Jiayuan Ye,Sajjad Zarifzadeh,Reza Shokri
How much of my dataset did you use? Quantitative Data Usage Inference in Machine Learning
How much of my data was used to train a machine learning model? This is a critical question for data owners assessing the risk of unauthorized usage of their data to train models. However, previous work mistakenly treats this as a binary problem—inferring whether all-or-none or any-or-none of the data was used—which is fragile when faced with real, non-binary data usage risks. To address this, we propose a fine-grained analysis called Dataset Usage Cardinality Inference (DUCI), which estimates the exact proportion of data used. Our algorithm, leveraging debiased membership guesses, matches the performance of the optimal MLE approach (with a maximum error <0.1) but with significantly lower (e.g., $300 \times$ less) computational cost.
****
Saman Kazemkhani,Aarav Pandya,Daphne Cornelisse,Brennan Shacklett,Eugene Vinitsky
GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS
Multi-agent learning algorithms have been successful at generating superhuman planning in various games but have had limited impact on the design of deployed multi-agent planners. A key bottleneck in applying these techniques to multi-agent planning is that they require billions of steps of experience. To enable the study of multi-agent planning at scale, we present GPUDrive, a GPU-accelerated, multi-agent simulator built on top of the Madrona Game Engine capable of generating over a million simulation steps per second. Observation, reward, and dynamics functions are written directly in C++, allowing users to define complex, heterogeneous agent behaviors that are lowered to high-performance CUDA. Despite these low-level optimizations, GPUDrive is fully accessible through Python, offering a seamless and efficient workflow for multi-agent, closed-loop simulation. Using GPUDrive, we train reinforcement learning agents on the Waymo Open Motion Dataset, achieving efficient goal-reaching in minutes and scaling to thousands of scenarios in hours. We open-source the code and pre-trained agents at \url{www.github.com/Emerge-Lab/gpudrive}.
****
Jingyu Zhang,Ahmed Elgohary,Ahmed Magooda,Daniel Khashabi,Benjamin Van Durme
Controllable Safety Alignment: Inference-Time Adaptation to Diverse Safety Requirements
The current paradigm for safety alignment of large language models (LLMs) follows a _one-size-fits-all_ approach: the model refuses to interact with any content deemed unsafe by the model provider. This approach lacks flexibility in the face of varying social norms across cultures and regions. In addition, users may have diverse safety needs, making a model with _static_ safety standards too restrictive to be useful, as well as too costly to be re-aligned.

We propose _Controllable Safety Alignment_ (CoSA), a framework designed to adapt models to diverse safety requirements without re-training. Instead of aligning a fixed model, we align models to follow _safety configs_—free-form natural language descriptions of the desired safety behaviors—that are provided as part of the system prompt. To adjust model safety behavior, authorized users only need to modify such safety configs at inference time. To enable that, we propose CoSAlign, a data-centric method for aligning LLMs to easily adapt to diverse safety configs. Furthermore, we devise a novel controllability evaluation protocol that considers both helpfulness and configured safety, summarizing them into CoSA-Score, and construct CoSApien, a _human-authored_ benchmark that consists of real-world LLM use cases with diverse safety requirements and corresponding evaluation prompts. We show that CoSAlign leads to substantial gains of controllability over strong baselines including in-context alignment. Our framework encourages better representation and adaptation to pluralistic human values in LLMs, and thereby increasing their practicality.
****
Ziran Qin,Yuchen Cao,Mingbao Lin,Wen Hu,Shixuan Fan,Ke Cheng,Weiyao Lin,Jianguo

Li
CAKE: Cascading and Adaptive KV Cache Eviction with Layer Preferences
Large language models (LLMs) excel at processing long sequences, boosting demand for key-value (KV) caching. While recent efforts to evict KV cache have alleviated the inference burden, they often fail to allocate resources rationally across layers with different attention patterns. In this paper, we introduce Cascading and Adaptive KV cache Eviction (CAKE), a novel approach that frames KV cache eviction as a ``cake-slicing problem.''
CAKE assesses layer-specific preferences by considering attention dynamics in both spatial and temporal dimensions, allocates rational cache size for layers accordingly, and manages memory constraints in a cascading manner. This approach enables a global view of cache allocation, adaptively distributing resources across diverse attention mechanisms while maintaining memory budgets.
CAKE also employs a new eviction indicator that considers the shifting importance of tokens over time, addressing limitations in existing methods that overlook temporal dynamics.
Comprehensive experiments on LongBench and NeedleBench show that CAKE maintains model performance with only 3.2\% of the KV cache and consistently outperforms current baselines across various models and memory constraints, particularly in low-memory settings. Additionally, CAKE achieves over 10$\times$ speedup in decoding latency compared to full cache when processing contexts of 128K tokens with FlashAttention-2. Our code is available at https://github.com/antgroup/cakekv.
****
Huanjian Zhou,Lingxiao Huang,Baoxiang Wang
Improved Approximation Algorithms for k-Submodular Maximization via Multilinear Extension
We investigate a generalized form of submodular maximization, referred to as k-submodular maximization, with applications across the domains of social networks and machine learning. In this work, we propose the multilinear extension of k-submodular functions and unified Frank-Wolfe-type frameworks based on that. This continuous framework accommodates 1) monotone or non-monotone functions, and 2) various constraint types including matroid constraints, knapsack constraints, and their combinations. Notably, we attain an asymptotically optimal $1/2$-approximation for monotone k-submodular maximization problems with knapsack constraints, surpassing previous $1/3$-approximation results, and a factor-$1/3$ approximation for non-monotone k-submodular maximization problems with knapsack constraints and matroid constraints which outperforms previous 0.245-approximation results. The foundation for our analysis stems from new insights into specific linear and monotone properties pertaining to the multilinear extension.
****
Yongxing Zhang,Donglin Yang,Renjie Liao
SymmetricDiffusers: Learning Discrete Diffusion on Finite Symmetric Groups
The group of permutations S_n, also known as the finite symmetric groups, are essential in fields such as combinatorics, physics, and chemistry. However, learning a probability distribution over S_n poses significant challenges due to its intractable size and discrete nature. In this paper, we introduce *SymmetricDiffusers*, a novel discrete diffusion model that simplifies the task of learning a complicated distribution over S_n by decomposing it into learning simpler transitions of the reverse diffusion using deep neural networks. We identify the riffle shuffle as an effective forward transition and provide empirical guidelines for selecting the diffusion length based on the theory of random walks on finite groups. Additionally, we propose a generalized Plackett-Luce (PL) distribution for the reverse transition, which is provably more expressive than the PL distribution. We further introduce a theoretically grounded "denoising schedule" to improve sampling and learning efficiency. Extensive experiments show that our model achieves state-of-the-art or comparable performance on solving tasks including sorting 4-digit MNIST images, jigsaw puzzles, and traveling salesman problems. Our code is released at <https://github.com/DSL-Lab/SymmetricDiffusers>.
****
Seungjun Lee,Yuyang Zhao,Gim Hee Lee

Segment Any 3D Object with Language
In this paper, we investigate Open-Vocabulary 3D Instance Segmentation (OV-3DIS) with free-form language instructions. Earlier works mainly rely on annotated base categories for training which leads to limited generalization to unseen novel categories. To mitigate the poor generalizability to novel categories, recent works generate class-agnostic masks or projecting generalized masks from 2D to 3D, subsequently classifying them with the assistance of 2D foundation model. However, these works often disregard semantic information in the mask generation, leading to sub-optimal performance. Instead, generating generalizable but semantic-aware masks directly from 3D point clouds would result in superior outcomes. To the end, we introduce Segment any 3D Object with LanguagE ($\textbf{SOLE}$), which is a semantic and geometric-aware visual-language learning framework with strong generalizability by generating semantic-related masks directly from 3D point clouds. Specifically, we propose a multimodal fusion network to incorporate multimodal semantics in both backbone and decoder. In addition, to align the 3D segmentation model with various language instructions and enhance the mask quality, we introduce three types of multimodal associations as supervision. Our SOLE outperforms previous methods by a large margin on ScanNetv2, ScanNet200, and Replica benchmarks, and the results are even closed to the fully-supervised counterpart despite the absence of class annotations in the training. Furthermore, extensive qualitative results demonstrate the versatility of our SOLE to language instructions. The code will be made publicly available.
****
Yuanchen Wu,Junlong Du,Ke Yan,Shouhong Ding,Xiaoqiang Li
ToVE: Efficient Vision-Language Learning via Knowledge Transfer from Vision Experts
Vision-language (VL) learning requires extensive visual perception capabilities, such as fine-grained object recognition and spatial perception. Recent works typically rely on training huge models on massive datasets to develop these capabilities. As a more efficient alternative, this paper proposes a new framework that Transfers the knowledge from a hub of Vision Experts (ToVE) for efficient VL learning, leveraging pre-trained vision expert models to promote visual perception capability. Specifically, building on a frozen CLIP image encoder that provides vision tokens for image-conditioned language generation, ToVE introduces a hub of multiple vision experts and a token-aware gating network that dynamically routes expert knowledge to vision tokens. In the transfer phase, we propose a "residual knowledge transfer" strategy, which not only preserves the generalizability of the vision tokens but also allows selective detachment of low-contributing experts to improve inference efficiency. Further, we explore to merge these expert knowledge to a single CLIP encoder, creating a knowledge-merged CLIP that produces more informative vision tokens without expert inference during deployment. Experiment results across various VL tasks demonstrate that the proposed ToVE achieves competitive performance with two orders of magnitude fewer training data.
****
Xin Li,Deshui Miao,Zhenyu He,Yaowei Wang,Huchuan Lu,Ming-Hsuan Yang
Learning Spatial-Semantic Features for Robust Video Object Segmentation
Tracking and segmenting multiple similar objects with distinct or complex parts in long-term videos is particularly challenging due to the ambiguity in identifying target components and the confusion caused by occlusion, background clutter, and changes in appearance or environment over time. In this paper, we propose a robust video object segmentation framework that learns spatial-semantic features and discriminative object queries to address the above issues. Specifically, we construct a spatial-semantic block comprising a semantic embedding component and a spatial dependency modeling part for associating global semantic features and local spatial features, providing a comprehensive target representation. In addition, we develop a masked cross-attention module to generate object queries that focus on the most discriminative parts of target objects during query propagation, alleviating noise accumulation to ensure effective long-term query propagation. The experimental results show that the proposed method sets new state-of-

the-art performance on multiple data sets, including the DAVIS2017 test (\textbf{87.8\%}), YoutubeVOS 2019 (\textbf{88.1\%}), MOSE val (\textbf{74.0\%}), and LVOS test (\textbf{73.0\%}), which demonstrate the effectiveness and generalization capacity of the proposed method. We will make all the source code and trained models publicly available.
****
Heming Xia,Yongqi Li,Jun Zhang,Cunxiao Du,Wenjie Li
SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration
Speculative decoding (SD) has emerged as a widely used paradigm to accelerate LLM inference without compromising quality. It works by first employing a compact model to draft multiple tokens efficiently and then using the target LLM to verify them in parallel. While this technique has achieved notable speedups, most existing approaches necessitate either additional parameters or extensive training to construct effective draft models, thereby restricting their applicability across different LLMs and tasks. To address this limitation, we explore a novel plug-and-play SD solution with layer-skipping, which skips intermediate layers of the target LLM as the compact draft model. Our analysis reveals that LLMs exhibit great potential for self-acceleration through layer sparsity and the task-specific nature of this sparsity. Building on these insights, we introduce SWIFT, an on-the-fly self-speculative decoding algorithm that adaptively selects intermediate layers of LLMs to skip during inference. SWIFT does not require auxiliary models or additional training, making it a plug-and-play solution for accelerating LLM inference across diverse input data streams. Our extensive experiments across a wide range of models and downstream tasks demonstrate that SWIFT can achieve over a $1.3\times$$\sim$$1.6\times$ speedup while preserving the original distribution of the generated text. We release our code in https://github.com/hemingkx/SWIFT.
****
Hyeonsu Jeong,Hye Won Chung
Rethinking Self-Distillation: Label Averaging and Enhanced Soft Label Refinement with Partial Labels
We investigate the mechanisms of self-distillation in multi-class classification, particularly in the context of linear probing with fixed feature extractors where traditional feature learning explanations do not apply. Our theoretical analysis reveals that multi-round self-distillation effectively performs label averaging among instances with high feature correlations, governed by the eigenvectors of the Gram matrix derived from input features. This process leads to clustered predictions and improved generalization, mitigating the impact of label noise by reducing the model's reliance on potentially corrupted labels. We establish conditions under which multi-round self-distillation achieves 100\% population accuracy despite label noise. Furthermore, we introduce a novel, efficient single-round self-distillation method using refined partial labels from the teacher's top two softmax outputs, referred to as the PLL student model. This approach replicates the benefits of multi-round distillation in a single round, achieving comparable or superior performance--especially in high-noise scenarios--while significantly reducing computational cost.
****
Yinqi Bai,Jie Wang,Lei Chen,Zhihai Wang,Yufei Kuang,Mingxuan Yuan,Jianye HAO,Feng Wu
A Graph Enhanced Symbolic Discovery Framework For Efficient Logic Optimization
The efficiency of Logic Optimization (LO) has become one of the key bottlenecks in chip design. To prompt efficient LO, previous studies propose using a key scoring function to predict and prune a large number of ineffective nodes of the LO heuristics. However, the existing scoring functions struggle to balance inference efficiency, interpretability, and generalization performance, which severely hinders their application to modern LO tools. To address this challenge, we propose a novel data-driven circuit symbolic learning framework, namely CMO, to learn lightweight, interpretable, and generalizable scoring functions. The major challenge of developing CMO is to discover symbolic functions that can well generalize to unseen circuits, i.e., the circuit symbolic generalization problem. Thus,

the major technical contribution of CMO is the novel Graph Enhanced Symbolic Di
scovery framework, which distills dark knowledge from a well-designed Graph Neur
al Network (GNN) to enhance the generalization capability of the learned symboli
c functions. To the best of our knowledge, CMO is *the first* graph-enhanced app
roach for discovering lightweight and interpretable symbolic functions that can
well generalize to unseen circuits in LO. Experiments on three challenging circu
it benchmarks show that the *interpretable* symbolic functions learned by CMO ou
tperform previous state-of-the-art (SOTA) GPU-based and human-designed approache
s in terms of *inference efficiency* and *generalization capability*. Moreover,
we integrate CMO with the Mfs2 heuristic---one of the most time-consuming LO heu
ristics. The empirical results demonstrate that CMO significantly improves its e
fficiency while keeping comparable optimization performance when executed on a C
PU-based machine, achieving up to 2.5× faster runtime.
****
Zehao Xiao,Shilin Yan,Jack Hong,Jiayin Cai,Xiaolong Jiang,Yao Hu,Jiayi Shen,Chee
ms Wang,Cees G. M. Snoek
DynaPrompt: Dynamic Test-Time Prompt Tuning
Test-time prompt tuning enhances zero-shot generalization of vision-language mod
els but tends to ignore the relatedness among test samples during inference. Onl
ine test-time prompt tuning provides a simple way to leverage the information in
previous test samples, albeit with the risk of prompt collapse due to error acc
umulation. To enhance test-time prompt tuning, we propose DynaPrompt, short for
dynamic test-time prompt tuning, exploiting relevant data distribution informati
on while reducing error accumulation. Built on an online prompt buffer, DynaProm
pt adaptively selects and optimizes the relevant prompts for each test sample du
ring tuning. Specifically, we introduce a dynamic prompt selection strategy base
d on two metrics: prediction entropy and probability difference. For unseen test
data information, we develop dynamic prompt appending, which allows the buffer
to append new prompts and delete the inactive ones. By doing so, the prompts are
optimized to exploit beneficial information on specific test data, while allevi
ating error accumulation. Experiments on fourteen datasets demonstrate the effec
tiveness of dynamic test-time prompt tuning.
****
Yiheng Xu,Dunjie Lu,Zhennan Shen,Junli Wang,Zekun Wang,Yuchen Mao,Caiming Xiong,
Tao Yu
AgentTrek: Agent Trajectory Synthesis via Guiding Replay with Web Tutorials
Graphical User Interface (GUI) agents hold great potential for automating comple
x tasks across diverse digital environments, from web applications to desktop so
ftware. However, the development of such agents is hindered by the lack of high-
quality, multi-step trajectory data required for effective training. Existing ap
proaches rely on expensive and labor-intensive human annotation, making them uns
ustainable at scale. To address this challenge, we propose AgentTrek, a scalable
data synthesis pipeline that generates high-quality web agent trajectories by l
everaging web tutorials. Our method automatically gathers tutorial-like texts fr
om the internet, transforms them into task goals with step-by-step instructions,
and employs a visual-language model (VLM) agent to simulate their execution in
a real digital environment. A VLM-based evaluator ensures the correctness of the
generated trajectories. We demonstrate that training GUI agents with these synt
hesized trajectories significantly improves their grounding and planning perform
ance over the current models. Moreover, our approach is more cost-efficient comp
ared to traditional human annotation methods. This work underscores the potentia
l of guided replay with web tutorials as a viable strategy for large-scale GUI a
gent training, paving the way for more capable and autonomous digital agents.
****
Zeyuan Ma,Jiacheng Chen,Hongshu Guo,Yue-Jiao Gong
Neural Exploratory Landscape Analysis for Meta-Black-Box-Optimization
Recent research in Meta-Black-Box-Optimization (MetaBBO) have shown that meta-tr
ained neural networks can effectively guide the design of black-box optimizers,
significantly reducing the need for expert tuning and delivering robust performa
nce across complex problem distributions. Despite their success, a paradox remai

ns: MetaBBO still rely on human-crafted Exploratory Landscape Analysis features to inform the meta-level agent about the low-level optimization progress. To address the gap, this paper proposes Neural Exploratory Landscape Analysis (NeurELA), a novel framework that dynamically profiles landscape features through a two-stage, attention-based neural network, executed in an entirely end-to-end fashion. NeurELA is pre-trained over a variety of MetaBBO algorithms using a multi-task neuroevolution strategy. Extensive experiments show that NeurELA achieves consistently superior performance when integrated into different and even unseen MetaBBO tasks and can be efficiently fine-tuned for further performance boost. This advancement marks a pivotal step in making MetaBBO algorithms more autonomous and broadly applicable. The source code of NeurELA can be accessed at https://anonymous.4open.science/r/Neur-ELA-303C.

****

Jack Merullo,Noah A. Smith,Sarah Wiegreffe,Yanai Elazar

On Linear Representations and Pretraining Data Frequency in Language Models

Pretraining data has a direct impact on the behaviors and quality of language models (LMs), but we only understand the most basic principles of this relationship. While most work focuses on pretraining data's effect on downstream task behavior, we investigate its relationship to LM representations. Previous work has discovered that, in language models, some concepts are encoded "linearly" in the representations, but what factors cause these representations to form (or not)? We study the connection between pretraining data frequency and models' linear representations of factual relations (e.g., mapping France to Paris in a capital prediction task). We find evidence that the formation of linear representations is strongly connected to pretraining term frequencies; specifically for subject-relation-object fact triplets, both subject-object co-occurrence frequency and in-context learning accuracy for the relation are highly correlated with linear representations. This is the case across all phases of pretraining, i.e., it is not affected by the model's underlying capability. In OLMo-7B and GPT-J (6B), we discover that a linear representation consistently (but not exclusively) forms when the subjects and objects within a relation co-occur at least 1k and 2k times, respectively, regardless of when these occurrences happen during pretraining (and around 4k times for OLMo-1B). Finally, we train a regression model on measurements of linear representation quality in fully-trained LMs that can predict how often a term was seen in pretraining. Our model achieves low error even on inputs from a different model with a different pretraining dataset, providing a new method for estimating properties of the otherwise-unknown training data of closed-data models. We conclude that the strength of linear representations in LMs contains signal about the models' pretraining corpora that may provide new avenues for controlling and improving model behavior: particularly, manipulating the models' training data to meet specific frequency thresholds. We release our code to support future work.

****

Aparna Elangovan,Lei Xu,Jongwoo Ko,Mahsa Elyasi,Ling Liu,Sravan Babu Bodapati,Dan Roth

Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge

The effectiveness of automatic evaluation of generative models is typically measured by comparing the labels generated via automation with human labels using correlation metrics.

However, metrics like Krippendorff's α and Randolph's κ were originally designed to measure the reliability of human labeling, thus make assumptions about typical human labeling behavior, and these assumptions may not be applicable to machine generated labels.

In this paper, we show how *relying on a single aggregate correlation score* can obscure fundamental differences between human labels and those from automatic evaluation, including LLM-as-a-Judge.

Specifically, we demonstrate that when the proportion of samples with variation or uncertainty in human assigned labels is relatively high, machine labels (generated by automatic evaluation methods) may superficially appear to have similar

or better correlation with the human majority label compared to the human-to-human (HH) correlation.
This can create the illusion that labels from automatic evaluation approximates the human majority label.
However, as the proportion of samples with consistent human labels increases, the correlation between machine and human labels fall well below HH correlation.

Based on these findings, we first propose *stratifying data by human label uncertainty* to provide a more robust analysis of automatic evaluation performance. Second, recognizing that uncertainty and variation are inherent in perception-based human evaluations, such as those involving attitudes or preferences, we introduce a new metric -*binned Jensen-Shannon Divergence for perception* for such scenarios to better measure the effectiveness of automatic evaluations. Third, we present visualization techniques -- *perception charts*, to contextualize correlation measures appropriately and to show the strengths and limitations of automatic evaluation. We have open-sourced our analysis and visualization tools at https://github.com/amazon-science/BeyondCorrelation.
****
Jing Yang,Minyue Jiang,Sen Yang,Xiao Tan,Yingying Li,Errui Ding,Jingdong Wang,Hanli Wang
MGMapNet: Multi-Granularity Representation Learning for End-to-End Vectorized HD Map Construction
The construction of vectorized high-definition map typically requires capturing both category and geometry information of map elements. Current state-of-the-art methods often adopt solely either point-level or instance-level representation, overlooking the strong intrinsic relationship between points and instances. In this work, we propose a simple yet efficient framework named MGMapNet (multi-granularity map network) to model map elements with multi-granularity representation, integrating both coarse-grained instance-level and fine-grained point-level queries. Specifically, these two granularities of queries are generated from the multi-scale bird's eye view features using a proposed multi-granularity aggregator. In this module, instance-level query aggregates features over the entire scope covered by an instance, and the point-level query aggregates features locally. Furthermore, a point-instance interaction module is designed to encourage information exchange between instance-level and point-level queries. Experimental results demonstrate that the proposed MGMapNet achieves state-of-the-art performances, surpassing MapTRv2 by 5.3 mAP on the nuScenes dataset and 4.4 mAP on the Argoverse2 dataset, respectively.
****
Hyungjin Chung,Jeongsol Kim,Geon Yeong Park,Hyelin Nam,Jong Chul Ye
CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models
Classifier-free guidance (CFG) is a fundamental tool in modern diffusion models for text-guided generation. Although effective, CFG has notable drawbacks. For instance, DDIM with CFG lacks invertibility, complicating image editing; furthermore, high guidance scales, essential for high-quality outputs, frequently result in issues like mode collapse. Contrary to the widespread belief that these are inherent limitations of diffusion models, this paper reveals that the problems actually stem from the off-manifold phenomenon associated with CFG, rather than the diffusion models themselves. More specifically, inspired by the recent advancements of diffusion model-based inverse problem solvers (DIS), we reformulate text-guidance as an inverse problem with a text-conditioned score matching loss and develop CFG++, a novel approach that tackles the off-manifold challenges inherent in traditional CFG. CFG++ features a surprisingly simple fix to CFG, yet it offers significant improvements, including better sample quality for text-to-image generation, invertibility, smaller guidance scales, reduced etc. Furthermore, CFG++ enables seamless interpolation between unconditional and conditional sampling at lower guidance scales, consistently outperforming traditional CFG at all scales. Moreover, CFG++ can be easily integrated into the high-order diffusion solvers and naturally extends to distilled diffusion models. Experimental results confirm that our method significantly enhances performance in text-to-image

generation, DDIM inversion, editing, and solving inverse problems, suggesting a wide-ranging impact and potential applications in various fields that utilize text guidance. Project Page: https://cfgpp-diffusion.github.io/anon
****

Wei Chen,Zichen Miao,Qiang Qiu
Large Convolutional Model Tuning via Filter Subspace
Efficient fine-tuning methods are critical to address the high computational and parameter complexity while adapting large pre-trained models to downstream tasks.
Our study is inspired by prior research that represents each convolution filter as a linear combination of a small set of filter subspace elements, referred to as filter atoms. In this paper, we propose to fine-tune pre-trained models by adjusting only filter atoms, which are responsible for spatial-only convolution, while preserving spatially-invariant channel combination knowledge in atom coefficients.
In this way, we bring a new filter subspace view for model tuning.
Furthermore, each filter atom can be recursively decomposed as a combination of another set of atoms, which naturally expands the number of tunable parameters in the filter subspace.
By only adapting filter atoms constructed by a small number of parameters, while maintaining the rest of model parameters constant, the proposed approach is highly parameter-efficient. It effectively preserves the capabilities of pre-trained models and prevents overfitting to downstream tasks.
Extensive experiments show that such a simple scheme surpasses previous tuning baselines for both discriminate and generative tasks.
****

Thibaud Gloaguen,Nikola Jovanovi■,Robin Staab,Martin Vechev
Black-Box Detection of Language Model Watermarks
Watermarking has emerged as a promising way to detect LLM-generated text, by augmenting LLM generations with later detectable signals. Recent work has proposed multiple families of watermarking schemes, several of which focus on preserving the LLM distribution. This distribution-preservation property is motivated by the fact that it is a tractable proxy for retaining LLM capabilities, as well as the inherently implied undetectability of the watermark by downstream users. Yet, despite much discourse around undetectability, no prior work has investigated the practical detectability of any of the current watermarking schemes in a realistic black-box setting. In this work we tackle this for the first time, developing rigorous statistical tests to detect the presence, and estimate parameters, of all three popular watermarking scheme families, using only a limited number of black-box queries. We experimentally confirm the effectiveness of our methods on a range of schemes and a diverse set of open-source models. Further, we validate the feasibility of our tests on real-world APIs. Our findings indicate that current watermarking schemes are more detectable than previously believed.
****

Erik Wijmans,Brody Huval,Alexander Hertzberg,Vladlen Koltun,Philipp Kraehenbuehl
Cut Your Losses in Large-Vocabulary Language Models
As language models grow ever larger, so do their vocabularies.
This has shifted the memory footprint of LLMs during training disproportionately to one single layer: the cross-entropy in the loss computation.
Cross-entropy builds up a logit matrix with entries for each pair of input tokens and vocabulary items and, for small models, consumes an order of magnitude more memory than the rest of the LLM combined.
We propose Cut Cross-Entropy (CCE), a method that computes the cross-entropy loss without materializing the logits for all tokens into global memory.
Rather, CCE only computes the logit for the correct token and evaluates the log-sum-exp over all logits on the fly.
We implement a custom kernel that performs the matrix multiplications and the log-sum-exp reduction over the vocabulary in flash memory, making global memory consumption for the cross-entropy computation negligible. This has a dramatic effect. Taking the Gemma 2 (2B) model as an example, CCE reduces the memory footprin

t of the loss computation from 24 GB to 1 MB, and the total training-time memory consumption of the classifier head from 28 GB to 1 GB.
To improve the throughput of CCE, we leverage the inherent sparsity of softmax and propose to skip elements of the gradient computation that have a negligible (i.e. below numerical precision) contribution to the gradient.
Experiments demonstrate that the dramatic reduction in memory consumption is accomplished without sacrificing training speed or convergence.
****
Alan Nawzad Amin,Nate Gruver,Yilun Kuang,Yucen Lily Li,Hunter Elliott,Calvin McCarter,Aniruddh Raghu,Peyton Greenside,Andrew Gordon Wilson
Bayesian Optimization of Antibodies Informed by a Generative Model of Evolving Sequences
To build effective therapeutics, biologists iteratively mutate antibody sequences to improve binding and stability. Proposed mutations can be informed by previous measurements or by learning from large antibody databases to predict only typical antibodies. Unfortunately, the space of typical antibodies is enormous to search, and experiments often fail to find suitable antibodies on a budget. We introduce Clone-informed Bayesian Optimization (CloneBO), a Bayesian optimization procedure that efficiently optimizes antibodies in the lab by teaching a generative model how our immune system optimizes antibodies. Our immune system makes antibodies by iteratively evolving specific portions of their sequences to bind their target strongly and stably, resulting in a set of related, evolving sequences known as a *clonal family*. We train a large language model, CloneLM, on hundreds of thousands of clonal families and use it to design sequences with mutations that are most likely to optimize an antibody within the human immune system. We propose to guide our designs to fit previous measurements with a twisted sequential Monte Carlo procedure. We show that CloneBO optimizes antibodies substantially more efficiently than previous methods in realistic *in silico* experiments and designs stronger and more stable binders in *in vitro* wet lab experiments.
****
Lili Zhao,Yang Wang,Qi Liu,Mengyun Wang,Wei Chen,Zhichao Sheng,Shijin Wang
Evaluating Large Language Models through Role-Guide and Self-Reflection: A Comparative Study
Large Language Models fine-tuned with Reinforcement Learning from Human Feedback (RLHF-LLMs) can over-rely on aligned preferences without truly gaining self-knowledge, leading to hallucination and biases. If an LLM can better access its knowledge and know what it knows, it can avoid making false or unsupported claims. Therefore, it is crucial to evaluate whether LLMs have the ability to know what they know, as it can help to ensure accuracy and faithfulness in real-world applications. Inspired by research in Educational Psychology, surface learners who don’t really know are easily affected by teacher and peer guidance, we treat LLM as a student, incorporate role guidance in prompts to explore whether LLMs really know. Specifically, we propose a novel strategy called Role-Guided and Self-Reflection (RoSe) to fully assess whether LLM “knows it knows”. We introduce multiple combinations of different roles and strong reminder in prompts combined with self-reflection to explore what local information in prompt LLMs rely on and whether LLMs remain unaffected by external guidance with varying roles. Our findings reveal that LLMs are very sensitive to the strong reminder information. Role guidance can help LLMs reduce their reliance on strong reminder. Meanwhile, LLMs tend to trust the role of authority more when guided by different roles. Following these findings, we propose a double-calibrated strategy with verbalized confidence to extract well-calibrated data from closed-source LLM and fine-tune open-source LLMs. Extensive experiments conducted on fine-tuning open-source LLMs demonstrate the effectiveness of double-calibrated strategy in mitigating the reliance of LLMs on local information. For a thorough comparison, we not only employ public JEC-QA and openBookQA datasets, but also construct EG-QA which contains English Grammar multiple-choice question-answering and 14 key knowledge points for assessing self-knowledge and logical reasoning.
****
Shuo Li,Tao Ji,Xiaoran Fan,Linsheng Lu,Leyi Yang,Yuming Yang,Zhiheng Xi,Rui Zhen

g,Yuran Wang,xh.zhao,Tao Gui,Qi Zhang,Xuanjing Huang
Have the VLMs Lost Confidence? A Study of Sycophancy in VLMs
In the study of LLMs, sycophancy represents a prevalent hallucination that poses significant challenges to these models. Specifically, LLMs often fail to adhere to original correct responses, instead blindly agreeing with users' opinions, even when those opinions are incorrect or malicious. However, research on sycophancy in visual language models (VLMs) has been scarce. In this work, we extend the exploration of sycophancy from LLMs to VLMs, introducing the MM-SY benchmark to evaluate this phenomenon. We present evaluation results from multiple representative models, addressing the gap in sycophancy research for VLMs. To mitigate sycophancy, we propose a synthetic dataset for training and employ methods based on prompts, supervised fine-tuning, and DPO. Our experiments demonstrate that these methods effectively alleviate sycophancy in VLMs. Additionally, we probe VLMs to assess the semantic impact of sycophancy and analyze the attention distribution of visual tokens. Our findings indicate that the ability to prevent sycophancy is predominantly observed in higher layers of the model. The lack of attention to image knowledge in these higher layers may contribute to sycophancy, and enhancing image attention at high layers proves beneficial in mitigating this issue.
****
Dominik Scheuer,Frederic Runge,Jörg K.H. Franke,Michael T. Wolfinger,Christoph Flamm,Frank Hutter
KinPFN: Bayesian Approximation of RNA Folding Kinetics using Prior-Data Fitted Networks
RNA is a dynamic biomolecule crucial for cellular regulation, with its function largely determined by its folding into complex structures, while misfolding can lead to multifaceted biological sequelae. During the folding process, RNA traverses through a series of intermediate structural states, with each transition occurring at variable rates that collectively influence the time required to reach the functional form. Understanding these folding kinetics is vital for predicting RNA behavior and optimizing applications in synthetic biology and drug discovery. While in silico kinetic RNA folding simulators are often computationally intensive and time-consuming, accurate approximations of the folding times can already be very informative to assess the efficiency of the folding process. In this work, we present KinPFN, a novel approach that leverages prior-data fitted networks to directly model the posterior predictive distribution of RNA folding times. By training on synthetic data representing arbitrary prior folding times, KinPFN efficiently approximates the cumulative distribution function of RNA folding times in a single forward pass, given only a few initial folding time examples. Our method offers a modular extension to existing RNA kinetics algorithms, promising significant computational speed-ups orders of magnitude faster, while achieving comparable results. We showcase the effectiveness of KinPFN through extensive evaluations and real-world case studies, demonstrating its potential for RNA folding kinetics analysis, its practical relevance, and generalization to other biological data.
****
Tianchen Zhao,Tongcheng Fang,Haofeng Huang,Rui Wan,Widyadewi Soedarmadji,Enshu Liu,Shiyao Li,Zinan Lin,Guohao Dai,Shengen Yan,Huazhong Yang,Xuefei Ning,Yu Wang
ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation
Diffusion transformers have demonstrated remarkable performance in visual generation tasks, such as generating realistic images or videos based on textual instructions. However, larger model sizes and multi-frame processing for video generation lead to increased computational and memory costs, posing challenges for practical deployment on edge devices. Post-Training Quantization (PTQ) is an effective method for reducing memory costs and computational complexity.
When quantizing diffusion transformers, we find that existing quantization methods face challenges when applied to text-to-image and video tasks. To address these challenges, we begin by systematically analyzing the source of quantization error and conclude with the unique challenges posed by DiT quantization. Accordin

gly, we design an improved quantization scheme: ViDiT-Q (**V**ideo \& **I**mage **Di**ffusion **T**ransformer **Q**uantization), tailored specifically for DiT models. We validate the effectiveness of ViDiT-Q across a variety of text-to-image and video models, achieving W8A8 and W4A8 with negligible degradation in visual quality and metrics. Additionally, we implement efficient GPU kernels to achieve practical 2-2.5x memory optimization and a 1.4-1.7x end-to-end latency speedup.
****
Maxim Fishman,Brian Chmiel,Ron Banner,Daniel Soudry
Scaling FP8 training to trillion-token LLMs
We train, for the first time, large language models using FP8 precision on datasets up to 2 trillion tokens --- a 20-fold increase over previous limits. Through these extended training runs, we uncover critical instabilities in FP8 training that were not observable in earlier works with shorter durations. We trace these instabilities to outlier amplification by the SwiGLU activation function. Interestingly, we show, both analytically and empirically, that this amplification happens only over prolonged training periods, and link it to a SwiGLU weight alignment process. To address this newly identified issue, we introduce Smooth-SwiGLU, a novel modification that ensures stable FP8 training without altering function behavior. We also demonstrate, for the first time, FP8 quantization of both Adam optimizer moments. Combining these innovations, we successfully train a 7B parameter model using FP8 precision on 256 Intel Gaudi2 accelerators, achieving on-par results with the BF16 baseline while delivering up to a $\sim$ 34 % throughput improvement. A reference implementation is supplied in https://github.com/Anonymous1252022/Megatron-DeepSpeed
****
Nico Messikommer,Jiaxu Xing,Elie Aljalbout,Davide Scaramuzza
Student-Informed Teacher Training
Imitation learning with a privileged teacher has proven effective for learning complex control behaviors from high-dimensional inputs, such as images. In this framework, a teacher is trained with privileged task information, while a student tries to predict the actions of the teacher with more limited observations, e.g., in a robot navigation task, the teacher might have access to distances to nearby obstacles, while the student only receives visual observations of the scene. However, privileged imitation learning faces a key challenge: the student might be unable to imitate the teacher's behavior due to partial observability. This problem arises because the teacher is trained without considering if the student is capable of imitating the learned behavior. To address this teacher-student asymmetry, we propose a framework for joint training of the teacher and student policies, encouraging the teacher to learn behaviors that can be imitated by the student despite the latters' limited access to information and its partial observability. Based on the performance bound in imitation learning, we add (i) the approximated action difference between teacher and student as a penalty term to the reward function of the teacher, and (ii) a supervised teacher-student alignment step. We motivate our method with a maze navigation task and demonstrate its effectiveness on complex vision-based quadrotor flight and manipulation tasks.
****
Yuan Bian,Grace Yi,Wenqing He
Boosting Methods for Interval-censored Data with Regression and Classification
Boosting has garnered significant interest across both machine learning and statistical communities. Traditional boosting algorithms, designed for fully observed random samples, often struggle with real-world problems, particularly with interval-censored data. This type of data is common in survival analysis and time-to-event studies where exact event times are unobserved but fall within known intervals. Effective handling of such data is crucial in fields like medical research, reliability engineering, and social sciences. In this work, we introduce novel nonparametric boosting methods for regression and classification tasks with interval-censored data. Our approaches leverage censoring unbiased transformations to adjust loss functions and impute transformed responses while maintaining model accuracy. Implemented via functional gradient descent, these methods ensure

scalability and adaptability. We rigorously establish their theoretical properti
es, including optimality and mean squared error trade-offs. Our proposed methods
 not only offer a robust framework for enhancing predictive accuracy in domains
where interval-censored data are common but also complement existing work, expan
ding the applicability of existing boosting techniques. Empirical studies demons
trate robust performance across various finite-sample scenarios, highlighting th
e practical utility of our approaches.
****
Thomas Bush,Stephen Chung,Usman Anwar,Adrià Garriga-Alonso,David Krueger
Interpreting Emergent Planning in Model-Free Reinforcement Learning
We present the first mechanistic evidence that model-free reinforcement learning
 agents can learn to plan. This is achieved by applying a methodology based on c
oncept-based interpretability to a model-free agent in Sokoban -- a commonly use
d benchmark for studying planning. Specifically, we demonstrate that DRC, a gene
ric model-free agent introduced by [Guez et al. (2019)](https://arxiv.org/abs/19
01.03559), uses learned concept representations to internally formulate plans th
at both predict the long-term effects of actions on the environment and influenc
e action selection. Our methodology involves: (1) probing for planning-relevant
concepts, (2) investigating plan formation within the agent's representations, a
nd (3) verifying that discovered plans (in the agent's representations) have a c
ausal effect on the agent's behavior through interventions. We also show that th
e emergence of these plans coincides with the emergence of a planning-like prope
rty: the ability to benefit from additional test-time compute. Finally, we perfo
rm a qualitative analysis of the planning algorithm learned by the agent and dis
cover a strong resemblance to parallelized bidirectional search. Our findings a
dvance understanding of the internal mechanisms underlying planning behavior in
agents, which is important given the recent trend of emergent planning and reaso
ning capabilities in LLMs through RL.
****
Zhizheng Liu,Joe Lin,Wayne Wu,Bolei Zhou
Learning to Generate Diverse Pedestrian Movements from Web Videos with Noisy Lab
els
Understanding and modeling pedestrian movements in the real world is crucial for
 applications like motion forecasting and scene simulation. Many factors influen
ce pedestrian movements, such as scene context, individual characteristics, and
goals, which are often ignored by the existing human generation methods. Web vid
eos contain natural pedestrian behavior and rich motion context, but annotating
them with pre-trained predictors leads to noisy labels. In this work, we propose
 learning diverse pedestrian movements from web videos. We first curate a large-
scale dataset called CityWalkers that captures diverse real-world pedestrian mov
ements in urban scenes. Then, based on CityWalkers, we propose a generative mod
el called PedGen for diverse pedestrian movement generation. PedGen introduces a
utomatic label filtering to remove the low-quality labels and a mask embedding t
o train with partial labels. It also contains a novel context encoder that lifts
 the 2D scene context to 3D and can incorporate various context factors in gener
ating realistic pedestrian movements in urban scenes. Experiments show that PedG
en outperforms existing baseline methods for pedestrian movement generation by l
earning from noisy labels and incorporating the context factors. In addition, Pe
dGen achieves zero-shot generalization in both real-world and simulated environm
ents. The code, model, and data are available at https://genforce.github.io/PedG
en/.
****
Milad Nasr,Thomas Steinke,Borja Balle,Christopher A. Choquette-Choo,Arun Ganesh,
Matthew Jagielski,Jamie Hayes,Abhradeep Guha Thakurta,Adam Smith,Andreas Terzis
The Last Iterate Advantage: Empirical Auditing and Principled Heuristic Analysis
 of Differentially Private SGD
We propose a simple heuristic privacy analysis of noisy clipped stochastic gradi
ent descent (DP-SGD) in the setting where only the last iterate is released and
the intermediate iterates remain hidden. Namely, our heuristic assumes a linear
structure for the model.

We show experimentally that our heuristic is predictive of the outcome of privac
y auditing applied to various training procedures. Thus it can be used prior to
training as a rough estimate of the final privacy leakage. We also probe the lim
itations of our heuristic by providing some artificial counterexamples where it
underestimates the privacy leakage.

The standard composition-based privacy analysis of DP-SGD effectively assumes th
at the adversary has access to all intermediate iterates, which is often unreali
stic. However, this analysis remains the state of the art in practice. While our
 heuristic does not replace a rigorous privacy analysis, it illustrates the larg
e gap between the best theoretical upper bounds and the privacy auditing lower b
ounds and sets a target for further work to improve the theoretical privacy anal
yses.
****
Vladimír Boža,Vladimír Macko
Two Sparse Matrices are Better than One: Sparsifying Neural Networks with Double
 Sparse Factorization
Neural networks are often challenging to work with due to their large size and c
omplexity. To address this, various methods aim to reduce model size by sparsify
ing or decomposing weight matrices, such as magnitude pruning and low-rank or bl
ock-diagonal factorization. In this work, we present Double Sparse Factorization
 (DSF), where we factorize each weight matrix into two sparse matrices. Although
 solving this problem exactly is computationally infeasible, we propose an effic
ient heuristic based on alternating minimization via ADMM that achieves state-of
-the-art results, enabling unprecedented sparsification of neural networks. For
instance, in a one-shot pruning setting, our method can reduce the size of the L
LaMA2-13B model by 50% while maintaining better performance than the dense LLaMA
2-7B model. We also compare favorably with Optimal Brain Compression, the state-
of-the-art layer-wise pruning approach for convolutional neural networks. Furthe
rmore, accuracy improvements of our method persist even after further model fine
-tuning.
Code available at: https://github.com/usamec/double_sparse
****
Patrick Pynadath,Ruqi Zhang
Controlled LLM Decoding via Discrete Auto-regressive Biasing
Controlled text generation allows for enforcing user-defined constraints on larg
e language model outputs, an increasingly important field as LLMs become more pr
evalent in everyday life. One common approach uses energy-based decoding, which
defines a target distribution through an energy function that combines multiple
constraints into a weighted average. However, these methods often struggle to ba
lance fluency with constraint satisfaction, even with extensive tuning of the en
ergy function's coefficients. In this paper, we identify that this suboptimal ba
lance arises from sampling in continuous space rather than the natural discrete
space of text tokens. To address this, we propose \emph{Discrete Auto-regressive
 Biasing}, a controlled decoding algorithm that leverages gradients while operat
ing entirely in the discrete text domain.
Specifically, we introduce a new formulation for controlled text generation by d
efining a joint distribution over the generated sequence and an auxiliary bias s
equence. To efficiently sample from this joint distribution, we propose a Langev
in-within-Gibbs sampling algorithm using gradient-based discrete MCMC. Our metho
d significantly improves constraint satisfaction while maintaining comparable or
 better fluency, all with lower computational costs. We demonstrate the advantag
es of our controlled decoding method on sentiment control, language detoxificati
on, and keyword-guided generation.
****
Yaxuan Huang,Xili Dai,Jianan Wang,Xianbiao Qi,Yixing Yuan,Xiangyu Yue
Unposed Sparse Views Room Layout Reconstruction in the Age of Pretrain Model
Room layout estimation from multiple-perspective images is poorly investigated d
ue to the complexities that emerge from multi-view geometry, which requires muti

-step solutions such as camera intrinsic and extrinsic estimation, image matchin g, and triangulation. However, in 3D reconstruction, the advancement of recent 3 D foundation models such as DUSt3R has shifted the paradigm from the traditional multi-step structure-from-motion process to an end-to-end single-step approach. To this end, we introduce Plane-DUSt3R, a novel method for multi-view room layou t estimation leveraging the 3D foundation model DUSt3R. Plane-DUSt3R incorporate s the DUSt3R framework and fine-tunes on a room layout dataset (Structure3D) wit h a modified objective to estimate structural planes. By generating uniform and parsimonious results, Plane-DUSt3R enables room layout estimation with only a si ngle post-processing step and 2D detection results.
Unlike previous methods that rely on single-perspective or panorama image, Plane -DUSt3R extends the setting to handle multiple-perspective images. Moreover, it offers a streamlined, end-to-end solution that simplifies the process and reduce s error accumulation.
Experimental results demonstrate that Plane-DUSt3R not only outperforms state-of -the-art methods on the synthetic dataset but also proves robust and effective o n in the wild data with different image styles such as cartoon. Our code is avai lable at: https://github.com/justacar/Plane-DUSt3R
****

Hongru Yan,Yu Zheng,Yueqi Duan
Gaussian-Det: Learning Closed-Surface Gaussians for 3D Object Detection
Skins wrapping around our bodies, leathers covering over the sofa, sheet metal c oating the car – it suggests that objects are enclosed by a series of continuous surfaces, which provides us with informative geometry prior for objectness dedu ction. In this paper, we propose Gaussian-Det which leverages Gaussian Splatting as surface representation for multi-view based 3D object detection. Unlike exis ting monocular or NeRF-based methods which depict the objects via discrete posit ional data, Gaussian-Det models the objects in a continuous manner by formulatin g the input Gaussians as feature descriptors on a mass of partial surfaces. Furt hermore, to address the numerous outliers inherently introduced by Gaussian spla tting, we accordingly devise a Closure Inferring Module (CIM) for the comprehens ive surface-based objectness deduction. CIM firstly estimates the probabilistic feature residuals for partial surfaces given the underdetermined nature of Gauss ian Splatting, which are then coalesced into a holistic representation on the ov erall surface closure of the object proposal. In this way, the surface informati on Gaussian-Det exploits serves as the prior on the quality and reliability of o bjectness and the information basis of proposal refinement. Experiments on both synthetic and real-world datasets demonstrate that Gaussian-Det outperforms vari ous existing approaches, in terms of both average precision and recall.
****

Yongwei Che,Benjamin Eysenbach
The "Law'' of the Unconscious Contrastive Learner: Probabilistic Alignment of Un paired Modalities
While internet-scale data often come in pairs (e.g., audio+image, image+text), w e often want to perform inferences over modalities unseen together in the traini ng data (e.g., audio+text). Prior work has addressed this issue by learning mult iple contrastive embedding spaces between existing modality pairs, implicitly ho ping that unseen modality pairs will end up being aligned. This theoretical pape r proves that this hope is well founded, under certain assumptions. Starting wit h the proper Bayesian approach of integrating out intermediate modalities, we sh ow that directly comparing the representations of data from unpaired modalities can recover the same likelihood ratio. Our analysis builds on prior work on the geometry and probabilistic interpretation of contrastive representations, showin g how these representations can answer many of the same inferences as probabilis tic graphical models. Our analysis suggests two new ways of using contrastive re presentations: in settings with pre-trained contrastive models, and for handling language ambiguity in reinforcement learning. Our numerical experiments study t he importance of our assumptions and demonstrate these new applications.
****

KE WANG,Shay B Cohen

DEPfold: RNA Secondary Structure Prediction as Dependency Parsing.
RNA secondary structure prediction is critical for understanding RNA function but remains challenging due to complex structural elements like pseudoknots and limited training data. We introduce DEPfold, a novel deep learning approach that re-frames RNA secondary structure prediction as a dependency parsing problem. DEPfold presents three key innovations: (1) a biologically motivated transformation of RNA structures into labeled dependency trees, (2) a biaffine attention mechanism for joint prediction of base pairings and their types, and (3) an optimal
tree decoding algorithm that enforces valid RNA structural constraints. Unlike traditional energy-based methods, DEPfold learns directly from annotated data and leverages pretrained language models to predict RNA structure. We evaluate DEPfold on both within-family and cross-family RNA datasets, demonstrating significant performance improvements over existing methods. DEPfold shows strong performance in cross-family generalization when trained on data augmented by traditional energy-based models, outperforming existing methods on the bpRNAnew dataset. This demonstrates DEPfold’s ability to effectively learn structural information beyond what traditional methods capture. Our approach bridges natural language processing (NLP) with RNA biology, providing a computationally efficient and adaptable tool for advancing RNA structure prediction and analysis
****
Md Imtiaz Hossain,Sharmen Akhter,Choong Seon Hong,Eui-Nam Huh
Single Teacher, Multiple Perspectives: Teacher Knowledge Augmentation for Enhanced Knowledge Distillation
Do diverse perspectives help students learn better? Multi-teacher knowledge distillation, which is a more effective technique than traditional single-teacher methods, supervises the student from different perspectives (i.e., teacher). While effective, multi-teacher, teacher ensemble, or teaching assistant-based approaches are computationally expensive and resource-intensive, as they require training multiple teacher networks. These concerns raise a question: can we supervise the student with diverse perspectives using only a single teacher? We, as the pioneer, demonstrate TeKAP, a novel teacher knowledge augmentation technique that generates multiple synthetic teacher knowledge by perturbing the knowledge of a single pretrained teacher i.e., Teacher Knowledge Augmentation via Perturbation, at both the feature and logit levels. These multiple augmented teachers simulate an ensemble of models together. The student model is trained on both the actual and augmented teacher knowledge, benefiting from the diversity of an ensemble without the need to train multiple teachers. TeKAP significantly reduces training time and computational resources, making it feasible for large-scale applications and easily manageable. Experimental results demonstrate that our proposed method helps existing state-of-the-art knowledge distillation techniques achieve better performance, highlighting its potential as a cost-effective alternative. The source code can be found in the supplementary.
****
Ahmed Abdulaal,Chen Jin,Nina Montaña-Brown,Aryo Pradipta Gema,Daniel C. Castro,Daniel C. Alexander,Philip Alexander Teare,Tom Diethe,Dino Oglic,Amrutha Saseendran
Balancing Act: Diversity and Consistency in Large Language Model Ensembles
Ensembling strategies for Large Language Models (LLMs) have demonstrated significant potential in improving performance across various tasks by combining the strengths of individual models. However, identifying the most effective ensembling method remains an open challenge, as neither maximizing output consistency through self-consistency decoding nor enhancing model diversity via frameworks like "Mixture of Agents" has proven universally optimal. Motivated by this, we propose a unified framework to examine the trade-offs between task performance, model diversity, and output consistency in ensembles. More specifically, we introduce a consistency score that defines a gating mechanism for mixtures of agents and an algorithm for mixture refinement to investigate these trade-offs at the semantic and model levels, respectively. We incorporate our insights into a novel inference-time LLM ensembling strategy called the Dynamic Mixture of Agents (DMoA)

and demonstrate that it achieves a new state-of-the-art result in the challengin
g Big Bench Hard mixed evaluations benchmark. Our analysis reveals that cross-va
lidation bias can enhance performance, contingent on the expertise of the consti
tuent models. We further demonstrate that distinct reasoning tasks—such as arith
metic reasoning, commonsense reasoning, and instruction following—require differ
ent model capabilities, leading to inherent task-dependent trade-offs that DMoA
balances effectively.
****
Ryan Missel,Linwei Wang
Continual Slow-and-Fast Adaptation of Latent Neural Dynamics (CoSFan): Meta-Lear
ning What-How & When to Adapt
An increasing interest in learning to forecast for time-series of high-dimension
al observations is the ability to adapt to systems with diverse underlying dynam
ics. Access to observations that define a stationary distribution of these syste
ms is often unattainable, as the underlying dynamics may change over time. Naive
ly training or retraining models at each shift may lead to catastrophic forgetti
ng about previously-seen systems. We present a new continual meta-learning (CML)
 framework to realize continual slow-and fast adaptation of latent dynamics (CoS
Fan). We leverage a feed-forward meta-model to infer *what* the current system i
s and *how* to adapt a latent dynamics function to it, enabling *fast adaptation
* to specific dynamics. We then develop novel strategies to automatically detect
 when a shift of data distribution occurs, with which to identify its underlyi
ng dynamics and its relation with previously-seen dynamics. In combination with
fixed-memory experience replay mechanisms, this enables continual *slow update*
of the *what-how* meta-model. Empirical studies demonstrated that both the meta-
 and continual-learning component was critical for learning to forecast across n
on-stationary distributions of diverse dynamics systems, and the feed-forward me
ta-model combined with task-aware/-relational continual learning strategies sign
ificantly outperformed existing CML alternatives.
****
Xize Cheng,Siqi Zheng,Zehan Wang,Minghui Fang,Ziang Zhang,Rongjie Huang,Shengpen
g Ji,Jialong Zuo,Tao Jin,Zhou Zhao
OmniSep: Unified Omni-Modality Sound Separation with Query-Mixup
Query-based sound separation (QSS) effectively isolate sound signals that match
the content of a given query, enhancing the understanding of audio data. However
, most existing QSS methods rely on a single modality for separation, lacking th
e ability to fully leverage homologous but heterogeneous information across mult
iple modalities for the same sound signal. To address this limitation, we introd
uce Omni-modal Sound Separation (**OmniSep**), a novel framework capable of isol
ating clean soundtracks based on omni-modal queries, encompassing both single-mo
dal and multi-modal composed queries. Specifically, we introduce the **Query-Mix
up** strategy, which blends query features from different modalities during trai
ning. This enables OmniSep to optimize multiple modalities concurrently, effecti
vely bringing all modalities under a unified framework for sound separation. We
further enhance this flexibility by allowing queries to influence sound separati
on positively or negatively, facilitating the retention or removal of specific s
ounds as desired. Finally, OmniSep employs a retrieval-augmented approach known
as **Query-Aug**, which enables open-vocabulary sound separation. Experimental e
valuations on MUSIC, VGGSOUND-CLEAN+, and MUSIC-CLEAN+ datasets demonstrate effe
ctiveness of OmniSep, achieving state-of-the-art performance in text-, image-, a
nd audio-queried sound separation tasks. For samples and further information, pl
ease visit the demo page at \url{https://omnisep.github.io/}.
****
Yuke Zhu,Yue Zhang,Dongdong Liu,Chi Xie,Zihua Xiong,Bo Zheng,Sheng Guo
Enhancing Document Understanding with Group Position Embedding: A Novel Approach
 to Incorporate Layout Information
Recent advancements in document understanding have been dominated by leveraging
large language models (LLMs) and multimodal large models. However, enabling LLMs
 to comprehend complex document layouts and structural information often necessi
tates intricate network modifications or costly pre-training, limiting their pra

ctical applicability. In this paper, we introduce Group Position Embedding (GPE), a novel and efficient technique to enhance the layout understanding capabilities of LLMs without architectural changes or additional pre-training. GPE achieves this by strategically grouping the attention heads and feeding each group with distinct positional embeddings, effectively encoding layout information relevant to document comprehension. This simple yet powerful method allows for effective integration of layout information within the existing LLM framework. We evaluate GPE against several competitive baselines across five mainstream document tasks. We also introduce a challenging benchmark called BLADE, specifically designed to assess layout comprehension.
Extensive experiments on both established and BLADE benchmarks confirm the efficacy of GPE in significantly advancing the state-of-the-art in document understanding. Our code is available at https://github.com/antgroup/GroupPositionEmbedding.git
****
Nikola Zubic,Federico Soldà,Aurelio Sulser,Davide Scaramuzza
Limits of Deep Learning: Sequence Modeling through the Lens of Complexity Theory
Despite their successes, deep learning models struggle with tasks requiring complex reasoning and function composition. We present a theoretical and empirical investigation into the limitations of Structured State Space Models (SSMs) and Transformers in such tasks. We prove that one-layer SSMs cannot efficiently perform function composition over large domains without impractically large state sizes, and even with Chain-of-Thought prompting, they require a number of steps that scale unfavorably with the complexity of the function composition. Also, the language of a finite-precision SSM is within the class of regular languages. Our experiments corroborate these theoretical findings. Evaluating models on tasks including various function composition settings, multi-digit multiplication, dynamic programming, and Einstein's puzzle, we find significant performance degradation even with advanced prompting techniques. Models often resort to shortcuts, leading to compounding errors. These findings highlight fundamental barriers within current deep learning architectures rooted in their computational capacities. We underscore the need for innovative solutions to transcend these constraints and achieve reliable multi-step reasoning and compositional task-solving, which is critical for advancing toward general artificial intelligence.
****
Minsi Ren,Yan-Ming Zhang,yi chen
Decoupling Layout from Glyph in Online Chinese Handwriting Generation
Text plays a crucial role in the transmission of human civilization, and teaching machines to generate online handwritten text in various styles presents an interesting and significant challenge. However, most prior work has concentrated on generating individual Chinese fonts, leaving complete text line generation largely unexplored. In this paper, we identify that text lines can naturally be divided into two components: layout and glyphs. Based on this division, we designed a text line layout generator coupled with a diffusion-based stylized font synthesizer to address this challenge hierarchically. More concretely, the layout generator performs in-context-like learning based on the text content and the provided style references to generate positions for each glyph autoregressively. Meanwhile, the font synthesizer which consists of a character embedding dictionary, a multi-scale calligraphy style encoder and a 1D U-Net based diffusion denoiser will generate each font on its position while imitating the calligraphy style extracted from the given style references. Qualitative and quantitative experiments on the CASIA-OLHWDB demonstrate that our method is capable of generating structurally correct and indistinguishable imitation samples.
****
Lily Xu,Bryan Wilder,Elias Boutros Khalil,Milind Tambe
Reinforcement learning with combinatorial actions for coupled restless bandits
Reinforcement learning (RL) has increasingly been applied to solve real-world planning problems, with progress in handling large state spaces and time horizons. However, a key bottleneck in many domains is that RL methods cannot accommodate large, combinatorially structured action spaces. In such settings, even represe

nting the set of feasible actions at a single step may require a complex discrete optimization formulation. We leverage recent advances in embedding trained neural networks into optimization problems to propose SEQUOIA, an RL algorithm that directly optimizes for long-term reward over the feasible action space. Our approach embeds a Q-network into a mixed-integer program to select a combinatorial action in each timestep. Here, we focus on planning over restless bandits, a class of planning problems which capture many real-world examples of sequential decision making. We introduce coRMAB, a broader class of restless bandits with combinatorial actions that cannot be decoupled across the arms of the restless bandit, requiring direct solving over the joint, exponentially large action space. We empirically validate SEQUOIA on four novel restless bandit problems with combinatorial constraints: multiple interventions, path constraints, bipartite matching, and capacity constraints. Our approach significantly outperforms existing methods—which cannot address sequential planning and combinatorial selection simultaneously—by an average of 24.8% on these difficult instances.
****

Jiarui Zhang,Mahyar Khayatkhoei,Prateek Chhikara,Filip Ilievski
MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs
Multimodal Large Language Models (MLLMs) have experienced rapid progress in visual recognition tasks in recent years. Given their potential integration into many critical applications, it is important to understand the limitations of their visual perception. In this work, we study whether MLLMs can perceive small visual details as effectively as large ones when answering questions about images. We observe that their performance is very sensitive to the size of the visual subject of the question, and further show that this effect is in fact causal by conducting an intervention study. Next, we study the attention patterns of MLLMs when answering visual questions, and intriguingly find that they consistently know where to look, even when they provide the wrong answer. Based on these findings, we then propose training-free visual intervention methods that leverage the internal knowledge of any MLLM itself, in the form of attention and gradient maps, to enhance its perception of small visual details. We evaluate our proposed methods on two widely-used MLLMs and seven visual question answering benchmarks and show that they can significantly improve MLLMs' accuracy without requiring any training. Our results elucidate the risk of applying MLLMs to visual recognition tasks concerning small details and indicate that visual intervention using the model's internal state is a promising direction to mitigate this risk. Our code is available at: https://github.com/saccharomycetes/mllms_know.
****

Claire Chen,Shuze Liu,Shangtong Zhang
Efficient Policy Evaluation with Safety Constraint for Reinforcement Learning
In reinforcement learning, classic on-policy evaluation methods often suffer from high variance and require massive online data to attain the desired accuracy. Previous studies attempt to reduce evaluation variance by searching for or designing proper behavior policies to collect data. However, these approaches ignore the safety of such behavior policies---the designed behavior policies have no safety guarantee and may lead to severe damage during online executions. In this paper, to address the challenge of reducing variance while ensuring safety simultaneously, we propose an optimal variance-minimizing behavior policy under safety constraints. Theoretically, while ensuring safety constraints, our evaluation method is unbiased and has lower variance than on-policy evaluation. Empirically, our method is the only existing method to achieve both substantial variance reduction and safety constraint satisfaction. Furthermore, we show our method is even superior to previous methods in both variance reduction and execution safety.
****

Ganzhao Yuan
ADMM for Structured Fractional Minimization
This paper considers a class of structured fractional minimization problems. The numerator consists of a differentiable function, a simple nonconvex nonsmooth function, a concave nonsmooth function, and a convex nonsmooth function composed

with a linear operator. The denominator is a continuous function that is either weakly convex or has a weakly convex square root. These problems are prevalent in various important applications in machine learning and data science. Existing methods, primarily based on subgradient methods and smoothing proximal gradient methods, often suffer from slow convergence and numerical stability issues. In this paper, we introduce {\sf FADMM}, the first Alternating Direction Method of Multipliers tailored for this class of problems. {\sf FADMM} decouples the original problem into linearized proximal subproblems, featuring two variants: one using Dinkelbach's parametric method ({\sf FADMM-D}) and the other using the quadratic transform method ({\sf FADMM-Q}). By introducing a novel Lyapunov function, we establish that {\sf FADMM} converges to ϵ-approximate critical points of the problem within an oracle complexity of $\mathcal{O}(1/\epsilon^{3})$. Extensive experiments on synthetic and real-world datasets, including sparse Fisher discriminant analysis, robust Sharpe ratio minimization, and robust sparse recovery, demonstrate the effectiveness of our approach.
****
Kevis-kokitsi Maninis,Kaifeng Chen,Soham Ghosh,Arjun Karpur,Koert Chen,Ye Xia,Bingyi Cao,Daniel Salz,Guangxing Han,Jan Dlabal,Dan Gnanapragasam,Mojtaba Seyedhosseini,Howard Zhou,Andre Araujo
TIPS: Text-Image Pretraining with Spatial awareness
While image-text representation learning has become very popular in recent years, existing models tend to lack spatial awareness and have limited direct applicability for dense understanding tasks.
For this reason, self-supervised image-only pretraining is still the go-to method for many dense vision applications (e.g. depth estimation, semantic segmentation), despite the lack of explicit supervisory signals.
In this paper, we close this gap between image-text and self-supervised learning, by proposing a novel general-purpose image-text model, which can be effectively used off the shelf for dense and global vision tasks.
Our method, which we refer to as Text-Image Pretraining with Spatial awareness (TIPS), leverages two simple and effective insights.
First, on textual supervision: we reveal that replacing noisy web image captions by synthetically generated textual descriptions boosts dense understanding performance significantly, due to a much richer signal for learning spatially aware representations.
We propose an adapted training method that combines noisy and synthetic captions, resulting in improvements across both dense and global understanding tasks.
Second, on the learning technique: we propose to combine contrastive image-text learning with self-supervised masked image modeling, to encourage spatial coherence, unlocking substantial enhancements for downstream applications.
Building on these two ideas, we scale our model using the transformer architecture, trained on a curated set of public images.
Our experiments are conducted on 8 tasks involving 16 datasets in total, demonstrating strong off-the-shelf performance on both dense and global understanding, for several image-only and image-text tasks.
Code and models are released at https://github.com/google-deepmind/tips .
****
Omar Chehab,Anna Korba,Austin J Stromme,Adrien Vacher
Provable Convergence and Limitations of Geometric Tempering for Langevin Dynamics
Geometric tempering is a popular approach to sampling from challenging multi-modal probability distributions by instead sampling from a sequence of distributions which interpolate, using the geometric mean, between an easier proposal distribution and the target distribution. In this paper, we theoretically investigate the soundness of this approach when the sampling algorithm is Langevin dynamics, proving both upper and lower bounds. Our upper bounds are the first analysis in the literature under functional inequalities. They assert the convergence of tempered Langevin in continuous and discrete-time, and their minimization leads to closed-form optimal tempering schedules for some pairs of proposal and target distributions. Our lower bounds demonstrate a simple case where the geometric tem

pering takes exponential time, and further reveal that the geometric tempering c
an suffer from poor functional inequalities and slow convergence, even when the
target distribution is well-conditioned. Overall, our results indicate that the
geometric tempering may not help, and can even be harmful for convergence.
****

Pierre Marion,Raphaël Berthier,Gérard Biau,Claire Boyer
Attention layers provably solve single-location regression
Attention-based models, such as Transformer, excel across various tasks but lack
 a comprehensive theoretical understanding, especially regarding token-wise spar
sity and internal linear representations. To address this gap, we introduce the
single-location regression task, where only one token in a sequence determines t
he output, and its position is a latent random variable, retrievable via a linea
r projection of the input. To solve this task, we propose a dedicated predictor,
 which turns out to be a simplified version of a non-linear self-attention layer
. We study its theoretical properties, by showing its asymptotic Bayes optimalit
y and analyzing its training dynamics. In particular, despite the non-convex nat
ure of the problem, the predictor effectively learns the underlying structure. T
his work highlights the capacity of attention mechanisms to handle sparse token
information and internal linear structures.
****

Hyunji Jung,Hanseul Cho,Chulhee Yun
Convergence and Implicit Bias of Gradient Descent on Continual Linear Classifica
tion
We study continual learning on multiple linear classification tasks by sequentia
lly running gradient descent (GD) for a fixed budget of iterations per each give
n task. When all tasks are jointly linearly separable and are presented in a cyc
lic/random order, we show the directional convergence of the trained linear clas
sifier to the joint (offline) max-margin solution. This is surprising because GD
 training on a single task is implicitly biased towards the individual max-margi
n solution for the task, and the direction of the joint max-margin solution can
be largely different from these individual solutions. Additionally, when tasks a
re given in a cyclic order, we present a non-asymptotic analysis on cycle-averag
ed forgetting, revealing that (1) alignment between tasks is indeed closely tied
 to catastrophic forgetting and backward knowledge transfer and (2) the amount o
f forgetting vanishes to zero as the cycle repeats. Lastly, we analyze the case
where the tasks are no longer jointly separable and show that the model trained
in a cyclic order converges to the unique minimum of the joint loss function.
****

Yuchen Zhu,Tianrong Chen,Lingkai Kong,Evangelos Theodorou,Molei Tao
Trivialized Momentum Facilitates Diffusion Generative Modeling on Lie Groups
The generative modeling of data on manifolds is an important task, for which dif
fusion models in flat spaces typically need nontrivial adaptations. This article
 demonstrates how a technique called `trivialization' can transfer the effective
ness of diffusion models in Euclidean spaces to Lie groups. In particular, an au
xiliary momentum variable was algorithmically introduced to help transport the p
osition variable between data distribution and a fixed, easy-to-sample distribut
ion. Normally, this would incur further difficulty for manifold data because mom
entum lives in a space that changes with the position. However, our trivializati
on technique creates a new momentum variable that stays in a simple fixed vector
 space. This design, together with a manifold preserving integrator, simplifies
implementation and avoids inaccuracies created by approximations such as project
ions to tangent space and manifold, which were typically used in prior work, hen
ce facilitating generation with high-fidelity and efficiency. The resulting meth
od achieves state-of-the-art performance on protein and RNA torsion angle genera
tion and sophisticated torus datasets. We also, arguably for the first time, tac
kle the generation of data on high-dimensional Special Orthogonal and Unitary gr
oups, the latter essential for quantum problems. Code is available at https://gi
thub.com/yuchen-zhu-zyc/TDM.
****

Calarina Muslimani,Matthew E. Taylor

Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning
To create useful reinforcement learning (RL) agents, step zero is to design a suitable reward function that captures the nuances of the task. However, reward engineering can be a difficult and time-consuming process.
Instead, human-in-the-loop RL methods hold the promise of learning reward functions from human feedback. Despite recent successes, many of the human-in-the-loop RL methods still require numerous human interactions to learn successful reward functions.
To improve the feedback efficiency of human-in-the-loop RL methods (i.e., require less human interaction), this paper introduces Sub-optimal Data Pre-training, SDP, an approach that leverages reward-free, sub-optimal data to improve scalar- and preference-based RL algorithms. In SDP, we start by pseudo-labeling all low-quality data with the minimum environment reward. Through this process, we obtain reward labels
to pre-train our reward model without requiring human labeling or preferences.
This pre-training phase provides the reward model a head start in learning, enabling it to recognize that low-quality transitions should be assigned low rewards. Through extensive experiments with both simulated and human teachers, we find that SDP can at least meet, but often significantly improve, state of the art human-in-the-loop RL performance across a variety of simulated robotic tasks.
****
Liqiang Jing,Zhehui Huang,Xiaoyang Wang,Wenlin Yao,Wenhao Yu,Kaixin Ma,Hongming Zhang,Xinya Du,Dong Yu
DSBench: How Far Are Data Science Agents from Becoming Data Science Experts?
Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) have demonstrated impressive language/vision reasoning abilities, igniting the recent trend of building agents for targeted applications such as shopping assistants or AI software engineers. Recently, many data science benchmarks have been proposed to investigate their performance in the data science domain. However, existing data science benchmarks still fall short when compared to real-world data science applications due to their simplified settings. To bridge this gap, we introduce DSBench, a comprehensive benchmark designed to evaluate data science agents with realistic tasks. This benchmark includes 466 data analysis tasks and 74 data modeling tasks, sourced from Eloquence and Kaggle competitions. DSBench offers a realistic setting by encompassing long contexts, multimodal task backgrounds, reasoning with large data files and multi-table structures, and performing end-to-end data modeling tasks. Our evaluation of state-of-the-art LLMs, LVLMs, and agents shows that they struggle with most tasks, with the best agent solving only 34.12% of data analysis tasks and achieving a 34.74% Relative Performance Gap (RPG). These findings underscore the need for further advancements in developing more practical, intelligent, and autonomous data science agents.
****
Francesco Insulla,Shuo Huang,Lorenzo Rosasco
Towards a learning theory of representation alignment
It has recently been argued that AI models' representations are becoming aligned as their scale and performance increase. Empirical analyses have been designed to support this idea and conjecture the possible alignment of different representations toward a shared statistical model of reality. In this paper, we propose a learning-theoretic perspective to representation alignment. First, we review and connect different notions of alignment based on metric, probabilistic, and spectral ideas. Then, we focus on stitching, a particular approach to understanding the interplay between different representations in the context of a task. Our main contribution here is to relate the properties of stitching to the kernel alignment of the underlying representation. Our results can be seen as a first step toward casting representation alignment as a learning-theoretic problem.
****
Eliot Xing,Vernon Luk,Jean Oh
Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation
Recent advances in GPU-based parallel simulation have enabled practitioners to collect large amounts of data and train complex control policies using deep reinf

orcement learning (RL), on commodity GPUs. However, such successes for RL in robotics have been limited to tasks sufficiently simulated by fast rigid-body dynamics. Simulation techniques for soft bodies are comparatively several orders of magnitude slower, thereby limiting the use of RL due to sample complexity requirements. To address this challenge, this paper presents both a novel RL algorithm and a simulation platform to enable scaling RL on tasks involving rigid bodies and deformables. We introduce Soft Analytic Policy Optimization (SAPO), a maximum entropy first-order model-based actor-critic RL algorithm, which uses first-order analytic gradients from differentiable simulation to train a stochastic actor to maximize expected return and entropy. Alongside our approach, we develop Rewarped, a parallel differentiable multiphysics simulation platform that supports simulating various materials beyond rigid bodies. We re-implement challenging manipulation and locomotion tasks in Rewarped, and show that SAPO outperforms baselines over a range of tasks that involve interaction between rigid bodies, articulations, and deformables. Additional details at https://rewarped.github.io/.
****

Abdolmehdi Behroozi,Chaopeng Shen,Daniel Kifer

Sensitivity-Constrained Fourier Neural Operators for Forward and Inverse Problems in Parametric Differential Equations

Parametric differential equations of the form $\frac{\partial u}{\partial t} = f(u, x, t, p)$ are fundamental in science and engineering. While deep learning frameworks like the Fourier Neural Operator (FNO) efficiently approximate differential equation solutions, they struggle with inverse problems, sensitivity calculations $\frac{\partial u}{\partial p}$, and concept drift. We address these challenges by introducing a novel sensitivity loss regularizer, demonstrated through Sensitivity-Constrained Fourier Neural Operators (SC-FNO). Our approach maintains high accuracy for solution paths and outperforms both standard FNO and FNO with Physics-Informed Neural Network regularization. SC-FNO exhibits superior performance in parameter inversion tasks, accommodates more complex parameter spaces (tested with up to 82 parameters), reduces training data requirements, and decreases training time while maintaining accuracy. These improvements apply across various differential equations and neural operators, enhancing their reliability without significant computational overhead (30%–130% extra training time per epoch). Models and selected experiment code are available at: https://github.com/AMBehroozi/SC_Neural_Operators.
****

Ruben Wiedemann,Antoine Jacquier,Lukas Gonon

Operator Deep Smoothing for Implied Volatility

We devise a novel method for nowcasting implied volatility based on neural operators.

Better known as implied volatility smoothing in the financial industry, nowcasting of implied volatility means constructing a smooth surface that is consistent with the prices presently observed on a given option market.

Option price data arises highly dynamically in ever-changing spatial configurations, which poses a major limitation to foundational machine learning approaches using classical neural networks.

While large models in language and image processing deliver breakthrough results on vast corpora of raw data, in financial engineering the generalization from big historical datasets has been hindered by the need for considerable data pre-processing.

In particular, implied volatility smoothing has remained an instance-by-instance, hands-on process both for neural network-based and traditional parametric strategies.

Our general *operator deep smoothing* approach, instead, directly maps observed data to smoothed surfaces.

We adapt the graph neural operator architecture to do so with high accuracy on ten years of raw intraday S&P 500 options data, using a single model instance.

The trained operator adheres to critical no-arbitrage constraints and is robust with respect to subsampling of inputs (occurring in practice in the context of o

utlier removal).
We provide extensive historical benchmarks and showcase the generalization capab
ility of our approach in a comparison with classical neural networks and SVI, an
 industry standard parametrization for implied volatility.
The operator deep smoothing approach thus opens up the use of neural networks on
 large historical datasets in financial engineering.
****
Jung-hun Kim,Min-hwan Oh
Dynamic Assortment Selection and Pricing with Censored Preference Feedback
In this study, we investigate the problem of dynamic multi-product selection and
 pricing by introducing a novel framework based on a *censored multinomial logit
* (C-MNL) choice model. In this model, sellers present a set of products with pr
ices, and buyers filter out products priced above their valuation, purchasing at
 most one product from the remaining options based on their preferences. The goa
l is to maximize seller revenue by dynamically adjusting product offerings and p
rices, while learning both product valuations and buyer preferences through purc
hase feedback. To achieve this, we propose a Lower Confidence Bound (LCB) pricin
g strategy. By combining this pricing strategy with either an Upper Confidence B
ound (UCB) or Thompson Sampling (TS) product selection approach, our algorithms
achieve regret bounds of $\tilde{O}(d^{\frac{3}{2}}\sqrt{T/\kappa})$ and $\tilde
{O}(d^{2}\sqrt{T/\kappa})$, respectively. Finally, we validate the performance o
f our methods through simulations, demonstrating their effectiveness.
****
Stephen Zhang,Vardan Papyan
OATS: Outlier-Aware Pruning Through Sparse and Low Rank Decomposition
The recent paradigm shift to large-scale foundation models has brought about a n
ew era for deep learning that, while has found great success in practice, has al
so been plagued by prohibitively expensive costs in terms of high memory consump
tion and compute. To mitigate these issues, there has been a concerted effort in
 post-hoc neural network pruning techniques that do not require costly retrainin
g. Despite the considerable progress being made, existing methods often exhibit
a steady drop in model performance as the compression increases. In this paper,
we present a novel approach to compressing large transformers, coined OATS, that
 compresses the model weights by approximating each weight matrix as the sum of
a sparse matrix and a low-rank matrix. Prior to the decomposition, the weights a
re first scaled by the second moment of their input embeddings, so as to ensure
the preservation of outlier features recently observed in large transformer mod
els. Without retraining, OATS achieves state-of-the-art performance when compres
sing large language models, such as Llama-3 and Phi-3, and vision transformers,
such as Google's ViT and DINOv2, by up to $60\\%$, all while speeding up the mod
el's inference on a CPU by up to $1.37\times$ compared to prior pruning methods.
****
Anish Abhijit Diwan,Julen Urain,Jens Kober,Jan Peters
Noise-conditioned Energy-based Annealed Rewards (NEAR): A Generative Framework f
or Imitation Learning from Observation
This paper introduces a new imitation learning framework based on energy-based g
enerative models capable of learning complex, physics-dependent, robot motion po
licies through state-only expert motion trajectories. Our algorithm, called Nois
e-conditioned Energy-based Annealed Rewards (NEAR), constructs several perturbed
 versions of the expert's motion data distribution and learns smooth, and well-d
efined representations of the data distribution's energy function using denoisin
g score matching. We propose to use these learnt energy functions as reward func
tions to learn imitation policies via reinforcement learning. We also present a
strategy to gradually switch between the learnt energy functions, ensuring that
the learnt rewards are always well-defined in the manifold of policy-generated s
amples. We evaluate our algorithm on complex humanoid tasks such as locomotion a
nd martial arts and compare it with state-only adversarial imitation learning al
gorithms like Adversarial Motion Priors (AMP). Our framework sidesteps the optim
isation challenges of adversarial imitation learning techniques and produces res
ults comparable to AMP in several quantitative metrics across multiple imitation

settings.
****
Jinlong Pang,Jiaheng Wei,Ankit Shah,Zhaowei Zhu,Yaxuan Wang,Chen Qian,Yang Liu,Yujia Bao,Wei Wei
Improving Data Efficiency via Curating LLM-Driven Rating Systems
Instruction tuning is critical for adapting large language models (LLMs) to downstream tasks, and recent studies have demonstrated that small amounts of human-curated data can outperform larger datasets, challenging traditional data scaling laws. While LLM-based data quality rating systems offer a cost-effective alternative to human annotation, they often suffer from inaccuracies and biases, even in powerful models like GPT-4. In this work, we introduce DS^2, a **D**iversity-aware **S**core curation method for **D**ata **S**election. By systematically modeling error patterns through a score transition matrix, DS^2 corrects LLM-based scores and promotes diversity in the selected data samples. Our approach shows that a curated subset (just 3.3\% of the original dataset) outperforms full-scale datasets (300k samples) across various machine-alignment benchmarks, and matches or surpasses human-aligned datasets such as LIMA with the same sample size (1k samples). These findings challenge conventional data scaling assumptions, highlighting that redundant, low-quality samples can degrade performance and reaffirming that ``more can be less''.
****
Minkyoung Cho,Yulong Cao,Jiachen Sun,Qingzhao Zhang,Marco Pavone,Jeong Joon Park,Heng Yang,Zhuoqing Mao
Cocoon: Robust Multi-Modal Perception with Uncertainty-Aware Sensor Fusion
An important paradigm in 3D object detection is the use of multiple modalities to enhance accuracy in both normal and challenging conditions, particularly for long-tail scenarios. To address this, recent studies have explored two directions of adaptive approaches: MoE-based adaptive fusion, which struggles with uncertainties arising from distinct object configurations, and late fusion for output-level adaptive fusion, which relies on separate detection pipelines and limits comprehensive understanding. In this work, we introduce Cocoon, an object- and feature-level uncertainty-aware fusion framework. The key innovation lies in uncertainty quantification for heterogeneous representations, enabling fair comparison across modalities through the introduction of a feature aligner and a learnable surrogate ground truth, termed feature impression. We also define a training objective to ensure that their relationship provides a valid metric for uncertainty quantification. Cocoon consistently outperforms existing static and adaptive methods in both normal and challenging conditions, including those with natural and artificial corruptions. Furthermore, we show the validity and efficacy of our uncertainty metric across diverse datasets.
****
Sihyun Yu,Sangkyung Kwak,Huiwon Jang,Jongheon Jeong,Jonathan Huang,Jinwoo Shin,Saining Xie
Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think
Recent studies have shown that the denoising process in (generative) diffusion models can induce meaningful (discriminative) representations inside the model, though the quality of these representations still lags behind those learned through recent self-supervised learning methods. We argue that one main bottleneck in training large-scale diffusion models for generation lies in effectively learning these representations. Moreover, training can be made easier by incorporating high-quality external visual representations, rather than relying solely on the diffusion models to learn them independently. We study this by introducing a straightforward regularization called REPresentation Alignment (REPA), which aligns the projections of noisy input hidden states in denoising networks with clean image representations obtained from external, pretrained visual encoders. The results are striking: our simple strategy yields significant improvements in both training efficiency and generation quality when applied to popular diffusion and flow-based transformers, such as DiTs and SiTs. For instance, our method can speed up SiT training by over 17.5$\times$, matching the performance (without clas

sifier-free guidance) of a SiT-XL model trained for 7M steps in less than 400K s
teps. In terms of final generation quality, our approach achieves state-of-the-a
rt results of FID=1.42 using classifier-free guidance with the guidance interval
.
****
Jen-tse Huang,Eric John Li,Man Ho LAM,Tian Liang,Wenxuan Wang,Youliang Yuan,Wenx
iang Jiao,Xing Wang,Zhaopeng Tu,Michael Lyu
Competing Large Language Models in Multi-Agent Gaming Environments
Decision-making is a complex process requiring diverse abilities, making it an e
xcellent framework for evaluating Large Language Models (LLMs). Researchers have
 examined LLMs' decision-making through the lens of Game Theory. However, existi
ng evaluation mainly focus on two-player scenarios where an LLM competes against
 another. Additionally, previous benchmarks suffer from test set leakage due to
their static design. We introduce GAMA(γ)-Bench, a new framework for eval
uating LLMs' Gaming Ability in Multi-Agent environments. It includes eight class
ical game theory scenarios and a dynamic scoring scheme specially designed to qu
antitatively assess LLMs' performance. γ-Bench allows flexible game setti
ngs and adapts the scoring system to different game parameters, enabling compreh
ensive evaluation of robustness, generalizability, and strategies for improvemen
t. Our results indicate that GPT-3.5 demonstrates strong robustness but limited
generalizability, which can be enhanced using methods like Chain-of-Thought. We
also evaluate 13 LLMs from 6 model families, including GPT-3.5, GPT-4, Gemini, L
LaMA-3.1, Mixtral, and Qwen-2. Gemini-1.5-Pro outperforms others, scoring of $69
.8$ out of 100, followed by LLaMA-3.1-70B (65.9) and Mixtral-8x22B (62.4).
 Our code and experimental results are publicly available at https://github.com/
CUHK-ARISE/GAMABench.
****
Indradyumna Roy,Eeshaan Jain,Soumen Chakrabarti,Abir De
Clique Number Estimation via Differentiable Functions of Adjacency Matrix Permut
ations
Estimating the clique number in a graph is central to various applications, e.g.
, community detection, graph retrieval, etc.
Existing estimators often rely on non-differentiable combinatorial components. H
ere, we propose a full differentiable estimator for clique number estimation, wh
ich can be trained from distant supervision of clique numbers, rather than demon
strating actual cliques.
Our key insight is a formulation of the maximum clique problem (MCP) as a maximi
zation of the size of fully dense square submatrix, within a suitably row-column
-permuted adjacency matrix.
We design a differentiable mechanism to search for permutations that lead to the
 discovery of such dense blocks.
However, the optimal permutation is not unique, which leads to the learning of s
purious permutations. To tackle this problem, we view the MCP problem as a seque
nce of subgraph matching tasks, each detecting progressively larger cliques in a
 nested manner. This allows effective navigation through suitable node permutati
ons.
These steps result in MxNet, an end-to-end differentiable model, which learns to
 predict clique number without explicit clique demonstrations, with the added be
nefit of interpretability. Experiments on eight datasets show the superior accu
racy of our approach.
****
Yuzhou Gu,Zhao Song,Lichen Zhang
Faster Algorithms for Structured Linear and Kernel Support Vector Machines
Quadratic programming is a ubiquitous prototype in convex programming. Many mach
ine learning problems can be formulated as quadratic programming, including the
famous Support Vector Machines (SVMs). Linear and kernel SVMs have been among th
e most popular models in machine learning over the past three decades, prior to
the deep learning era.

Generally, a quadratic program has an input size of $\Theta(n^2)$, where n is

the number of variables. Assuming the Strong Exponential Time Hypothesis ($\textsf{SETH}$), it is known that no $O(n^{2-o(1)})$ time algorithm exists when the quadratic objective matrix is positive semidefinite (Backurs, Indyk, and Schmidt, NeurIPS'17). However, problems such as SVMs usually admit much smaller input sizes: one is given n data points, each of dimension d, and d is oftentimes much smaller than n. Furthermore, the SVM program has only $O(1)$ equality linear constraints. This suggests that faster algorithms are feasible, provided the program exhibits certain structures.

In this work, we design the first nearly-linear time algorithm for solving quadratic programs whenever the quadratic objective admits a low-rank factorization, and the number of linear constraints is small. Consequently, we obtain results for SVMs:

* For linear SVM when the input data is d-dimensional, our algorithm runs in time $\widetilde O(nd^{(\omega+1)/2}\log(1/\epsilon))$ where $\omega\approx 2.37$ is the fast matrix multiplication exponent;

* For Gaussian kernel SVM, when the data dimension $d = O(\log n)$ and the squared dataset radius is sub-logarithmic in n, our algorithm runs in time $O(n^{1+o(1)}\log(1/\epsilon))$. We also prove that when the squared dataset radius is at least $\Omega(\log^2 n)$, then $\Omega(n^{2-o(1)})$ time is required. This improves upon the prior best lower bound in both the dimension d and the squared dataset radius.
****
Amir Mohammad Karimi Mamaghan,Samuele Papa,Karl Henrik Johansson,Stefan Bauer,Andrea Dittadi
Exploring the Effectiveness of Object-Centric Representations in Visual Question Answering: Comparative Insights with Foundation Models
Object-centric (OC) representations, which model visual scenes as compositions of discrete objects, have the potential to be used in various downstream tasks to achieve systematic compositional generalization and facilitate reasoning. However, these claims have yet to be thoroughly validated empirically.
Recently, foundation models have demonstrated unparalleled capabilities across diverse domains, from language to computer vision, positioning them as a potential cornerstone of future research for a wide range of computational tasks.
In this paper, we conduct an extensive empirical study on representation learning for downstream Visual Question Answering (VQA), which requires an accurate compositional understanding of the scene.
We thoroughly investigate the benefits and trade-offs of OC models and alternative approaches including large pre-trained foundation models on both synthetic and real-world data, ultimately identifying a promising path to leverage the strengths of both paradigms.
The extensiveness of our study, encompassing over 600 downstream VQA models and 15 different types of upstream representations, also provides several additional insights that we believe will be of interest to the community at large.
****
Jianhua Sun,Yuxuan Li,Liang Chai,Cewu Lu
Interactive Adjustment for Human Trajectory Prediction with Individual Feedback
Human trajectory prediction is fundamental for autonomous driving and service robot. The research community has studied various important aspects of this task and made remarkable progress recently. However, there is an essential perspective which is not well exploited in previous research all along, namely individual feedback. Individual feedback exists in the sequential nature of trajectory prediction, where earlier predictions of a target can be verified over time by his ground-truth trajectories to obtain feedback which provides valuable experience for subsequent predictions on the same agent. In this paper, we show such feedback can reveal the strengths and weaknesses of the model's predictions on a specific target and heuristically guide to deliver better predictions on him. We present an interactive adjustment network to effectively model and leverage the feedba

ck. This network first exploits the feedback from previous predictions to dynami
cally generate an adjuster which then interactively makes appropriate adjustment
s to current predictions for more accurate ones. We raise a novel displacement e
xpectation loss to train this interactive architecture. Through experiments on r
epresentative prediction methods and widely-used benchmarks, we demonstrate the
great value of individual feedback and the superior effectiveness of proposed in
teractive adjustment network.
****
Yihang Chen,Qianyi Wu,Mengyao Li,Weiyao Lin,Mehrtash Harandi,Jianfei Cai
Fast Feedforward 3D Gaussian Splatting Compression
With 3D Gaussian Splatting (3DGS) advancing real-time and high-fidelity renderin
g for novel view synthesis, storage requirements pose challenges for their wides
pread adoption. Although various compression techniques have been proposed, prev
ious art suffers from a common limitation: for any existing 3DGS, per-scene opti
mization is needed to achieve compression, making the compression sluggish and s
low. To address this issue, we introduce Fast Compression of 3D Gaussian Splatti
ng (FCGS), an optimization-free model that can compress 3DGS representations rap
idly in a single feed-forward pass, which significantly reduces compression time
 from minutes to seconds. To enhance compression efficiency, we propose a multi-
path entropy module that assigns Gaussian attributes to different entropy constr
aint paths for balance between size and fidelity. We also carefully design both
inter- and intra-Gaussian context models to remove redundancies among the unstru
ctured Gaussian blobs. Overall, FCGS achieves a compression ratio of over 20X wh
ile maintaining fidelity, surpassing most per-scene SOTA optimization-based meth
ods. Code: github.com/YihangChen-ee/FCGS.
****
Gabriel Wu,Jacob Hilton
Estimating the Probabilities of Rare Outputs in Language Models
We consider the problem of *low probability estimation*: given a machine learnin
g model and a formally-specified input distribution, how can we estimate the pro
bability of a binary property of the model's output, even when that probability
is too small to estimate by random sampling? This problem is motivated by the ne
ed to improve worst-case performance, which distribution shift can make much mor
e likely. We study low probability estimation in the context of argmax sampling
from small transformer language models. We compare two types of methods: importa
nce sampling, which involves searching for inputs giving rise to the rare output
, and activation extrapolation, which involves extrapolating a probability distr
ibution fit to the model's logits. We find that importance sampling outperforms
activation extrapolation, but both outperform naive sampling. Finally, we explai
n how minimizing the probability estimate of an undesirable behavior generalizes
 adversarial training, and argue that new methods for low probability estimation
 are needed to provide stronger guarantees about worst-case performance.
****
Yizhi Song,Liu He,Zhifei Zhang,Soo Ye Kim,He Zhang,Wei Xiong,Zhe Lin,Brian L. Pr
ice,Scott Cohen,Jianming Zhang,Daniel Aliaga
Refine-by-Align: Reference-Guided Artifacts Refinement through Semantic Alignmen
t
Personalized image generation has emerged from the recent advancements in genera
tive models. However, these generated personalized images often suffer from loca
lized artifacts such as incorrect logos, reducing fidelity and fine-grained iden
tity details of the generated results. Furthermore, there is little prior work t
ackling this problem. To help improve these identity details in the personalized
 image generation, we introduce a new task: reference-guided artifacts refinemen
t. We present Refine-by-Align, a first-of-its-kind model that employs a diffusio
n-based framework to address this challenge. Our model consists of two stages: A
lignment Stage and Refinement Stage, which share weights of a unified neural net
work model. Given a generated image, a masked artifact region, and a reference i
mage, the alignment stage identifies and extracts the corresponding regional fea
tures in the reference, which are then used by the refinement stage to fix the
artifacts. Our model-agnostic pipeline requires no test-time tuning or optimizat

ion. It automatically enhances image fidelity and reference identity in the gene
rated image, generalizing well to existing models on various tasks including but
 not limited to customization, generative compositing, view synthesis, and virtu
al try-on. Extensive experiments and comparisons demonstrate that our pipeline g
reatly pushes the boundary of fine details in the image synthesis models.
****

Lequan Lin,Dai Shi,Andi Han,Zhiyong Wang,Junbin Gao

Diffusing to the Top: Boost Graph Neural Networks with Minimal Hyperparameter Tu
ning

Graph Neural Networks (GNNs) are proficient in graph representation learning and
 achieve promising performance on versatile tasks such as node classification an
d link prediction.
Usually, a comprehensive hyperparameter tuning is essential for fully unlocking
GNN's top performance, especially for complicated tasks such as node classificat
ion on large graphs and long-range graphs. This is usually associated with high
computational and time costs and careful design of appropriate search spaces.
This work introduces a graph-conditioned latent diffusion framework (GNN-Diff) t
o generate high-performing GNNs based on the model checkpoints of sub-optimal hy
perparameters selected by a light-tuning coarse search. We validate our method t
hrough 166 experiments across four graph tasks: node classification on small, la
rge, and long-range graphs, as well as link prediction. Our experiments involve
10 classic and state-of-the-art target models and 20 publicly available datasets
. The results consistently demonstrate that GNN-Diff: (1) boosts the performance
 of GNNs with efficient hyperparameter tuning; and (2) presents high stability a
nd generalizability on unseen data across multiple generation runs. The code is
available at https://github.com/lequanlin/GNN-Diff.
****

Xuannan Liu,Zekun Li,Pei Pei Li,Huaibo Huang,Shuhan Xia,Xing Cui,Linzhi Huang,We
ihong Deng,Zhaofeng He

MMFakeBench: A Mixed-Source Multimodal Misinformation Detection Benchmark for LV
LMs

Current multimodal misinformation detection (MMD) methods often assume a single
source and type of forgery for each sample, which is insufficient for real-world
 scenarios where multiple forgery sources coexist. The lack of a benchmark for m
ixed-source misinformation has hindered progress in this field. To address this,
 we introduce MMFakeBench, the first comprehensive benchmark for mixed-source MM
D. MMFakeBench includes 3 critical sources: textual veracity distortion, visual
veracity distortion, and cross-modal consistency distortion, along with 12 sub-c
ategories of misinformation forgery types. We further conduct an extensive evalu
ation of 6 prevalent detection methods and 15 Large Vision-Language Models (LVLM
s) on MMFakeBench under a zero-shot setting. The results indicate that current m
ethods struggle under this challenging and realistic mixed-source MMD setting. A
dditionally, we propose MMD-Agent, a novel approach to integrate the reasoning,
action, and tool-use capabilities of LVLM agents, significantly enhancing accura
cy and generalization. We believe this study will catalyze future research into
more realistic mixed-source multimodal misinformation and provide a fair evaluat
ion of misinformation detection methods.
****

Sanghyuk Chun,Wonjae Kim,Song Park,Sangdoo Yun

Probabilistic Language-Image Pre-Training

Vision-language models (VLMs) embed aligned image-text pairs into a joint space
but often rely on deterministic embeddings, assuming a one-to-one correspondence
 between images and texts. This oversimplifies real-world relationships, which a
re inherently many-to-many, with multiple captions describing a single image and
 vice versa. We introduce Probabilistic Language-Image Pre-training (ProLIP), th
e first probabilistic VLM pre-trained on a billion-scale image-text dataset usin
g only probabilistic objectives, achieving a strong zero-shot capability (e.g.,
74.6% ImageNet zero-shot accuracy with ViT-B/16). ProLIP efficiently estimates u
ncertainty by an ``uncertainty token'' without extra parameters. We also introdu
ce a novel inclusion loss that enforces distributional inclusion relationships b

etween image-text pairs and between original and masked inputs. Experiments demo
nstrate that, by leveraging uncertainty estimates, ProLIP benefits downstream ta
sks and aligns with intuitive notions of uncertainty, e.g., shorter texts being
more uncertain and more general inputs including specific ones. Utilizing text u
ncertainties, we further improve ImageNet accuracy from 74.6% to 75.8% (under a
few-shot setting), supporting the practical advantages of our probabilistic appr
oach. The code is available at https://github.com/naver-ai/prolip
****

Huan Ren,Wenfei Yang,Xiang Liu,Shifeng Zhang,Tianzhu Zhang
Learning Shape-Independent Transformation via Spherical Representations for Cate
gory-Level Object Pose Estimation
Category-level object pose estimation aims to determine the pose and size of nov
el objects in specific categories. Existing correspondence-based approaches typi
cally adopt point-based representations to establish the correspondences between
 primitive observed points and normalized object coordinates. However, due to th
e inherent shape-dependence of canonical coordinates, these methods suffer from
semantic incoherence across diverse object shapes. To resolve this issue, we inn
ovatively leverage the sphere as a shared proxy shape of objects to learn shape-
independent transformation via spherical representations. Based on this insight,
 we introduce a novel architecture called SpherePose, which yields precise corre
spondence prediction through three core designs. Firstly, We endow the point-wis
e feature extraction with SO(3)-invariance, which facilitates robust mapping bet
ween camera coordinate space and object coordinate space regardless of rotation
transformation. Secondly, the spherical attention mechanism is designed to propa
gate and integrate features among spherical anchors from a comprehensive perspec
tive, thus mitigating the interference of noise and incomplete point cloud. Last
ly, a hyperbolic correspondence loss function is designed to distinguish subtle
distinctions, which can promote the precision of correspondence prediction. Expe
rimental results on CAMERA25, REAL275 and HouseCat6D benchmarks demonstrate the
superior performance of our method, verifying the effectiveness of spherical rep
resentations and architectural innovations.
****

Putra Manggala,Atalanti A. Mastakouri,Elke Kirschbaum,Shiva Kasiviswanathan,Aadi
tya Ramdas
QA-Calibration of Language Model Confidence Scores
To use generative question-and-answering (QA) systems for decision-making and in
 any critical application, these systems need to provide well-calibrated confide
nce scores that reflect the correctness of their answers. Existing calibration m
ethods aim to ensure that the confidence score is, *on average*, indicative of t
he likelihood that the answer is correct. We argue, however, that this standard
 (average-case) notion of calibration is difficult to interpret for decision-mak
ing in generative QA. To address this, we generalize the standard notion of aver
age calibration and introduce QA-calibration, which ensures calibration holds ac
ross different question-and-answer groups. We then propose discretized posthoc c
alibration schemes for achieving QA-calibration. We establish distribution-free
guarantees on the performance of this method and validate our method on confiden
ce scores returned by elicitation prompts across multiple QA benchmarks and larg
e language models (LLMs).
****

Marine Le Morvan,Gael Varoquaux
Imputation for prediction: beware of diminishing returns.
Missing values are prevalent across various fields, posing challenges for traini
ng and deploying predictive models. In this context, imputation is a common prac
tice, driven by the hope that accurate imputations will enhance predictions. How
ever, recent theoretical and empirical studies indicate that simple constant imp
utation can be consistent and competitive. This empirical study aims at clarifyi
ng
if and *when* investing in advanced imputation methods yields significantly be
tter predictions. Relating imputation and predictive accuracies across combinati
ons of imputation and predictive models on 19 datasets, we show that imputation

accuracy matters less i) when using expressive models, ii) when incorporating mi
ssingness indicators as complementary inputs, iii) matters much more for generat
ed linear outcomes than for real-data outcomes. Interestingly, we also show that
 the use of the missingness indicator is beneficial to the prediction performanc
e, even in MCAR scenarios. Overall, on real-data with powerful models, imputatio
n quality has only a minor effect on prediction performance. Thus, investing in
better imputations for improved predictions often offers limited benefits.
****

Yushu Li,Yongyi Su,Adam Goodge,Kui Jia,Xun Xu

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free A
daptation of Vision-Language Model

Vision-language models (VLMs) have revolutionized machine learning by leveraging
 large pre-trained models to tackle various downstream tasks. Although label, tr
aining, and data efficiency have improved, many state-of-the-art VLMs still requ
ire task-specific hyperparameter tuning and fail to fully exploit test samples.
To overcome these challenges, we propose a graph-based approach for label-effici
ent adaptation and inference. Our method dynamically constructs a graph over tex
t prompts, few-shot examples, and test samples, using label propagation for infe
rence without task-specific tuning. Unlike existing zero-shot label propagation
techniques, our approach requires no additional unlabeled support set and effect
ively leverages the test sample manifold through dynamic graph expansion. We fur
ther introduce a context-aware feature re-weighting mechanism to improve task ad
aptation accuracy. Additionally, our method supports efficient graph expansion,
enabling real-time inductive inference. Extensive evaluations on downstream task
s, such as fine-grained categorization and out-of-distribution generalization, d
emonstrate the effectiveness of our approach. The source code is available at ht
tps://github.com/Yushu-Li/ECALP.
****

Muhammad A Shah,David Solans Noguero,Mikko A. Heikkilä,Bhiksha Raj,Nicolas Kourt
ellis

Speech Robust Bench: A Robustness Benchmark For Speech Recognition

As Automatic Speech Recognition (ASR) models become ever more pervasive, it is i
mportant to ensure that they make reliable predictions under corruptions present
 in the physical and digital world. We propose Speech Robust Bench (SRB), a com
prehensive benchmark for evaluating the robustness of ASR models to diverse corr
uptions. SRB is composed of 114 input perturbations which simulate an heterogene
ous range of corruptions that ASR models may encounter when deployed in the wild
. We use SRB to evaluate the robustness of several state-of-the-art ASR models a
nd observe that model size and certain modeling choices such as the use of discr
ete representations, or self-training appear to be conducive to robustness. We e
xtend this analysis to measure the robustness of ASR models on data from various
 demographic subgroups, namely English and Spanish speakers, and males and femal
es. Our results revealed noticeable disparities in the model's robustness across
 subgroups. We believe that SRB will significantly facilitate future research to
wards robust ASR models, by making it easier to conduct comprehensive and compar
able robustness evaluations.
****

Zeyu Li,Hongkun Dou,Shen Fang,Wang Han,Yue Deng,Lijun Yang

Physics-aligned field reconstruction with diffusion bridge

The reconstruction of physical fields from sparse measurements is pivotal in bot
h scientific research and engineering applications. Traditional methods are incr
easingly supplemented by deep learning models due to their efficacy in extractin
g features from data. However, except for the low accuracy on complex physical s
ystems, these models often fail to comply with essential physical constraints, s
uch as governing equations and boundary conditions. To overcome this limitation,
 we introduce a novel data-driven field reconstruction framework, termed the Phy
sics-aligned Schr\"{o}dinger Bridge (PalSB). This framework leverages a diffusio
n bridge mechanism that is specifically tailored to align with physical constrai
nts. The PalSB approach incorporates a dual-stage training process designed to a
ddress both local reconstruction mapping and global physical principles. Additio

nally, a boundary-aware sampling technique is implemented to ensure adherence to physical boundary conditions. We demonstrate the effectiveness of PalSB through its application to three complex nonlinear systems: cylinder flow from Particle Image Velocimetry experiments, two-dimensional turbulence, and a reaction-diffusion system. The results reveal that PalSB not only achieves higher accuracy but also exhibits enhanced compliance with physical constraints compared to existing methods. This highlights PalSB's capability to generate high-quality representations of intricate physical interactions, showcasing its potential for advancing field reconstruction techniques. The source code can be found at https://github.com/lzy12301/PalSB.
****

Minghui Chen,Ruinan Jin,Wenlong Deng,Yuanyuan Chen,Zhi Huang,Han Yu,Xiaoxiao Li
Can Textual Gradient Work in Federated Learning?
Recent studies highlight the promise of LLM-based prompt optimization, especially with TextGrad, which automates ``differentiation'' via texts and backpropagates textual feedback provided by LLMs. This approach facilitates training in various real-world applications that do not support numerical gradient propagation or loss calculation. It opens new avenues for optimization in decentralized, resource-constrained environments, suggesting that users of black-box LLMs (e.g., ChatGPT) could enhance components of LLM agentic systems (such as prompt optimization) through collaborative paradigms like federated learning (FL). In this paper, we systematically explore the potential and challenges of incorporating textual gradient into FL. Our contributions are fourfold.
Firstly, we introduce a novel FL paradigm, Federated Textual Gradient (FedTextGrad), that allows FL clients to upload their locally optimized prompts derived from textual gradients, while the FL server aggregates the received prompts through text summarization. Unlike traditional FL frameworks, which are designed for numerical aggregation, FedTextGrad is specifically tailored for handling textual data, expanding the applicability of FL to a broader range of problems that lack well-defined numerical loss functions.
Secondly, building on this design, we conduct extensive experiments to explore the feasibility of federated textual gradients. Our findings highlight the importance of properly tuning key factors (e.g., local steps) in FL training to effectively integrate textual gradients.
Thirdly, we highlight a major challenge in federated textual gradient aggregation: retaining essential information from distributed prompt updates. Concatenation often produces prompts that exceed the LLM API’s context window, while summarization can degrade performance by generating overly condensed or complex text that lacks key context.
Last but not least, in response to this issue, we improve the vanilla variant of FedTextGrad by providing actionable guidance to the LLM when summarizing client prompts by leveraging the Uniform Information Density principle. Such a design reduces the complexity of the aggregated global prompt, thereby better incentivizing the LLM's reasoning ability. Through this principled study, we enable the adoption of textual gradients in FL for optimizing LLMs, identify important issues, and pinpoint future directions, thereby opening up a new research area that warrants further investigation.
****

Yibo Yang,Justus Will,Stephan Mandt
Progressive Compression with Universally Quantized Diffusion Models
Diffusion probabilistic models have achieved mainstream success in many generative modeling tasks, from image generation to inverse problem solving. A distinct feature of these models is that they correspond to deep hierarchical latent variable models optimizing a variational evidence lower bound (ELBO) on the data likelihood. Drawing on a basic connection between likelihood modeling and compression, we explore the potential of diffusion models for progressive coding, resulting in a sequence of bits that can be incrementally transmitted and decoded with progressively improving reconstruction quality. Unlike prior work based on Gaussian diffusion or conditional diffusion models, we propose a new form of diffusion model with uniform noise in the forward process, whose negative ELBO correspon

ds to the end-to-end compression cost using universal quantization. We obtain pr
omising first results on image compression, achieving competitive rate-distortio
n-realism results on a wide range of bit-rates with a single model, bringing neu
ral codecs a step closer to practical deployment. Our code can be found at https
://github.com/mandt-lab/uqdm.
****

Haiyan Zhao,Heng Zhao,Bo Shen,Ali Payani,Fan Yang,Mengnan Du

Beyond Single Concept Vector: Modeling Concept Subspace in LLMs with Gaussian Di
stribution

Probing learned concepts in large language models (LLMs) is crucial for understa
nding how semantic knowledge is encoded internally. Training linear classifiers
on probing tasks is a principle approach to denote the vector of a certain conce
pt in the representation space. However, the single vector identified for a conc
ept varies with both data and training, making it less robust and weakening its
effectiveness in real-world applications. To address this challenge, we propose
an approach to approximate the subspace representing a specific concept. Built o
n linear probing classifiers, we extend the concept vectors into Gaussian Concep
t Subspace (GCS). We demonstrate GCS's effectiveness through measuring its faith
fulness and plausibility across multiple LLMs with different sizes and architect
ures. Additionally, we use representation intervention tasks to showcase its eff
icacy in real-world applications such as emotion steering. Experimental results
indicate that GCS concept vectors have the potential to balance steering perform
ance and maintaining the fluency in natural language generation tasks.
****

Chun Jung Chen,Chung-Chin Shih,Ti-Rong Wu

Strength Estimation and Human-Like Strength Adjustment in Games

Strength estimation and adjustment are crucial in designing human-AI interaction
s, particularly in games where AI surpasses human players. This paper introduces
 a novel strength system, including a *strength estimator* (SE) and an SE-based
Monte Carlo tree search, denoted as *SE-MCTS*, which predicts strengths from gam
es and offers different playing strengths with human styles. The strength estima
tor calculates strength scores and predicts ranks from games without direct huma
n interaction. SE-MCTS utilizes the strength scores in a Monte Carlo tree search
 to adjust playing strength and style. We first conduct experiments in Go, a cha
llenging board game with a wide range of ranks. Our strength estimator significa
ntly achieves over 80% accuracy in predicting ranks by observing 15 games only,
whereas the previous method reached 49% accuracy for 100 games. For strength adj
ustment, SE-MCTS successfully adjusts to designated ranks while achieving a 51.3
3% accuracy in aligning to human actions, outperforming a previous state-of-the-
art, with only 42.56% accuracy. To demonstrate the generality of our strength sy
stem, we further apply SE and SE-MCTS to chess and obtain consistent results. Th
ese results show a promising approach to strength estimation and adjustment, enh
ancing human-AI interactions in games. Our code is available at https://rlg.iis.
sinica.edu.tw/papers/strength-estimator.
****

Mohammadreza Pourreza,Hailong Li,Ruoxi Sun,Yeounoh Chung,Shayan Talaei,Gaurav Ta
rlok Kakkar,Yu Gan,Amin Saberi,Fatma Ozcan,Sercan O Arik

CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in
Text-to-SQL

We present CHASE-SQL, a novel framework addressing large language model (LLM) pe
rformance challenges for Text-to-SQL tasks by leveraging multi-agent modeling an
d test-time compute for improved candidate generation and selection. CHASE-SQL u
ses LLMs to generate diverse SQL candidates with: (1) a divide-and-conquer appro
ach to break down complex queries, (2) chain-of-thought reasoning based on query
 execution plans, and (3) instance-aware synthetic example generation for tailor
ed few-shot demonstrations. A selection agent ranks candidates via pairwise comp
arisons using a fine-tuned binary selection LLM, offering robust performance. Th
is framework improves SQL query quality and diversity, achieving state-of-the-ar
t execution accuracy of 73.0% on the BIRD Text-to-SQL benchmark test set, toppin
g the leaderboard at the time of submission.

****
Priya Kasimbeg,Frank Schneider,Runa Eschenhagen,Juhan Bae,Chandramouli Shama Sastry,Mark Saroufim,BOYUAN FENG,Less Wright,Edward Z. Yang,Zachary Nado,Sourabh Medapati,Philipp Hennig,Michael Rabbat,George E. Dahl
Accelerating neural network training: An analysis of the AlgoPerf competition
The goal of the AlgoPerf: Training Algorithms competition is to evaluate practical speed-ups in neural network training achieved solely by improving the underlying training algorithms. In the external tuning ruleset, submissions must provide workload-agnostic hyperparameter search spaces, while in the self-tuning ruleset they must be completely hyperparameter-free. In both rulesets, submissions are compared on time-to-result across multiple deep learning workloads, training on fixed hardware. This paper presents the inaugural AlgoPerf competition's results, which drew 18 diverse submissions from 10 teams. Our investigation reveals several key findings: (1) The winning submission in the external tuning ruleset, using Distributed Shampoo, demonstrates the effectiveness of non-diagonal preconditioning over popular methods like Adam, even when compared on wall-clock runtime. (2) The winning submission in the self-tuning ruleset, based on the Schedule Free AdamW algorithm, demonstrates a new level of effectiveness for completely hyperparameter-free training algorithms. (3) The top-scoring submissions were surprisingly robust to workload changes. We also discuss the engineering challenges encountered in ensuring a fair comparison between different training algorithms. These results highlight both the significant progress so far, and the considerable room for further improvements.
****
Aming WU,Cheng Deng
CFD: Learning Generalized Molecular Representation via Concept-Enhanced Feedback Disentanglement
To accelerate biochemical research, e.g., drug and protein discovery, molecular representation learning (MRL) has attracted much attention. However, most existing methods follow the closed-set assumption that training and testing data share identical distribution, which limits their generalization abilities in out-of-distribution (OOD) cases. In this paper, we explore designing a new disentangled mechanism for learning generalized molecular representation that exhibits robustness against distribution shifts. And an approach of Concept-Enhanced Feedback Disentanglement (CFD) is proposed, whose goal is to exploit the feedback mechanism to learn distribution-agnostic representation. Specifically, we first propose two dedicated variational encoders to separately decompose distribution-agnostic and spurious features. Then, a set of molecule-aware concepts are tapped to focus on invariant substructure characteristics. By fusing these concepts into the disentangled distribution-agnostic features, the generalization ability of the learned molecular representation could be further enhanced. Next, we execute iteratively the disentangled operations based on a feedback received from the previous output. Finally, based on the outputs of multiple feedback iterations, we construct a self-supervised objective to promote the variational encoders to possess the disentangled capability. In the experiments, our method is verified on multiple real-world molecular datasets. The significant performance gains over state-of-the-art baselines demonstrate that our method can effectively disentangle generalized molecular representation in the presence of various distribution shifts. The source code will be released at https://github.com/AmingWu/MoleculeCFD.
****
YOUHE JIANG,Ran Yan,Binhang Yuan
HexGen-2: Disaggregated Generative Inference of LLMs in Heterogeneous Environment
Disaggregating the prefill and decoding phases represents an effective new paradigm for generative inference of large language models (LLM). This approach offers some significant system advantages, such as eliminating prefill-decoding interference and optimizing resource allocation. However, it is still an challenging open problem about how to deploy the disaggregated inference paradigm across a group of heterogeneous GPUs, which can be an economic alternative of the deployment over the homogeneous high performance GPUs.

Towards this end, we introduce HexGen-2, a distributed system for high throughpu
t and cost-efficient LLM serving on heterogeneous GPUs following the disaggragat
ed paradigm. Built on top of HexGen, the core component of HexGen-2 is a sophist
icated scheduling algorithm that formalizes the allocation of disaggregated LLM
inference computations and communications over heterogeneous GPUs and network co
nnections as a constraint optimization problem. We leverage the graph partitioni
ng and max-flow algorithm to co-optimize resource allocation, parallel strategie
s for distinct inference phases, and the efficiency of inter-phase key-value (KV
) cache communications. We conduct extensive experiments to evaluate HexGen-2, i
.e., on OPT (30B) and Llama-2 (70B) models in various real-world settings, the r
esults reveal that HexGen-2 delivers up to a 2.0$\times$ and on average a 1.3$\t
imes$ improvement in serving throughput, reduces the average inference latency b
y 1.5$\times$ compared with state-of-the-art systems given the same price budget
, and achieves comparable inference performance with a 30% lower price budget.
****
Hai Zhang,Boyuan Zheng,Tianying Ji,JinHang Liu,Anqi Guo,Junqiao Zhao,Lanqing Li
Scrutinize What We Ignore: Reining In Task Representation Shift Of Context-Based
Offline Meta Reinforcement Learning
Offline meta reinforcement learning (OMRL) has emerged as a promising approach f
or interaction avoidance and strong generalization performance by leveraging pre
-collected data and meta-learning techniques.
Previous context-based approaches predominantly rely on the intuition that alter
nating optimization between the context encoder and the policy can lead to perfo
rmance improvements, as long as the context encoder follows the principle of max
imizing the mutual information between the task variable M and its latent repr
esentation Z ($I(Z;M)$) while the policy adopts the standard offline reinforce
ment learning (RL) algorithms conditioning on the learned task representation.
Despite promising results, the theoretical justification of performance improvem
ents for such intuition remains underexplored.
Inspired by the return discrepancy scheme in the model-based RL field, we find t
hat the previous optimization framework can be linked with the general RL object
ive of maximizing the expected return, thereby explaining performance improvemen
ts.
Furthermore, after scrutinizing this optimization framework, we observe that the
condition for monotonic performance improvements does not consider the variatio
n of the task representation. When these variations are considered, the previous
ly established condition may no longer be sufficient to ensure monotonicity, the
reby impairing the optimization process.
We name this issue \underline{task representation shift} and theoretically prove
that the monotonic performance improvements can be guaranteed with appropriate
context encoder updates.
We use different settings to rein in the task representation shift on three wide
ly adopted training objectives concerning maximizing $I(Z;M)$ across different d
ata qualities.
Empirical results show that reining in the task representation shift can indeed
improve performance.
Our work opens up a new avenue for OMRL, leading to a better understanding betwe
en the task representation and performance improvements.
****
Dmitry Kamzolov,Artem Agafonov,Dmitry Pasechnyuk,Alexander Gasnikov,Martin Taká■
OPTAMI: Global Superlinear Convergence of High-order Methods
Second-order methods for convex optimization outperform first-order methods in t
erms of theoretical iteration convergence, achieving rates up to $O(k^{-5})$ for
highly-smooth functions. However, their practical performance and applications
are limited due to their multi-level structure and implementation complexity. In
this paper, we present new results on high-order optimization methods, supporte
d by their practical performance. First, we show that the basic high-order metho
ds, such as the Cubic Regularized Newton Method, exhibit global superlinear conv
ergence for μ-strongly star-convex functions, a class that includes μ-st
rongly convex functions and some non-convex functions. Theoretical convergence r

esults are both inspired and supported by the practical performance of these met
hods. Secondly, we propose a practical version of the Nesterov Accelerated Tenso
r method, called NATA. It significantly outperforms the classical variant and ot
her high-order acceleration techniques in practice. The convergence of NATA is a
lso supported by theoretical results. Finally, we introduce an open-source compu
tational library for high-order methods, called OPTAMI. This library includes va
rious methods, acceleration techniques, and subproblem solvers, all implemented
as PyTorch optimizers, thereby facilitating the practical application of high-or
der methods to a wide range of optimization problems. We hope this library will
simplify research and practical comparison of methods beyond first-order.
****

Shanglin Li,Motoaki Kawanabe,Reinmar J Kobler
SPDIM: Source-Free Unsupervised Conditional and Label Shift Adaptation in EEG
The non-stationary nature of electroencephalography (EEG) introduces distributio
n shifts across domains (e.g., days and subjects), posing a significant challeng
e to EEG-based neurotechnology generalization.
Without labeled calibration data for target domains, the problem is a source-fre
e unsupervised domain adaptation (SFUDA) problem.
For scenarios with constant label distribution, Riemannian geometry-aware statis
tical alignment frameworks on the symmetric positive definite (SPD) manifold are
 considered state-of-the-art.
However, many practical scenarios, including EEG-based sleep staging, exhibit la
bel shifts.
Here, we propose a geometric deep learning framework for SFUDA problems under sp
ecific distribution shifts, including label shifts.
We introduce a novel, realistic generative model and show that prior Riemannian
statistical alignment methods on the SPD manifold can compensate for specific ma
rginal and conditional distribution shifts but hurt generalization under label s
hifts.
As a remedy, we propose a parameter-efficient manifold optimization strategy ter
med SPDIM.
SPDIM uses the information maximization principle to learn a single SPD-manifold
-constrained parameter per target domain.
In simulations, we demonstrate that SPDIM can compensate for the shifts under ou
r generative model.
Moreover, using public EEG-based brain-computer interface and sleep staging data
sets, we show that SPDIM outperforms prior approaches.
****

Xueru Wen,Jie Lou,Yaojie Lu,Hongyu Lin,XingYu,Xinyu Lu,Ben He,Xianpei Han,Debing
 Zhang,Le Sun
Rethinking Reward Model Evaluation: Are We Barking up the Wrong Tree?
Reward Models (RMs) are crucial for aligning language models with human preferen
ces.
Currently, the evaluation of RMs depends on measuring accuracy against a validat
ion set of manually annotated preference data.
Although this method is straightforward and widely adopted, the relationship bet
ween RM accuracy and downstream policy performance remains under-explored.
In this work, we conduct experiments in a synthetic setting to investigate how d
ifferences in RM measured by accuracy translate into gaps in optimized policy pe
rformance.
Our findings reveal that while there is a weak positive correlation between accu
racy and downstream performance, policies optimized towards RMs with similar acc
uracy can exhibit quite different performance.
Moreover, we discover that the way of measuring accuracy significantly impacts i
ts ability to predict the final policy performance.
Through the lens of the Regressional Goodhart effect, we recognize that accuracy
, when used for measuring RM quality, can fail to fully capture the potential RM
 overoptimization.
This underscores the inadequacy of relying solely on accuracy to reflect their i
mpact on policy optimization.

****
Mintong Kang,Bo Li
R^2-Guard: Robust Reasoning Enabled LLM Guardrail via Knowledge-Enhanced Logic
al Reasoning
As large language models (LLMs) become increasingly prevalent across various app
lications, it is critical to establish safety guardrails to moderate input/outpu
t content of LLMs and ensure compliance with safety policies. Existing guardrail
 models, such as OpenAI Mod and LlamaGuard, treat various safety categories (e.g
., self-harm, self-harm/instructions) independently and fail to explicitly captu
re the intercorrelations among them. This has led to limitations such as ineffec
tiveness due to inadequate training on long-tail data from correlated safety cat
egories, susceptibility to jailbreaking attacks, and inflexibility regarding new
 safety categories.
To address these limitations, we propose R^2-Guard, a robust reasoning enabled
 LLM guardrail via knowledge-enhanced logical reasoning. Specifically, R^2-Gua
rd comprises two parts: data-driven guardrail models and reasoning components. T
he data-driven guardrail models provide unsafety probabilities of moderated cont
ent on different safety categories.
We then encode safety knowledge among different categories as first-order logica
l rules and embed them into a probabilistic graphic model (PGM) based reasoning
component. The unsafety probabilities of different categories from data-driven g
uardrail models are sent to the reasoning component for final inference. We empl
oy two types of PGMs: Markov logic networks (MLNs) and probabilistic circuits (P
Cs), and optimize PCs to achieve precision-efficiency balance via improved graph
 structure. We also propose different methods to optimize the weights of knowled
ge. To further perform stress tests for guardrail models, we employ a pairwise c
onstruction method to construct a new safety benchmark TwinSafety, which feature
s principled categories and presents new challenges for moderation. We show that
 R^2-Guard is effective even given unrepresentative categories or challenging
jailbreaking prompts. We demonstrate the effectiveness of R^2-Guard by compari
sons with eight strong guardrail models on six standard moderation datasets, and
 demonstrate the robustness of R^2-Guard against four SOTA jailbreaking attack
s. R^2-Guard significantly surpasses SOTA method LlamaGuard by 12.6% on standa
rd moderation datasets and by 59.9% against jailbreaking attacks.
We further reveal that R^2-Guard can effectively adapt to safety category upda
tes by simply editing the PGM reasoning graph.
****
Zechen Zhang,Haim Sompolinsky
When narrower is better: the narrow width limit of Bayesian parallel branching n
eural networks
The infinite width limit of random neural networks is known to result in Neural
Networks as Gaussian Process (NNGP) (Lee et al. (2018)), characterized by task-i
ndependent kernels. It is widely accepted that larger network widths contribute
to improved generalization (Park et al. (2019)). However, this work challenges t
his notion by investigating the narrow width limit of the Bayesian Parallel Bran
ching
Neural Network (BPB-NN), an architecture that resembles neural networks with res
idual blocks. We demonstrate that when the width of a BPB-NN is significantly sm
aller compared to the number of training examples, each branch exhibits more rob
ust learning due to a symmetry breaking of branches in kernel renormalization. S
urprisingly, the performance of a BPB-NN in the narrow width limit is generally
superior to or comparable to that achieved in the wide width limit in bias-limit
ed scenarios. Furthermore, the readout norms of each branch in the narrow width
limit are mostly independent of the architectural hyperparameters but generally
reflective of the nature of the data. We demonstrate such phenomenon primarily i
n the branching graph neural networks, where each branch represents a different
order of convolutions of the graph; we also extend the results to other more gen
eral architectures such as the residual-MLP and demonstrate that the narrow widt
h effect is a general feature of the branching networks. Our results characteriz
e a newly defined narrow-width regime for parallel branching networks in general

.
****
Xinnan Dai,Haohao Qu,Yifei Shen,Bohang Zhang,Qihao Wen,Wenqi Fan,Dongsheng Li,Jiliang Tang,Caihua Shan
How Do Large Language Models Understand Graph Patterns? A Benchmark for Graph Pattern Comprehension
Benchmarking the capabilities and limitations of large language models (LLMs) in graph-related tasks is becoming an increasingly popular and crucial area of research. Recent studies have shown that LLMs exhibit a preliminary ability to understand graph structures and node features. However, the potential of LLMs in graph pattern mining remains largely unexplored. This is a key component in fields such as computational chemistry, biology, and social network analysis. To bridge this gap, this work introduces a comprehensive benchmark to assess LLMs' capabilities in graph pattern tasks. We have developed a benchmark that evaluates whether LLMs can understand graph patterns based on either terminological or topological descriptions. Additionally, our benchmark tests the LLMs' capacity to autonomously discover graph patterns from data. The benchmark encompasses both synthetic and real datasets, and a variety of models, with a total of 11 tasks and 7 models. Our experimental framework is designed for easy expansion to accommodate new models and datasets. Our findings reveal that: (1) LLMs have preliminary abilities to understand graph patterns, with O1-mini outperforming in the majority of tasks; (2) Formatting input graph data to align with the knowledge acquired during pretraining can enhance performance; (3) LLMs employ diverse potential algorithms to solve one task, with performance varying based on their execution capabilities.
****
Ziye Wang,Yiran Qin,Lin Zeng,Ruimao Zhang
High-Dynamic Radar Sequence Prediction for Weather Nowcasting Using Spatiotemporal Coherent Gaussian Representation
Weather nowcasting is an essential task that involves predicting future radar echo sequences based on current observations, offering significant benefits for disaster management, transportation, and urban planning. Current prediction methods are limited by training and storage efficiency, mainly focusing on 2D spatial predictions at specific altitudes. Meanwhile, 3D volumetric predictions at each timestamp remain largely unexplored. To address such a challenge, we introduce a comprehensive framework for 3D radar sequence prediction in weather nowcasting, using the newly proposed SpatioTemporal Coherent Gaussian Splatting (STC-GS) for dynamic radar representation and GauMamba for efficient and accurate forecasting. Specifically, rather than relying on a 4D Gaussian for dynamic scene reconstruction, STC-GS optimizes 3D scenes at each frame by employing a group of Gaussians while effectively capturing their movements across consecutive frames. It ensures consistent tracking of each Gaussian over time, making it particularly effective for prediction tasks. With the temporally correlated Gaussian groups established, we utilize them to train GauMamba, which integrates a memory mechanism into the Mamba framework. This allows the model to learn the temporal evolution of Gaussian groups while efficiently handling a large volume of Gaussian tokens. As a result, it achieves both efficiency and accuracy in forecasting a wide range of dynamic meteorological radar signals. The experimental results demonstrate that our STC-GS can efficiently represent 3D radar sequences with over $16\times$ higher spatial resolution compared with the existing 3D representation methods, while GauMamba outperforms state-of-the-art methods in forecasting a broad spectrum of high-dynamic weather conditions.
****
Aviral Kumar,Vincent Zhuang,Rishabh Agarwal,Yi Su,John D Co-Reyes,Avi Singh,Kate Baumli,Shariq Iqbal,Colton Bishop,Rebecca Roelofs,Lei M Zhang,Kay McKinney,Disha Shrivastava,Cosmin Paduraru,George Tucker,Doina Precup,Feryal Behbahani,Aleksandra Faust
Training Language Models to Self-Correct via Reinforcement Learning
Self-correction is a highly desirable capability of large language models (LLMs), yet it has consistently been found to be largely ineffective in modern LLMs.

Current methods for training self-correction typically depend on either multiple models, a more advanced model, or additional forms of supervision. To address these shortcomings, we develop a multi-turn online reinforcement learning (RL) approach, SCoRe, that significantly improves an LLM's self-correction ability using entirely self-generated data. To build SCoRe, we first show that variants of supervised fine-tuning (SFT) on offline model-generated correction traces are insufficient for instilling self-correction behavior. In particular, we observe that training via SFT either suffers from a distribution mismatch between the training data and the model's own responses or implicitly prefers only a certain mode of correction behavior that is often not effective at test time. SCoRe addresses these challenges by training under the model's own distribution of self-generated correction traces and using appropriate regularization to steer the learning process into learning a self-correction strategy that is effective at test time as opposed to simply fitting high-reward responses for a given prompt. This regularization prescribes running a first phase of RL on a base model to generate a policy initialization that is less susceptible to collapse and then using a reward bonus to amplify self-correction during training. When applied to Gemini 1.0 Pro and 1.5 Flash models, we find that SCoRe achieves state-of-the-art self-correction performance, improving the base models' self-correction by 15.6% and 9.1% respectively on the MATH and HumanEval benchmarks.
****

Hao Bai,Yifei Zhou,Li Erran Li,Sergey Levine,Aviral Kumar
Digi-Q: Learning VLM Q-Value Functions for Training Device-Control Agents
While a number of existing approaches for building foundation model agents rely on prompting or fine-tuning with human demonstrations, it is not sufficient in dynamic environments (e.g., mobile device control). On-policy reinforcement learning (RL) should address these limitations, but collecting actual rollouts in an environment is often undesirable in truly open-ended agentic problems such as mobile device control or interacting with humans, where each unit of interaction is associated with a cost. In such scenarios, a method for policy learning that can utilize off-policy experience by learning a trained action-value function is much more effective. In this paper, we develop an approach, called Digi-Q, to train VLM-based action-value Q-functions which are then used to extract the agent policy. We study our approach in the mobile device control setting. Digi-Q trains the Q-function using offline temporal-difference (TD) learning, on top of frozen, intermediate-layer features of a VLM. Compared to fine-tuning the whole VLM, this approach saves us compute and enhances scalability. To make the VLM features amenable for representing the Q-function, we need to employ an initial phase of fine-tuning to amplify coverage over actionable information needed for value function. Once trained, we use this Q-function via a Best-of-N policy extraction operator that imitates the best action out of multiple candidate actions from the current policy as ranked by the value function, enabling policy improvement without environment interaction. Digi-Q outperforms several prior methods on user-scale device control tasks in Android-in-the-Wild, attaining 21.2% improvement over prior best-performing method. In some cases, our Digi-Q approach already matches state-of-the-art RL methods that require interaction. The project is open-sourced at https://github.com/DigiRL-agent/digiq
****

Junyeong Park,Junmo Cho,Sungjin Ahn
MrSteve: Instruction-Following Agents in Minecraft with What-Where-When Memory
Significant advances have been made in developing general-purpose embodied AI in environments like Minecraft through the adoption of LLM-augmented hierarchical approaches. While these approaches, which combine high-level planners with low-level controllers, show promise, low-level controllers frequently become performance bottlenecks due to repeated failures. In this paper, we argue that the primary cause of failure in many low-level controllers is the absence of an episodic memory system. To address this, we introduce MrSteve (Memory Recall Steve), a novel low-level controller equipped with Place Event Memory (PEM), a form of episodic memory that captures what, where, and when information from episodes. This directly addresses the main limitation of the popular low-level controller, Steve

-1. Unlike previous models that rely on short-term memory, PEM organizes spatial and event-based data, enabling efficient recall and navigation in long-horizon tasks. Additionally, we propose an Exploration Strategy and a Memory-Augmented Task Solving Framework, allowing agents to alternate between exploration and task-solving based on recalled events. Our approach significantly improves task-solving and exploration efficiency compared to existing methods. We will release our code and demos on the project page: https://sites.google.com/view/mr-steve.
****
Jaemoo Choi,Yongxin Chen,Jaewoong Choi
Improving Neural Optimal Transport via Displacement Interpolation
Optimal Transport (OT) theory investigates the cost-minimizing transport map that moves a source distribution to a target distribution. Recently, several approaches have emerged for learning the optimal transport map for a given cost function using neural networks. We refer to these approaches as the OT Map. OT Map provides a powerful tool for diverse machine learning tasks, such as generative modeling and unpaired image-to-image translation. However, existing methods that utilize max-min optimization often experience training instability and sensitivity to hyperparameters. In this paper, we propose a novel method to improve stability and achieve a better approximation of the OT Map by exploiting displacement interpolation, dubbed Displacement Interpolation Optimal Transport Model (DIOTM). We derive the dual formulation of displacement interpolation at specific time t and prove how these dual problems are related across time. This result allows us to utilize the entire trajectory of displacement interpolation in learning the OT Map. Our method improves the training stability and achieves superior results in estimating optimal transport maps. We demonstrate that DIOTM outperforms existing OT-based models on image-to-image translation tasks.
****
Jayneel Parekh,Quentin Bouniot,Pavlo Mozharovskyi,Alasdair Newson,Florence d'Alché-Buc
Restyling Unsupervised Concept Based Interpretable Networks with Generative Models
Developing inherently interpretable models for prediction has gained prominence in recent years. A subclass of these models, wherein the interpretable network relies on learning high-level concepts, are valued because of closeness of concept representations to human communication. However, the visualization and understanding of the learnt unsupervised dictionary of concepts encounters major limitations, especially for large-scale images. We propose here a novel method that relies on mapping the concept features to the latent space of a pretrained generative model. The use of a generative model enables high quality visualization, and lays out an intuitive and interactive procedure for better interpretation of the learnt concepts by imputing concept activations and visualizing generated modifications. Furthermore, leveraging pretrained generative models has the additional advantage of making the training of the system more efficient. We quantitatively ascertain the efficacy of our method in terms of accuracy of the interpretable prediction network, fidelity of reconstruction, as well as faithfulness and consistency of learnt concepts. The experiments are conducted on multiple image recognition benchmarks for large-scale images. Project page available at https://jayneelparekh.github.io/VisCoIN_project_page/
****
Lunjun Zhang,Arian Hosseini,Hritik Bansal,Mehran Kazemi,Aviral Kumar,Rishabh Agarwal
Generative Verifiers: Reward Modeling as Next-Token Prediction
Verifiers or reward models are often used to enhance the reasoning performance of large language models (LLMs). A common approach is the Best-of-N method, where N candidate solutions generated by the LLM are ranked by a verifier, and the best one is selected. While LLM-based verifiers are typically trained as discriminative classifiers to score solutions, they do not utilize the text generation capabilities of pretrained LLMs. To overcome this limitation, we instead propose training verifiers using the ubiquitous next-token prediction objective, jointly on verification and solution generation. Compared to standard verifiers, such ge

nerative verifiers (GenRM) can benefit from several advantages of LLMs: they int
egrate seamlessly with instruction tuning, enable chain-of-thought reasoning, an
d can utilize additional test-time compute via majority voting for better verifi
cation. We demonstrate that GenRM outperforms discriminative, DPO verifiers, and
 LLM-as-a-Judge, resulting in large performance gains with Best-of-N, namely 5%
→ 45.3% on algorithmic tasks, 73% → 93.4% on GSM8K, and 28% →44.6% on easy-to-ha
rd generalization on MATH. Furthermore, we find that training GenRM with synthet
ic verification rationales is sufficient to pick out subtle errors on math probl
ems. Finally, we demonstrate that generative verifiers scale favorably with mode
l size and inference-time compute.
****

Zhijian Zhuo,Ya Wang,Yutao Zeng,Xiaoqing Li,Xun Zhou,Jinwen Ma
Polynomial Composition Activations: Unleashing the Dynamics of Large Language Mo
dels
Transformers have found extensive applications across various domains due to the
ir powerful fitting capabilities. This success can be partially attributed to th
eir inherent nonlinearity. Thus, in addition to the ReLU function employed in th
e original transformer architecture, researchers have explored alternative modul
es such as GeLU and SwishGLU to enhance nonlinearity and thereby augment represe
ntational capacity. In this paper, we propose a novel category of polynomial com
position activations (PolyCom), designed to optimize the dynamics of transformer
s. Theoretically, we provide a comprehensive mathematical analysis of PolyCom, h
ighlighting its enhanced expressivity and efficacy relative to other activation
functions. Notably, we demonstrate that networks incorporating PolyCom achieve t
he **optimal approximation rate**, indicating that PolyCom networks require mini
mal parameters to approximate general smooth functions in Sobolev spaces. We con
duct empirical experiments on the pre-training configurations of large language
models (LLMs), including both dense and sparse architectures. By substituting co
nventional activation functions with PolyCom, we enable LLMs to capture higher-o
rder interactions within the data, thus improving performance metrics in terms o
f accuracy and convergence rates. Extensive experimental results demonstrate th
e effectiveness of our method, showing substantial improvements over other activ
ation functions. Code is available at https://github.com/BryceZhuo/PolyCom.
****

Junkang Wu,Yuexiang Xie,Zhengyi Yang,Jiancan Wu,Jiawei Chen,Jinyang Gao,Bolin Di
ng,Xiang Wang,Xiangnan He
Towards Robust Alignment of Language Models: Distributionally Robustifying Direc
t Preference Optimization
This study addresses the challenge of noise in training datasets for Direct Pref
erence Optimization (DPO), a method for aligning Large Language Models (LLMs) wi
th human preferences. We categorize noise into pointwise noise, which includes l
ow-quality data points, and pairwise noise, which encompasses erroneous data pai
r associations that affect preference rankings. Utilizing Distributionally Robus
t Optimization (DRO), we enhance DPO's resilience to these types of noise. Our t
heoretical insights reveal that DPO inherently embeds DRO principles, conferring
 robustness to pointwise noise, with the regularization coefficient β play
ing a critical role in its noise resistance. Extending this framework, we introd
uce Distributionally Robustifying DPO (Dr. DPO), which integrates pairwise robus
tness by optimizing against worst-case pairwise scenarios. The novel hyperparame
ter β' in Dr. DPO allows for fine-tuned control over data pair reliability
, providing a strategic balance between exploration and exploitation in noisy tr
aining environments. Empirical evaluations demonstrate that Dr. DPO substantiall
y improves the quality of generated text and response accuracy in preference dat
asets, showcasing enhanced performance in both noisy and noise-free settings.
****

Dung Manh Nguyen,Thang Chau Phan,Nam Le Hai,Tien-Thong Doan,Nam V. Nguyen,Quang
Pham,Nghi D. Q. Bui
CodeMMLU: A Multi-Task Benchmark for Assessing Code Understanding & Reasoning Ca
pabilities of CodeLLMs
Recent advances in Code Large Language Models (CodeLLMs) have primarily focused

on open-ended code generation, often overlooking the crucial aspect of code unde
rstanding & reasoning. To bridge this gap, we introduce CodeMMLU, a comprehensiv
e multiple-choice benchmark designed to evaluate the depth of software and code
comprehension in LLMs. CodeMMLU includes nearly 20,000 questions spanning divers
e domains, including code analysis, defect detection, and software engineering p
rinciples across multiple programming languages. Unlike traditional benchmarks t
hat emphasize code generation, CodeMMLU assesses a model’s ability to reason abo
ut programs across a wide-range of tasks such as code repair, execution reasonin
g, and fill-in-the-blank challenges. Our extensive evaluation reveals that even
state-of-the-art models struggle with CodeMMLU, highlighting significant gaps in
 comprehension beyond generation. By emphasizing the essential connection betwee
n code understanding and effective AI-assisted development, CodeMMLU provides a
critical resource for advancing more reliable and capable coding assistants.
****

Dujian Ding,Bicheng Xu,Laks V. S. Lakshmanan
OCCAM: Towards Cost-Efficient and Accuracy-Aware Classification Inference
Classification tasks play a fundamental role in various applications, spanning d
omains such as healthcare, natural language processing and computer vision. With
 the growing popularity and capacity of machine learning models, people can easi
ly access trained classifiers as a service online or offline. However, model use
 comes with a cost and classifiers of higher capacity (such as large foundation
models) usually incur higher inference costs. To harness the respective strength
s of different classifiers, we propose a principled approach, OCCAM, to compute
the best classifier assignment strategy over classification queries (termed as t
he optimal model portfolio) so that the aggregated accuracy is maximized, under
user-specified cost budgets. Our approach uses an unbiased and low-variance accu
racy estimator and effectively computes the optimal solution by solving an integ
er linear programming problem. On a variety of real-world datasets, OCCAM achiev
es 40% cost reduction with little to no accuracy drop.
****

Kaiwen Zheng,Yongxin Chen,Hanzi Mao,Ming-Yu Liu,Jun Zhu,Qinsheng Zhang
Masked Diffusion Models are Secretly Time-Agnostic Masked Models and Exploit Ina
ccurate Categorical Sampling
Masked diffusion models (MDMs) have emerged as a popular research topic for gene
rative modeling of discrete data, thanks to their superior performance over othe
r discrete diffusion models, and are rivaling the auto-regressive models (ARMs)
for language modeling tasks. The recent effort in simplifying the masked diffusi
on framework further leads to alignment with continuous-space diffusion models a
nd more principled training and sampling recipes.
In this paper, however, we reveal that both training and sampling of MDMs are th
eoretically free from the time variable, arguably the key signature of diffusion
 models, and are instead equivalent to masked models. The connection on the samp
ling aspect is drawn by our proposed first-hitting sampler (FHS). Specifically,
we show that the FHS is theoretically equivalent to MDMs' original generation pr
ocess while significantly alleviating the time-consuming categorical sampling an
d achieving a 20$\times$ speedup. In addition, our investigation raises doubts a
bout whether MDMs can truly beat ARMs in text generation. We identify, for the f
irst time, an underlying numerical issue, even with the commonly used 32-bit flo
ating-point precision, which results in inaccurate categorical sampling.
We show that it lowers the effective temperature both theoretically and empirica
lly, and the resulting decrease in token diversity makes previous evaluations, w
hich assess the generation quality solely through the incomplete generative perp
lexity metric, somewhat unfair.
****

Meng Yu,Kun Zhan
Bias Mitigation in Graph Diffusion Models
Most existing graph diffusion models have significant bias problems. We observe
that the forward diffusion’s maximum perturbation distribution in most models de
viates from the standard Gaussian distribution, while reverse sampling consisten
tly starts from a standard Gaussian distribution, which results in a reverse-sta

rting bias. Together with the inherent exposure bias of diffusion models, this r
esults in degraded generation quality. This paper proposes a comprehensive appro
ach to mitigate both biases. To mitigate reverse-starting bias, we employ a newl
y designed Langevin sampling algorithm to align with the forward maximum perturb
ation distribution, establishing a new reverse-starting point. To address the ex
posure bias, we introduce a score correction mechanism based on a newly defined
score difference. Our approach, which requires no network modifications, is vali
dated across multiple models, datasets, and tasks, achieving state-of-the-art re
sults.
****
Ranajoy Sadhukhan,Jian Chen,Zhuoming Chen,Vashisth Tiwari,Ruihang Lai,Jinyuan Sh
i,Ian En-Hsu Yen,Avner May,Tianqi Chen,Beidi Chen
MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation w
ith Speculative Decoding
Large Language Models (LLMs) have become more prevalent in long-context applicat
ions such as interactive chatbots, document analysis, and agent workflows, but i
t is challenging to serve long-context requests with low latency and high throug
hput. Speculative decoding (SD) is a widely used technique to reduce latency los
slessly, but the conventional wisdom suggests that its efficacy is limited to sm
all batch sizes. In MagicDec, we show that surprisingly SD can achieve speedup e
ven for a high throughput inference regime for moderate to long sequences. More
interestingly, an intelligent drafting strategy can achieve better speedup with
increasing batch size based on our rigorous analysis. MagicDec first identifies
the bottleneck shifts with increasing batch size and sequence length, and uses t
hese insights to deploy SD more effectively for high throughput inference. We le
verage draft model with sparse KV cache to address the KV bottleneck, which scal
es with both sequence length and batch size. Additionally, we propose a theoreti
cal model to select the optimal drafting strategy for maximum speedup. Our work
highlights the broad applicability of speculative decoding in long-context servi
ng, as it can enhance throughput and reduce latency without compromising accurac
y. For moderate to long sequences, we demonstrate up to 2.51x speedup for LLaMA-
3.1-8B when serving batch sizes ranging from 32 to 256 on various types of hardw
are and tasks.
****
Mohamed El Amine Boudjoghra,Angela Dai,Jean Lahoud,Hisham Cholakkal,Rao Muhammad
 Anwer,Salman Khan,Fahad Shahbaz Khan
Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation
Recent works on open-vocabulary 3D instance segmentation show strong promise but
 at the cost of slow inference speed and high computation requirements. This hig
h computation cost is typically due to their heavy reliance on aggregated clip f
eatures from multi-view, which require computationally expensive 2D foundation m
odels like Segment Anything (SAM) and CLIP. Consequently, this hampers their app
licability in many real-world applications that require both fast and accurate p
redictions. To this end, we propose a novel open-vocabulary 3D instance segmenta
tion approach, named Open-YOLO 3D, that efficiently leverages only 2D object det
ection from multi-view RGB images for open-vocabulary 3D instance segmentation.
 We demonstrate that our proposed Multi-View Prompt Distribution (MVPDist) metho
d makes use of multi-view information to account for misclassification from the
object detector to predict a reliable label for 3D instance masks. Furthermore,
since projections of 3D object instances are already contained within the 2D bou
nding boxes, we show that our proposed low granularity label maps, which require
 only a 2D object detector to construct, are sufficient and very fast to predict
 prompt IDs for 3D instance masks when used with our proposed MVPDist.
 We validate our Open-YOLO 3D on two benchmarks, ScanNet200 and Replica,
 under two scenarios: (i) with ground truth masks, where labels are required for
 given object proposals, and (ii) with class-agnostic 3D proposals generated fro
m a 3D proposal network.
 Our Open-YOLO 3D achieves state-of-the-art performance on both datasets while o
btaining up to $\sim$16$\times$ speedup compared to the best existing method in
literature. On ScanNet200 val. set, our Open-YOLO 3D achieves mean average preci

sion (mAP) of 24.7% while operating at 22 seconds per scene. github.com/aminebdj/OpenYOLO3D
****

Yang You,Yixin Li,Congyue Deng,Yue Wang,Leonidas Guibas
Multiview Equivariance Improves 3D Correspondence Understanding with Minimal Feature Finetuning
Vision foundation models, particularly the ViT family, have revolutionized image understanding by providing rich semantic features. However, despite their success in 2D comprehension, their abilities on grasping 3D spatial relationships are still unclear.
In this work, we evaluate and enhance the 3D awareness of ViT-based models. We begin by systematically assessing their ability to learn 3D equivariant features, specifically examining the consistency of semantic embeddings across different viewpoints. Our findings indicate that improved 3D equivariance leads to better performance on various downstream tasks, including pose estimation, tracking, and semantic transfer. Building on this insight, we propose a simple yet effective finetuning strategy based on 3D correspondences, which significantly enhances the 3D understanding of existing vision models. Remarkably, even finetuning on a single object for just one iteration results in substantial performance gains. Code is available on https://github.com/qq456cvb/3DCorrEnhance.
****

Dong Un Kang,Hayeon Kim,Se Young Chun
Class Distribution-induced Attention Map for Open-vocabulary Semantic Segmentations
Open-vocabulary semantic segmentation is a challenging task that assigns seen or unseen class labels to individual pixels. While recent works with vision-language models (VLMs) have shown promising results in zero-shot semantic segmentation, they still struggle to accurately localize class-related objects. In this work, we argue that CLIP-based prior works yield patch-wise noisy class predictions while having highly correlated class distributions for each object. Then, we propose Class Distribution-induced Attention Map, dubbed CDAM, that is generated by the Jensen-Shannon divergence between class distributions of two patches that belong to the same (class) object. This CDAM can be used for open-vocabulary semantic segmentation by integrating it into the final layer of CLIP to enhance the capability to accurately localize desired classes. Our class distribution-induced attention scheme can easily work with multi-scale image patches as well as augmented text prompts for further enhancing attention maps. By exploiting class distribution, we also propose robust entropy-based background thresholding for the inference of semantic segmentation. Interestingly, the core idea of our proposed method does not conflict with other prior arts in zero-shot semantic segmentation, thus can be synergetically used together, yielding substantial improvements in performance across popular semantic segmentation benchmarks.
****

Rustem Islamov,Yuan Gao,Sebastian U Stich
Towards Faster Decentralized Stochastic Optimization with Communication Compression
Communication efficiency has garnered significant attention as it is considered the main bottleneck for large-scale decentralized Machine Learning applications in distributed and federated settings. In this regime, clients are restricted to transmitting small amounts of compressed information to their neighbors over a communication graph. Numerous endeavors have been made to address this challenging problem by developing algorithms with compressed communication for decentralized non-convex optimization problems. Despite considerable efforts, current theoretical understandings of the problem are still very limited, and existing algorithms all suffer from various limitations. In particular, these algorithms typically rely on strong, and often infeasible assumptions such as bounded data heterogeneity or require large batch access while failing to achieve linear speedup with the number of clients. In this paper, we introduce MoTEF, a novel approach that integrates communication compression with $\textbf{Mo}$mentum $\textbf{T}$racking and $\textbf{E}$rror $\textbf{F}$eedback. MoTEF is the first algorithm to

achieve an asymptotic rate matching that of distributed SGD under arbitrary data heterogeneity, hence resolving a long-standing theoretical obstacle in decentralized optimization with compressed communication. We provide numerical experiments to validate our theoretical findings and confirm the practical superiority of MoTEF.
****
Andreas Christian Schneider,Valentin Neuhaus,David Alexander Ehrlich,Abdullah Makkeh,Alexander S Ecker,Viola Priesemann,Michael Wibral
What should a neuron aim for? Designing local objective functions based on information theory
In modern deep neural networks, the learning dynamics of individual neurons are often obscure, as the networks are trained via global optimization. Conversely, biological systems build on self-organized, local learning, achieving robustness and efficiency with limited global information. Here, we show how self-organization between individual artificial neurons can be achieved by designing abstract bio-inspired local learning goals. These goals are parameterized using a recent extension of information theory, Partial Information Decomposition (PID), which decomposes the information that a set of information sources holds about an outcome into unique, redundant and synergistic contributions. Our framework enables neurons to locally shape the integration of information from various input classes, i.e., feedforward, feedback, and lateral, by selecting which of the three inputs should contribute uniquely, redundantly or synergistically to the output. This selection is expressed as a weighted sum of PID terms, which, for a given problem, can be directly derived from intuitive reasoning or via numerical optimization, offering a window into understanding task-relevant local information processing. Achieving neuron-level interpretability while enabling strong performance using local learning, our work advances a principled information-theoretic foundation for local learning strategies.
****
Xinran Li,Xiaolu Wang,Chenjia Bai,Jun Zhang
Exponential Topology-enabled Scalable Communication in Multi-agent Reinforcement Learning
In cooperative multi-agent reinforcement learning (MARL), well-designed communication protocols can effectively facilitate consensus among agents, thereby enhancing task performance. Moreover, in large-scale multi-agent systems commonly found in real-world applications, effective communication plays an even more critical role due to the escalated challenge of partial observability compared to smaller-scale setups. In this work, we endeavor to develop a scalable communication protocol for MARL. Unlike previous methods that focus on selecting optimal pairwise communication links—a task that becomes increasingly complex as the number of agents grows—we adopt a global perspective on communication topology design. Specifically, we propose utilizing the exponential topology to enable rapid information dissemination among agents by leveraging its small-diameter and small-size properties. This approach leads to a scalable communication protocol, named ExpoComm. To fully unlock the potential of exponential graphs as communication topologies, we employ memory-based message processors and auxiliary tasks to ground messages, ensuring that they reflect global information and benefit decision-making. Extensive experiments on large-scale cooperative benchmarks, including MAgent and Infrastructure Management Planning, demonstrate the superior performance and robust zero-shot transferability of ExpoComm compared to existing communication strategies. The
code is publicly available at https://github.com/LXXXXR/ExpoComm.
****
Aron Brenner,Rahman Khorramfar,Jennifer Z Sun,Saurabh Amin
A Deep Generative Learning Approach for Two-stage Adaptive Robust Optimization
Two-stage adaptive robust optimization (ARO) is a powerful approach for planning under uncertainty, balancing first-stage decisions with recourse decisions made after uncertainty is realized. To account for uncertainty, modelers typically define a simple uncertainty set over which potential outcomes are considered. How

ever, classical methods for defining these sets unintentionally capture a wide r
ange of unrealistic outcomes, resulting in overly-conservative and costly planni
ng in anticipation of unlikely contingencies. In this work, we introduce AGRO, a
 solution algorithm that performs adversarial generation for two-stage adaptive
robust optimization using a variational autoencoder. AGRO generates high-dimensi
onal contingencies that are simultaneously adversarial and realistic, improving
the robustness of first-stage decisions at a lower planning cost than standard m
ethods. To ensure generated contingencies lie in high-density regions of the unc
ertainty distribution, AGRO defines a tight uncertainty set as the image of "lat
ent" uncertainty sets under the VAE decoding transformation. Projected gradient
ascent is then used to maximize recourse costs over the latent uncertainty sets
by leveraging differentiable optimization methods. We demonstrate the cost-effic
iency of AGRO by applying it to both a synthetic production-distribution problem
 and a real-world power system expansion setting. We show that AGRO outperforms
the standard column-and-constraint algorithm by up to 1.8% in production-distrib
ution planning and up to 8% in power system expansion.
****

Hangting Ye,He Zhao,Wei Fan,Mingyuan Zhou,Dan dan Guo,Yi Chang
DRL: Decomposed Representation Learning for Tabular Anomaly Detection
Anomaly detection, indicating to identify the anomalies that significantly devia
te from the majority normal instances of data, has been an important role in mac
hine learning and related applications. Despite the significant success achieved
 in anomaly detection on image and text data, the accurate Tabular Anomaly Detec
tion (TAD) has still been hindered due to the lack of clear prior semantic infor
mation in the tabular data. Most state-of-the-art TAD studies are along the line
 of reconstruction, which first reconstruct training data and then use reconstru
ction errors to decide anomalies; however, reconstruction on training data can s
till hardly distinguish anomalies due to the data entanglement in their represen
tations. To address this problem, in this paper, we propose a novel approach Dec
omposed Representation Learning (DRL), to re-map data into a tailor-designed con
strained space, in order to capture the underlying shared patterns of normal sam
ples and differ anomalous patterns for TAD.
Specifically, we enforce the representation of each normal sample in the latent
space to be decomposed into a weighted linear combination of randomly generated
orthogonal basis vectors, where these basis vectors are both data-free and train
ing-free.
Furthermore, we enhance the discriminative capability between normal and anomalo
us patterns in the latent space by introducing a novel constraint that amplifies
 the discrepancy between these two categories, supported by theoretical analysis
.
Finally, extensive experiments on 40 tabular datasets and 16 competing tabular a
nomaly detection algorithms show that our method achieves state-of-the-art perfo
rmance.
****

Hanxiang Ren,Li Sun,Xulong Wang,Pei Zhou,Zewen Wu,Siyan Dong,Difan Zou,Youyi Zhe
ng,Yanchao Yang
HyPoGen: Optimization-Biased Hypernetworks for Generalizable Policy Generation
Policy learning through behavior cloning poses significant challenges, particula
rly when demonstration data is limited. In this work, we present HyPoGen, a nove
l optimization-biased hypernetwork for policy generation. The proposed hypernetw
ork learns to synthesize optimal policy parameters solely from task specificatio
ns -- without accessing training data -- by modeling policy generation as an ap
proximation of the optimization process executed over a finite number of steps a
nd assuming these specifications serve as a sufficient representation of the dem
onstration data. By incorporating structural designs that bias the hypernetwork
towards optimization, we can improve its generalization capability while only tr
aining on source task demonstrations. During the feed-forward prediction pass, t
he hypernetwork effectively performs an optimization in the latent (compressed)
policy space, which is then decoded into policy parameters for action prediction
. Experimental results on locomotion and manipulation benchmarks show that HyPoG

en significantly outperforms state-of-the-art methods in generating policies for unseen target tasks without any demonstrations, achieving higher success rates and underscoring the potential of optimization-biased hypernetworks in advancing generalizable policy generation. Our code and data are available at: https://github.com/ReNginx/HyPoGen.
****
Zengwei Yao,Wei Kang,Xiaoyu Yang,Fangjun Kuang,Liyong Guo,Han Zhu,Zengrui Jin,Zhaoqing Li,Long Lin,Daniel Povey
CR-CTC: Consistency regularization on CTC for improved speech recognition
Connectionist Temporal Classification (CTC) is a widely used method for automatic speech recognition (ASR), renowned for its simplicity and computational efficiency. However, it often falls short in recognition performance. In this work, we propose the Consistency-Regularized CTC (CR-CTC), which enforces consistency between two CTC distributions obtained from different augmented views of the input speech mel-spectrogram. We provide in-depth insights into its essential behaviors from three perspectives: 1) it conducts self-distillation between random pairs of sub-models that process different augmented views; 2) it learns contextual representation through masked prediction for positions within time-masked regions, especially when we increase the amount of time masking; 3) it suppresses the extremely peaky CTC distributions, thereby reducing overfitting and improving the generalization ability. Extensive experiments on LibriSpeech, Aishell-1, and GigaSpeech datasets demonstrate the effectiveness of our CR-CTC. It significantly improves the CTC performance, achieving state-of-the-art results comparable to those attained by transducer or systems combining CTC and attention-based encoder-decoder (CTC/AED). We release our code at \url{https://github.com/k2-fsa/icefall}.
****
Milena Gazdieva,Jaemoo Choi,Alexander Kolesov,Jaewoong Choi,Petr Mokrov,Alexander Korotin
Robust Barycenter Estimation using Semi-Unbalanced Neural Optimal Transport
Aggregating data from multiple sources can be formalized as an *Optimal Transport* (OT) barycenter problem, which seeks to compute the average of probability distributions with respect to OT discrepancies. However, in real-world scenarios, the presence of outliers and noise in the data measures can significantly hinder the performance of traditional statistical methods for estimating OT barycenters. To address this issue, we propose a novel scalable approach for estimating the *robust* continuous barycenter, leveraging the dual formulation of the *(semi-)unbalanced* OT problem. To the best of our knowledge, this paper is the first attempt to develop an algorithm for robust barycenters under the continuous distribution setup. Our method is framed as a $\min$-$\max$ optimization problem and is adaptable to *general* cost functions. We rigorously establish the theoretical underpinnings of the proposed method and demonstrate its robustness to outliers and class imbalance through a number of illustrative experiments. Our source code is publicly available at https://github.com/milenagazdieva/U-NOTBarycenters.
****
Futing Wang,Jianhao Yan,Yue Zhang,Tao Lin
ELICIT: LLM Augmentation Via External In-context Capability
Enhancing the adaptive capabilities of large language models is a critical pursuit in both research and application.
Traditional fine-tuning methods require substantial data, computational resources, and specific capabilities, while in-context learning is limited by the need for appropriate demonstrations and efficient token usage.

Inspired by the expression of in-context learned capabilities through task vectors and the concept of modular capability or knowledge, we propose ELICIT, a framework consisting of two modules designed to effectively store and reuse task vectors to enhance the diverse adaptive capabilities of models without additional training or inference tokens.

Our comprehensive experiments and analysis demonstrate that our pipeline is highly transferable across different input formats, tasks, and model architectures.

Externally storing and reusing vectors that represent in-context learned capabilities not only shows the potential to extract modular capabilities but also significantly enhances the performance, versatility, adaptability, and scalability of large language models, paving the way for more efficient and effective use of these models in a wide range of applications.
****
Han Wang,Yilin Zhao,Dian Li,Xiaohan Wang,sinbadliu,Xuguang Lan,Hui Wang
Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps
Humor is previously regarded as a gift exclusive to humans for the following reasons. Humor is a culturally nuanced aspect of human language, presenting challenges for its understanding and generation.
Humor generation necessitates a multi-hop reasoning process, with each hop founded on proper rationales. Although many studies, such as those related to GPT-o1, focus on logical reasoning with reflection and correction, they still fall short in humor generation. Due to the sparsity of the knowledge graph in creative thinking, it is arduous to achieve multi-hop reasoning.
Consequently, in this paper, we propose a more robust framework for addressing the humor reasoning task, named LoL. LoL aims to inject external information to mitigate the sparsity of the knowledge graph, thereby enabling multi-hop reasoning. In the first stage of LoL, we put forward an automatic instruction-evolution method to incorporate the deeper and broader thinking processes underlying humor.
Judgment-oriented instructions are devised to enhance the model's judgment capability, dynamically supplementing and updating the sparse knowledge graph. Subsequently, through reinforcement learning, the reasoning logic for each online-generated response is extracted using GPT-4o. In this process, external knowledge is re-introduced to aid the model in logical reasoning and the learning of human preferences.
Finally, experimental results indicate that the combination of these two processes can enhance both the model's judgment ability and its generative capacity.
These findings deepen our comprehension of the creative capabilities of large language models (LLMs) and offer approaches to boost LLMs' creative abilities for cross-domain innovative applications.
****
Diego Garcia Cerdas,Christina Sartzetaki,Magnus Petersen,Gemma Roig,Pascal Mettes,Iris Groen
BrainACTIV: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation
The human brain efficiently represents visual inputs through specialized neural populations that selectively respond to specific categories. Advancements in generative modeling have enabled data-driven discovery of neural selectivity using brain-optimized image synthesis. However, current methods independently generate one sample at a time, without enforcing structural constraints on the generations; thus, these individual images have no explicit point of comparison, making it hard to discern which image features drive neural response selectivity. To address this issue, we introduce Brain Activation Control Through Image Variation (BrainACTIV), a method for manipulating a reference image to enhance or decrease activity in a target cortical region using pretrained diffusion models. Starting from a reference image allows for fine-grained and reliable offline identification of optimal visuo-semantic properties, as well as producing controlled stimuli for novel neuroimaging studies. We show that our manipulations effectively modulate predicted fMRI responses and agree with hypothesized preferred categories in established regions of interest, while remaining structurally close to the reference image. Moreover, we demonstrate how our method accentuates differences between brain regions that are selective to the same category, and how it could be used to explore neural representation of brain regions with unknown selectivities. Hence, BrainACTIV holds the potential to formulate robust hypotheses about brain representation and to facilitate the production of naturalistic stimuli for neuroscientific experiments.

****
Zhao Mandi,Yijia Weng,Dominik Bauer,Shuran Song
Real2Code: Reconstruct Articulated Objects via Code Generation
We present Real2Code, a novel approach to reconstructing articulated objects via code generation. Given visual observations of an object, we first reconstruct its part geometry using image segmentation and shape completion. We represent these object parts with oriented bounding boxes, from which a fine-tuned large language model (LLM) predicts joint articulation as code. By leveraging pre-trained vision and language models, our approach scales elegantly with the number of articulated parts, and generalizes from synthetic training data to real world objects in unstructured environments. Experimental results demonstrate that Real2Code significantly outperforms the previous state-of-the-art in terms of reconstruction accuracy, and is the first approach to extrapolate beyond objects' structural complexity in the training set, as we show for objects with up to 10 articulated parts. When incorporated with a stereo reconstruction model, Real2Code moreover generalizes to real-world objects, given only a handful of multi-view RGB images and without the need for depth or camera information.
****
Devon R. Graham,Kevin Leyton-Brown
Utilitarian Algorithm Configuration for Infinite Parameter Spaces
Utilitarian algorithm configuration is a general-purpose technique for automatically searching the parameter space of a given algorithm to optimize its performance, as measured by a given utility function, on a given set of inputs. Recently introduced utilitarian configuration procedures offer optimality guarantees about the returned parameterization while provably adapting to the hardness of the underlying problem. However, the applicability of these approaches is severely limited by the fact that they only search a finite, relatively small set of parameters. They cannot effectively search the configuration space of algorithms with continuous or uncountable parameters. In this paper we introduce a new procedure, which we dub COUP (Continuous, Optimistic Utilitarian Procrastination). COUP is designed to search infinite parameter spaces efficiently to find good configurations quickly. Furthermore, COUP maintains the theoretical benefits of previous utilitarian configuration procedures when applied to finite parameter spaces but is significantly faster, both provably and experimentally.
****
Xin Xu,Tong Xiao,Zitong Chao,Zhenya Huang,Can Yang,Yang Wang
Can LLMs Solve Longer Math Word Problems Better?
Math Word Problems (MWPs) play a vital role in assessing the capabilities of Large Language Models (LLMs), yet current research primarily focuses on questions with concise contexts. The impact of longer contexts on mathematical reasoning remains under-explored. This study pioneers the investigation of Context Length Generalizability (CoLeG), which refers to the ability of LLMs to solve MWPs with extended narratives. We introduce Extended Grade-School Math (E-GSM), a collection of MWPs featuring lengthy narratives, and propose two novel metrics to evaluate the efficacy and resilience of LLMs in tackling these problems. Our analysis of existing zero-shot prompting techniques with proprietary LLMs along with open-source LLMs reveals a general deficiency in CoLeG. To alleviate these issues, we propose tailored approaches for different categories of LLMs. For proprietary LLMs, we introduce a new instructional prompt designed to mitigate the impact of long contexts. For open-source LLMs, we develop a novel auxiliary task for fine-tuning to enhance CoLeG. Our comprehensive results demonstrate the effectiveness of our proposed methods, showing improved performance on E-GSM. Additionally, we conduct an in-depth analysis to differentiate the effects of semantic understanding and reasoning efficacy, showing that our methods improves the latter. We also establish the generalizability of our methods across several other MWP benchmarks. Our findings highlight the limitations of current LLMs and offer practical solutions correspondingly, paving the way for further exploration of model generalizability and training methodologies.
****
Yunpeng Zhao,Jie Zhang

Does Training with Synthetic Data Truly Protect Privacy?
As synthetic data becomes increasingly popular in machine learning tasks, numerous methods---without formal differential privacy guarantees---use synthetic data for training. These methods often claim, either explicitly or implicitly, to protect the privacy of the original training data.
In this work, we explore four different training paradigms: coreset selection, dataset distillation, data-free knowledge distillation, and synthetic data generated from diffusion models. While all these methods utilize synthetic data for training, they lead to vastly different conclusions regarding privacy preservation. We caution that empirical approaches to preserving data privacy require careful and rigorous evaluation; otherwise, they risk providing a false sense of privacy.
****
Royina Karegoudra Jayanth,Yinshuang Xu,Ziyun Wang,Evangelos Chatzipantazis,Kostas Daniilidis,Daniel Gehrig
EqNIO: Subequivariant Neural Inertial Odometry
Neural network-based odometry using accelerometer and gyroscope readings from a single IMU can achieve robust, and low-drift localization capabilities, through the use of _neural displacement priors (NDPs)_. These priors learn to produce denoised displacement measurements but need to ignore data variations due to specific IMU mount orientation and motion directions, hindering generalization.
This work introduces EqNIO, which addresses this challenge with _canonical displacement priors_, i.e., priors that are invariant to the orientation of the gravity-aligned frame in which the IMU data is expressed. We train such priors on IMU measurements, that are mapped into a learnable canonical frame, which is uniquely defined via three axes: the first is gravity, making the frame gravity aligned, while the second and third are predicted from IMU data. The outputs (displacement and covariance) are mapped back to the original gravity-aligned frame. To maximize generalization, we find that these learnable frames must transform equivariantly with global gravity-preserving roto-reflections from the subgroup $O_g(3)\subset O(3)$, acting on the trajectory, rendering the NDP $O(3)$-_subequivariant_. We tailor specific linear, convolutional, and non-linear layers that commute with the actions of the group. Moreover, we introduce a bijective decomposition of angular rates into vectors that transform similarly to accelerations, allowing us to leverage both measurement types. Natively, angular rates would need to be inverted upon reflection, unlike acceleration, which hinders their joint processing. We highlight EqNIO's flexibility and generalization capabilities by applying it to both filter-based (TLIO), and end-to-end (RONIN) architectures, and outperforming existing methods that use _soft equivariance from auxiliary losses or data augmentation on various datasets. We believe this work paves the way for low-drift and generalizable neural inertial odometry on edge devices. The project details and code can be found at https://github.com/RoyinaJayanth/EqNIO.
****
Ori Yoran,Kunhao Zheng,Fabian Gloeckle,Jonas Gehring,Gabriel Synnaeve,Taco Cohen
The KoLMogorov Test: Compression by Code Generation
Compression is at the heart of intelligence. A theoretically optimal way to compress any sequence of data is to find the shortest program that outputs that sequence and then halts. However, such Kolmogorov compression is uncomputable, and code generating LLMs struggle to approximate this theoretical ideal, as it requires reasoning, planning and search capabilities beyond those of current models. In this work, we introduce the *KoLMogorov-Test* (KT), a compression-as-intelligence intelligence test for code generation LLMs. In KT a model is presented with a sequence of data at inference time, and asked to generate the shortest program that produces the sequence. We identify several benefits of KT for both evaluation and training: an essentially infinite number of problem instances of varying difficulty is readily available, strong baselines already exist, the evaluation metric (compression) cannot be gamed, and pretraining data contamination is highly unlikely. To evaluate current models, we use audio, text, and DNA data, as well as sequences produced by random synthetic programs. Current flagship models

perform poorly - both GPT4-o and Llama-3.1-405B struggle on our natural and syn
thetic sequences. On our synthetic distribution, we are able to train code gener
ation models with lower compression rates than previous approaches. Moreover, we
 show that gains on synthetic data generalize poorly to real data, suggesting th
at new innovations are necessary for additional gains on KT.
****
Xinbao Qiao,Meng Zhang,Ming Tang,Ermin Wei
Hessian-Free Online Certified Unlearning
Machine unlearning strives to uphold the data owners' right to be forgotten by e
nabling models to selectively forget specific data.
Recent advances suggest pre-computing and storing statistics extracted from seco
nd-order information and implementing unlearning through Newton-style updates.
However, the Hessian matrix operations are extremely costly and previous works c
onduct unlearning for empirical risk minimizer with the convexity assumption, pr
ecluding their applicability to high-dimensional over-parameterized models and t
he nonconvergence condition.
In this paper, we propose an efficient Hessian-free unlearning approach.
The key idea is to maintain a statistical vector for each training data, compute
d through affine stochastic recursion of the difference between the retrained an
d learned models.
We prove that our proposed method outperforms the state-of-the-art methods in te
rms of the unlearning and generalization guarantees, the deletion capacity, and
the time/storage complexity, under the same regularity conditions.
Through the strategy of recollecting statistics for removing data, we develop an
 online unlearning algorithm that achieves near-instantaneous data removal, as i
t requires only vector addition.
Experiments demonstrate that our proposed scheme surpasses existing results by o
rders of magnitude in terms of time/storage costs with millisecond-level unlearn
ing execution, while also enhancing test accuracy.
****
Chengwen Qi,Ren Ma,Bowen Li,He Du,Binyuan Hui,Jinwang Wu,Yuanjun Laili,Conghui H
e
Large Language Models Meet Symbolic Provers for Logical Reasoning Evaluation
First-order logic (FOL) reasoning, which involves sequential deduction, is pivot
al for intelligent systems and serves as a valuable task for evaluating reasonin
g capabilities, particularly in chain-of-thought (CoT) contexts. Existing benchm
arks often rely on extensive human annotation or handcrafted templates, making i
t difficult to achieve the necessary complexity, scalability, and diversity for
robust evaluation. To address these limitations, we propose a novel framework ca
lled ProverGen that synergizes the generative strengths of Large Language Models
 (LLMs) with the rigor and precision of symbolic provers, enabling the creation
of a scalable, diverse, and high-quality FOL reasoning dataset, ProverQA. Prover
QA is also distinguished by its inclusion of accessible and logically coherent i
ntermediate reasoning steps for each problem. Our evaluation shows that state-of
-the-art LLMs struggle to solve ProverQA problems, even with CoT prompting, high
lighting the dataset's challenging nature. We also finetune Llama3.1-8B-Instruct
 on a separate training set generated by our framework.
The finetuned model demonstrates consistent improvements on both in-distribution
 and out-of-distribution test sets, suggesting the value of our proposed data ge
neration framework. Code available at: \url{https://github.com/opendatalab/Prove
rGen}
****
Jiashun Jin,Jingming Wang
Fitting Networks with a Cancellation Trick
The degree-corrected block model (DCBM), latent space model (LSM), and β-m
odel are all popular network models. We combine their modeling ideas and propose
 the logit-DCBM as a new model. Similar as the β-model and LSM, the logit-
DCBM contains nonlinear factors, where fitting the parameters is a challenging o
pen problem. We resolve this problem by introducing a cancellation trick. We als
o propose R-SCORE as a recursive community detection algorithm, where in each it

eration, we first use the idea above to update our parameter estimation, and the n use the results to remove the nonlinear factors in the logit-DCBM so the renor malized model approximately satisfies a low-rank model, just like the DCBM. Our numerical study suggests that R-SCORE significantly improves over existing spect ral approaches in many cases. Also, theoretically, we show that the Hamming err or rate of R-SCORE is faster than that of SCORE in a specific sparse region, and is at least as fast outside this region.
****
Lin-Zhuo Chen,Kangjie Liu,Youtian Lin,Zhihao Li,Siyu Zhu,Xun Cao,Yao Yao
Flow Distillation Sampling: Regularizing 3D Gaussians with Pre-trained Matching Priors
3D Gaussian Splatting (3DGS) has achieved excellent rendering quality with fast training and rendering speed. However, its optimization process lacks explicit g eometric constraints, leading to suboptimal geometric reconstruction in regions with sparse or no observational input views. In this work, we try to mitigate th e issue by incorporating a pre-trained matching prior to the 3DGS optimization p rocess. We introduce Flow Distillation Sampling (FDS), a technique that leverage s pre-trained geometric knowledge to bolster the accuracy of the Gaussian radian ce field. Our method employs a strategic sampling technique to target unobserved views adjacent to the input views, utilizing the optical flow calculated from t he matching model (Prior Flow) to guide the flow analytically calculated from th e 3DGS geometry (Radiance Flow). Comprehensive experiments in depth rendering, m esh reconstruction, and novel view synthesis showcase the significant advantages of FDS over state-of-the-art methods. Additionally, our interpretive experiment s and analysis aim to shed light on the effects of FDS on geometric accuracy and rendering quality, potentially providing readers with insights into its perform ance.
****
Arnas Uselis,Seong Joon Oh
Intermediate Layer Classifiers for OOD generalization
Deep classifiers are known to be sensitive to data distribution shifts, primaril y due to their reliance on spurious correlations in training data. It has been s uggested that these classifiers can still find useful features in the network's last layer that hold up under such shifts. In this work, we question the use of last-layer representations for out-of-distribution (OOD) generalisation and expl ore the utility of intermediate layers. To this end, we introduce \textit{Interm ediate Layer Classifiers} (ILCs). We discover that intermediate layer representa tions frequently offer substantially better generalisation than those from the p enultimate layer. In many cases, zero-shot OOD generalisation using earlier-laye r representations approaches the few-shot performance of retraining on penultima te layer representations. This is confirmed across multiple datasets, architectu res, and types of distribution shifts. Our analysis suggests that intermediate l ayers are less sensitive to distribution shifts compared to the penultimate laye r. These findings highlight the importance of understanding how information is d istributed across network layers and its role in OOD generalisation, while also pointing to the limits of penultimate layer representation utility. Code is avai lable at https://github.com/oshapio/intermediate-layer-generalization.
****
Will Dorrell,Kyle Hsu,Luke Hollingsworth,Jin Hwa Lee,Jiajun Wu,Chelsea Finn,Pete r E. Latham,Timothy Edward John Behrens,James C. R. Whittington
Range, not Independence, Drives Modularity in Biologically Inspired Representati ons
Why do biological and artificial neurons sometimes modularise, each encoding a s ingle meaningful variable, and sometimes entangle their representation of many v ariables? In this work, we develop a theory of when biologically inspired networ ks---those that are nonnegative and energy efficient---modularise their represen tation of source variables (sources). We derive necessary and sufficient conditi ons on a sample of sources that determine whether the neurons in an optimal biol ogically-inspired linear autoencoder modularise. Our theory applies to any datas et, extending far beyond the case of statistical independence studied in previou

s work. Rather we show that sources modularise if their support is ``sufficientl
y spread''. From this theory, we extract and validate predictions in a variety o
f empirical studies on how data distribution affects modularisation in nonlinear
 feedforward and recurrent neural networks trained on supervised and unsupervise
d tasks. Furthermore, we apply these ideas to neuroscience data, showing that ra
nge independence can be used to understand the mixing or modularising of spatial
 and reward information in entorhinal recordings in seemingly conflicting experi
ments. Further, we use these results to suggest alternate origins of mixed-selec
tivity, beyond the predominant theory of flexible nonlinear classification. In s
um, our theory prescribes precise conditions on when neural activities modularis
e, providing tools for inducing and elucidating modular representations in brain
s and machines.
****
Wenyue Hua,Mengting Wan,JAGANNATH SHASHANK SUBRAMANYA SAI VADREVU,Ryan Nadel,Yon
gfeng Zhang,Chi Wang
Interactive Speculative Planning: Enhance Agent Efficiency through Co-design of
System and User Interface
Agents, as user-centric tools, are increasingly deployed for human task delegati
on, assisting with a broad spectrum of requests by generating thoughts, engaging
 with user proxies, and producing action plans. However, agents based on large l
anguage models often face substantial planning latency due to two primary factor
s: the efficiency limitations of the underlying LLMs due to their large size and
 high demand, and the structural complexity of the agents due to the extensive g
eneration of intermediate steps to produce the final output. Given that ineffici
ency in service provision can undermine the value of automation for users, this
paper presents a human-centered efficient agent planning method – Interactive Sp
eculative Planning – aiming at enhancing the efficiency of agent planning throug
h both system design and user interaction. Our approach advocates for the co-des
ign of the agent system and user interface, underscoring the importance of an ag
ent system that can fluidly manage user interactions and interruptions. By integ
rating human interruptions as a fundamental component of the system, we not only
 make it more user-centric but also expedite the entire process by leveraging hu
man-in-the-loop interactions to provide accurate intermediate steps.
****
Kiljae Lee,Yuan Zhang
Leave-One-Out Stable Conformal Prediction
Conformal prediction (CP) is an important tool for distribution-free predictive
uncertainty quantification.
Yet, a major challenge is to balance computational efficiency and prediction acc
uracy, particularly for multiple predictions.
We propose **L**eave-**O**ne-**O**ut **Stab**le **C**onformal **P**rediction (LO
O-StabCP), a novel method to speed up full conformal using algorithmic stability
 without sample splitting.
By leveraging *leave-one-out* stability, our method is much faster in handling a
 large number of prediction requests compared to existing method RO-StabCP based
 on *replace-one* stability.
We derived stability bounds for several popular machine learning tools: regulari
zed loss minimization (RLM) and stochastic gradient descent (SGD), as well as ke
rnel method, neural networks and bagging.
Our method is theoretically justified and demonstrates superior numerical perfor
mance on synthetic and real-world data.
We applied our method to a screening problem, where its effective exploitation o
f training data led to improved test power compared to state-of-the-art method b
ased on split conformal.
****
Yunfan LU,Yanlin Qian,Ziyang Rao,Junren Xiao,Liming Chen,Hui Xiong
RGB-Event ISP: The Dataset and Benchmark
Event-guided imaging has received significant attention due to its potential to
revolutionize instant imaging systems. However, the prior methods primarily focu
s on enhancing RGB images in a post-processing manner, neglecting the challenges

of image signal processor (ISP) dealing with event sensor and the benefits even ts provide for reforming the ISP process. To achieve this, we conduct the first research on event-guided ISP. First, we present a new event-RAW paired dataset, collected with a novel but still confidential sensor that records pixel-level al igned events and RAW images. This dataset includes 3373 RAW images with $2248\ti mes 3264$ resolution and their corresponding events, spanning 24 scenes with 3 e xposure modes and 3 lenses. Second, we propose a convential ISP pipeline to gene rate good RGB frames as reference. This convential ISP pipleline performs basic ISP operations, e.g., demosaicing, white balancing, denoising and color space tr ansforming, with a ColorChecker as reference. Third, we classify the existing le arnable ISP methods into 3 classes, and select multiple methods to train and eva luate on our new dataset. Lastly, since there is no prior work for reference, we propose a simple event-guided ISP method and test it on our dataset. We further put forward key technical challenges and future directions in RGB-Event ISP. In summary, to the best of our knowledge, this is the very first research focusing on event-guided ISP, and we hope it will inspire the community.
****

Dazhou Yu,Genpei Zhang,Liang Zhao

PolyhedronNet: Representation Learning for Polyhedra with Surface-attributed Gra ph

Ubiquitous geometric objects can be precisely and efficiently represented as pol yhedra. The transformation of a polyhedron into a vector, known as polyhedra rep resentation learning, is crucial for manipulating these shapes with mathematical and statistical tools for tasks like classification, clustering, and generation . Recent years have witnessed significant strides in this domain, yet most effor ts focus on the vertex sequence of a polyhedron, neglecting the complex surface modeling crucial in real-world polyhedral objects.

This study proposes \textbf{PolyhedronNet}, a general framework tailored for lea rning representations of 3D polyhedral objects. We propose the concept of the s urface-attributed graph to seamlessly model the vertices, edges, faces, and thei r geometric interrelationships within a polyhedron.

To effectively learn the representation of the entire surface-attributed graph, we first propose to break it down into local rigid representations to effectivel y learn each local region's relative positions against the remaining regions wit hout geometric information loss. Subsequently, we propose PolyhedronGNN to hiera rchically aggregate the local rigid representation via intra-face and inter-face geometric message passing modules, to obtain a global representation that minim izes information loss while maintaining rotation and translation invariance.

Our experimental evaluations on four distinct datasets, encompassing both classi fication and retrieval tasks, substantiate PolyhedronNet's efficacy in capturing comprehensive and informative representations of 3D polyhedral objects.
****

Zhen Han,Zeyinzi Jiang,Yulin Pan,Jingfeng Zhang,Chaojie Mao,Chen-Wei Xie,Yu Liu, Jingren Zhou

ACE: All-round Creator and Editor Following Instructions via Diffusion Transform er

Diffusion models have emerged as a powerful generative technology and have been found to be applicable in various scenarios. Most existing foundational diffusio n models are primarily designed for text-guided visual generation and do not sup port multi-modal conditions, which are essential for many visual editing tasks. This limitation prevents these foundational diffusion models from serving as a u nified model in the field of visual generation, like GPT-4 in the natural langua ge processing field. In this work, we propose ACE, an All-round Creator and Edit or, which achieves comparable performance compared to those expert models in a w ide range of visual generation tasks. To achieve this goal, we first introduce a unified condition format termed Long-context Condition Unit (LCU), and propose a novel Transformer-based diffusion model that uses LCU as input, aiming for joi nt training across various generation and editing tasks. Furthermore, we propose an efficient data collection approach to address the issue of the absence of av ailable training data. It involves acquiring pairwise images with synthesis-base

d or clustering-based pipelines and supplying these pairs with accurate textual instructions by leveraging a fine-tuned multi-modal large language model. To com prehensively evaluate the performance of our model, we establish a benchmark of manually annotated pairs data across a variety of visual generation tasks. The e xtensive experimental results demonstrate the superiority of our model in visual generation fields. Thanks to the all-in-one capabilities of our model, we can e asily build a multi-modal chat system that responds to any interactive request f or image creation using a single model to serve as the backend, avoiding the cum bersome pipeline typically employed in visual agents.
****

Yizhuo Lu,Changde Du,Chong Wang,Xuanliu Zhu,Liuyun Jiang,Xujin Li,Huiguang He
Animate Your Thoughts: Reconstruction of Dynamic Natural Vision from Human Brain Activity
Reconstructing human dynamic vision from brain activity is a challenging task wi th great scientific significance. Although prior video reconstruction methods h ave made substantial progress, they still suffer from several limitations, inclu ding: (1) difficulty in simultaneously reconciling semantic (e.g. categorical de scriptions), structure (e.g. size and color), and consistent motion information (e.g. order of frames); (2) low temporal resolution of fMRI, which poses a chall enge in decoding multiple frames of video dynamics from a single fMRI frame; (3) reliance on video generation models, which introduces ambiguity regarding wheth er the dynamics observed in the reconstructed videos are genuinely derived from fMRI data or are hallucinations from generative model. To overcome these limitat ions, we propose a two-stage model named Mind-Animator. During the fMRI-to-feat ure stage, we decouple semantic, structure, and motion features from fMRI. Speci fically, we employ fMRI-vision-language tri-modal contrastive learning to decode semantic feature from fMRI and design a sparse causal attention mechanism for d ecoding multi-frame video motion features through a next-frame-prediction task. In the feature-to-video stage, these features are integrated into videos using a n inflated Stable Diffusion, effectively eliminating external video data interfe rence. Extensive experiments on multiple video-fMRI datasets demonstrate that o ur model achieves state-of-the-art performance. Comprehensive visualization anal yses further elucidate the interpretability of our model from a neurobiological perspective. Project page: https://mind-animator-design.github.io/.
****

Yudi Zhang,Pei Xiao,Lu Wang,Chaoyun Zhang,Meng Fang,Yali Du,Yevgeniy Puzyrev,Ran dolph Yao,Si Qin,Qingwei Lin,Mykola Pechenizkiy,Dongmei Zhang,Saravan Rajmohan,Q i Zhang
RuAG: Learned-rule-augmented Generation for Large Language Models
In-context learning (ICL) and Retrieval-Augmented Generation (RAG) have gained a ttention for their ability to enhance LLMs' reasoning by incorporating external knowledge but suffer from limited contextual window size, leading to insufficien t information injection. To this end, we propose a novel framework to automatica lly distill large volumes of offline data into interpretable first-order logic r ules, which are injected into LLMs to boost their reasoning capabilities. Our me thod begins by formulating the search process relying on LLMs' commonsense, wher e LLMs automatically define head and body predicates. Then, we apply Monte Carlo Tree Search (MCTS) to address the combinational searching space and efficientl y discover logic rules from data. The resulting logic rules are translated into natural language, allowing targeted knowledge injection and seamless integration into LLM prompts for LLM's downstream task reasoning. We evaluate our framework on public and private industrial tasks, including Natural Language Processing (NLP), time-series, decision-making, and industrial tasks, demonstrating its effe ctiveness in enhancing LLM's capability over diverse tasks.
****

Parjanya Prajakta Prashant,Ignavier Ng,Kun Zhang,Biwei Huang
Differentiable Causal Discovery for Latent Hierarchical Causal Models
Discovering causal structures with latent variables from observational data is a fundamental challenge in causal discovery. Existing methods often rely on const raint-based, iterative discrete searches, limiting their scalability for large n

umbers of variables. Moreover, these methods frequently assume linearity or inve
rtibility, restricting their applicability to real-world scenarios. We present n
ew theoretical results on the identifiability of non-linear latent hierarchical
causal models, relaxing previous assumptions in the literature about the determi
nistic nature of latent variables and exogenous noise. Building on these insight
s, we develop a novel differentiable causal discovery algorithm that efficiently
 estimates the structure of such models. To the best of our knowledge, this is t
he first work to propose a differentiable causal discovery method for non-linear
 latent hierarchical models. Our approach outperforms existing methods in both a
ccuracy and scalability. Furthermore, we demonstrate its practical utility by le
arning interpretable hierarchical latent structures from high-dimensional image
data and demonstrate its effectiveness on downstream tasks such as transfer lear
ning.
****
Yiming Zhang,Jianfeng Chi,Hailey Nguyen,Kartikeya Upasani,Daniel M. Bikel,Jason
E Weston,Eric Michael Smith
Backtracking Improves Generation Safety
Text generation has a fundamental limitation almost by definition: there is no t
aking back tokens that have been generated, even when they are clearly problemat
ic.
In the context of language model safety, when a partial unsafe generation is pro
duced, language models by their nature tend to happily keep on generating simila
rly unsafe additional text.
This is in fact how safety alignment of frontier models gets circumvented in the
 wild, despite great efforts in improving their safety.
Deviating from the paradigm of approaching safety alignment as prevention (decre
asing the probability of harmful responses), we propose backtracking, a techniqu
e that allows language models to "undo" and recover from their own unsafe genera
tion through the introduction of a special [RESET] token.
Our method can be incorporated into either SFT or DPO training to optimize helpf
ulness and harmlessness.
We show that models trained to backtrack are consistently safer than baseline mo
dels: backtracking Llama-3-8B is four times more safe than the baseline model (6
.1\% $\to$ 1.5\%) in our evaluations without regression in helpfulness.
Our method additionally provides protection against four adversarial attacks inc
luding an adaptive attack, despite not being trained to do so.
****
Yang Zhou,Hao Shao,Letian Wang,Steven L. Waslander,Hongsheng Li,Yu Liu
SmartPretrain: Model-Agnostic and Dataset-Agnostic Representation Learning for M
otion Prediction
Predicting the future motion of surrounding agents is essential for autonomous v
ehicles (AVs) to operate safely in dynamic, human-robot-mixed environments. Howe
ver, the scarcity of large-scale driving datasets has hindered the development o
f robust and generalizable motion prediction models, limiting their ability to c
apture complex interactions and road geometries. Inspired by recent advances in
natural language processing (NLP) and computer vision (CV), self-supervised lear
ning (SSL) has gained significant attention in the motion prediction community f
or learning rich and transferable scene representations. Nonetheless, existing p
re-training methods for motion prediction have largely focused on specific model
 architectures and single dataset, limiting their scalability and generalizabili
ty.
To address these challenges, we propose SmartPretrain, a general and scalable SS
L framework for motion prediction that is both model-agnostic and dataset-agnost
ic. Our approach integrates contrastive and reconstructive SSL, leveraging the s
trengths of both generative and discriminative paradigms to effectively represen
t spatiotemporal evolution and interactions without imposing architectural const
raints. Additionally, SmartPretrain employs a dataset-agnostic scenario sampling
 strategy that integrates multiple datasets, enhancing data volume, diversity, a
nd robustness.
Extensive experiments on multiple datasets demonstrate that SmartPretrain consis

tently improves the performance of state-of-the-art prediction models across dat
asets, data splits and main metrics. For instance, SmartPretrain significantly r
educes the MissRate of Forecast-MAE by 10.6\%. These results highlight SmartPret
rain's effectiveness as a unified, scalable solution for motion prediction, brea
king free from the limitations of the small-data regime.
****
Haobin Li,Peng Hu,Qianjun Zhang,Xi Peng,XitingLiu,Mouxing Yang
Test-time Adaptation for Cross-modal Retrieval with Query Shift
The success of most existing cross-modal retrieval methods heavily relies on the
 assumption that the given queries follow the same distribution of the source do
main.
However, such an assumption is easily violated in real-world scenarios due to th
e complexity and diversity of queries, thus leading to the query shift problem.
Specifically, query shift refers to the online query stream originating from the
 domain that follows a different distribution with the source one.
In this paper, we observe that query shift would not only diminish the uniformit
y (namely, within-modality scatter) of the query modality but also amplify the g
ap between query and gallery modalities.
Based on the observations, we propose a novel method dubbed Test-time adaptation
 for Cross-modal Retrieval (TCR).
In brief, TCR employs a novel module to refine the query predictions (namely, re
trieval results of the query) and a joint objective to prevent query shift from
disturbing the common space, thus achieving online adaptation for the cross-moda
l retrieval models with query shift.
Expensive experiments demonstrate the effectiveness of the proposed TCR against
query shift.
Code is available at https://github.com/XLearning-SCU/2025-ICLR-TCR.
****
Timothee Leleu,Sam Reifenstein
Non-Equilibrium Dynamics of Hybrid Continuous-Discrete Ground-State Sampling
We propose a general framework for a hybrid continuous-discrete algorithm that i
ntegrates continuous-time deterministic dynamics with Metropolis-Hastings (MH) s
teps to combine search dynamics that either preserve or break detailed balance.
Our purpose is to study the non-equilibrium dynamics that leads to the ground st
ate of rugged energy landscapes in this general setting. Our results show that M
H-driven dynamics reach ``easy'' ground states more quickly, indicating a strong
er bias toward these solutions in algorithms using reversible transition probabi
lities. To validate this, we construct a set of Ising problem instances with a c
ontrollable bias in the energy landscape that makes certain degenerate solutions
 more accessible than others. The constructed hybrid algorithm demonstrates sign
ificant improvements in convergence and ground-state sampling accuracy, achievin
g a 100x speedup on GPU compared to simulated annealing, making it well-suited f
or large-scale applications.
****
Zeou Hu,Yaoliang Yu
Leveraging Variable Sparsity to Refine Pareto Stationarity in Multi-Objective Op
timization
Gradient-based multi-objective optimization (MOO) is essential in modern machine
 learning, with applications in e.g., multi-task learning, federated learning,
algorithmic fairness and reinforcement learning. In this work, we first reveal s
ome limitations of Pareto stationarity, a widely accepted first-order condition
for Pareto optimality, in the presence of sparse function-variable structures. N
ext, to account for such sparsity, we propose a novel solution concept termed Re
fined Pareto Stationarity (RPS), which we prove is always sandwiched between Par
eto optimality and Pareto stationarity. We give an efficient partitioning algori
thm to automatically mine the function-variable dependency and substantially tri
m non-optimal Pareto stationary solutions. Then, we show that gradient-based des
cent algorithms in MOO can be enhanced with our refined partitioning. In particu
lar, we propose Multiple Gradient Descent Algorithm with Refined Partition (RP-M
GDA) as an example method that converges to RPS, while still enjoying a similar

per-step complexity and convergence rate. Lastly, we validate our approach throu gh experiments on both synthetic examples and realistic application scenarios wh ere distinct function-variable dependency structures appear. Our results highlig ht the importance of exploiting function-variable structure in gradient-based MO O, and provide a seamless enhancement to existing approaches.
****

Linda He,Jue WANG,Maurice Weber,Shang Zhu,Ben Athiwaratkun,Ce Zhang
Scaling Instruction-tuned LLMs to Million-token Contexts via Hierarchical Synthe tic Data Generation
Large Language Models (LLMs) struggle with long-context reasoning, not only due to the quadratic scaling of computational complexity with sequence length but al so because of the scarcity and expense of annotating long-context data. There ha s been barely any open-source work that systematically ablates long-context data , nor is there any openly available instruction tuning dataset with contexts sur passing 100K tokens. To bridge this gap, we introduce a novel post-training synt hetic data generation strategy designed to efficiently extend the context window of LLMs while preserving their general task performance. Our approach scalably extends to arbitrarily long context lengths, unconstrained by the length of avai lable real-world data, which effectively addresses the scarcity of raw long-cont ext data.
Through a step-by-step rotary position embedding (RoPE) scaling training strateg y, we demonstrate that our model, with a context length of up to 1M tokens, perf orms well on the RULER benchmark and InfiniteBench and maintains robust performa nce on general language tasks.
****

Fei YE,Zaixiang Zheng,Dongyu Xue,Yuning Shen,Lihao Wang,Yiming Ma,Yan Wang,Xinyo u Wang,Xiangxin Zhou,Quanquan Gu
ProteinBench: A Holistic Evaluation of Protein Foundation Models
Recent years have witnessed a surge in the development of protein foundation mod els, significantly improving performance in protein prediction and generative ta sks ranging from 3D structure prediction and protein design to conformational dy namics. However, the capabilities and limitations associated with these models r emain poorly understood due to the absence of a unified evaluation framework. To fill this gap, we introduce ProteinBench, a holistic evaluation framework desig ned to enhance the transparency of protein foundation models. Our approach consi sts of three key components: (i) A taxonomic classification of tasks that broadl y encompass the main challenges in the protein domain, based on the relationship s between different protein modalities; (ii) A multi-metric evaluation approach that assesses performance across four key dimensions: quality, novelty, diversit y, and robustness; and (iii) In-depth analyses from various user objectives, pro viding a holistic view of model performance. Our comprehensive evaluation of pro tein foundation models reveals several key findings that shed light on their cur rent capabilities and limitations. To promote transparency and facilitate furthe r research, we release the evaluation dataset, code, and a public leaderboard pu blicly for further analysis and a general modular toolkit. We intend for Protein Bench to be a living benchmark for establishing a standardized, in-depth evaluat ion framework for protein foundation models, driving their development and appli cation while fostering collaboration within the field.
****

Peng Jing,Meiqi Yang,Qiong Zhang,Xiaoxiao Li
S4M: S4 for multivariate time series forecasting with Missing values
Multivariate time series data play a pivotal role in a wide range of real-world applications, such as finance, healthcare, and meteorology, where accurate forec asting is critical for informed decision-making and proactive interventions. How ever, the presence of block missing data introduces significant challenges, ofte n compromising the performance of predictive models. Traditional two-step approa ches, which first impute missing values and then perform forecasting, are prone to error accumulation, particularly in complex multivariate settings characteriz ed by high missing ratios and intricate dependency structures. In this work, we introduce S4M, an end-to-end time series forecasting framework that seamlessly i

ntegrates missing data handling into the Structured State Space Sequence (S4) mo
del architecture. Unlike conventional methods that treat imputation as a separat
e preprocessing step, S4M leverages the latent space of S4 models to directly re
cognize and represent missing data patterns, thereby more effectively capturing
the underlying temporal and multivariate dependencies. Our framework comprises t
wo key components: the Adaptive Temporal Prototype Mapper (ATPM) and the Missing
-Aware Dual Stream S4 (MDS-S4). The ATPM employs a prototype bank to derive robu
st and informative representations from historical data patterns, while the MDS-
S4 processes these representations alongside missingness masks as dual input str
eams to enable accurate forecasting. Through extensive empirical evaluations on
diverse real-world datasets, we demonstrate that S4M consistently achieves state
-of-the-art performance. These results underscore the efficacy of our integrated
 approach in handling missing data, showcasing its robustness and superiority ov
er traditional imputation-based methods. Our findings highlight the potential of
 S4M to advance reliable time series forecasting in practical applications, offe
ring a promising direction for future research and deployment. Code is available
 at https://github.com/WINTERWEEL/S4M.git.
****

Varun Mulchandani,Jung-Eun Kim

Severing Spurious Correlations with Data Pruning

Deep neural networks have been shown to learn and rely on spurious correlations
present in the data that they are trained on. Reliance on such correlations can
cause these networks to malfunction when deployed in the real world, where these
 correlations may no longer hold. To overcome the learning of and reliance on su
ch correlations, recent studies propose approaches that yield promising results.
 These works, however, study settings where the strength of the spurious signal
is significantly greater than that of the core, invariant signal, making it easi
er to detect the presence of spurious features in individual training samples an
d allow for further processing. In this paper, we identify new settings where th
e strength of the spurious signal is relatively weaker, making it difficult to d
etect any spurious information while continuing to have catastrophic consequence
s. We also discover that spurious correlations are learned primarily due to only
 a handful of all the samples containing the spurious feature and develop a nove
l data pruning technique that identifies and prunes small subsets of the trainin
g data that contain these samples. Our proposed technique does not require infer
red domain knowledge, information regarding the sample-wise presence or nature o
f spurious information, or human intervention. Finally, we show that such data p
runing attains state-of-the-art performance on previously studied settings where
 spurious information is identifiable.
****

Tianyun Yang,Ziniu Li,Juan Cao,Chang Xu

Understanding and Mitigating Hallucination in Large Vision-Language Models via M
odular Attribution and Intervention

Large Vision-Language Models (LVLMs) exhibit impressive capabilities in complex
visual tasks but are prone to hallucination, especially in open-ended generation
 tasks. This paper explores why LVLMs tend to hallucinate and how to mitigate it
. First, we conduct causal mediation analysis through counterfactual edits on sp
ecific modules in LVLMs. Our results disclose that Multi-Head Attention (MHA) mo
dules contribute more to the probability of generating hallucination words than
multi-layer perceptron modules. We then identify specific heads that are respons
ible for hallucination, referred to as hallucination heads. Second, we examine t
he behavior of hallucination heads. We find that they are concentrated in the mi
ddle and deeper layers, displaying a strong attention bias toward text tokens. F
urther, we show that the attention patterns of certain hallucination heads exhib
it greater similarity to the base language model and change slowly during the in
struction tuning process. Finally, we propose two simple yet effective methods t
o mitigate hallucination: one is training-free and can be applied directly durin
g decoding, while the other involves fine-tuning. Both methods are targeted for
hallucination heads to reduce their reliance on text tokens. Notably, our method
s achieve up to 1.7x reduction in hallucination rate for the LLaVA-v1.5-7B model

in COCO captioning task, outperforming existing baselines. Overall, our finding s suggest that hallucinations in LVLMs are likely to stem from certain modules, and targeted interventions can effectively mitigate these issues.
****

Dongmin Park,Sebin Kim,Taehong Moon,Minkyu Kim,Kangwook Lee,Jaewoong Cho
Rare-to-Frequent: Unlocking Compositional Generation Power of Diffusion Models o n Rare Concepts with LLM Guidance
State-of-the-art text-to-image (T2I) diffusion models often struggle to generate rare compositions of concepts, e.g., objects with unusual attributes. In this p aper, we show that the compositional generation power of diffusion models on suc h rare concepts can be significantly enhanced by the Large Language Model (LLM) guidance. We start with empirical and theoretical analysis, demonstrating that e xposing frequent concepts relevant to the target rare concepts during the diffus ion sampling process yields more accurate concept composition. Based on this, we propose a training-free approach, R2F, that plans and executes the overall rare -to-frequent concept guidance throughout the diffusion inference by leveraging t he abundant semantic knowledge in LLMs. Our framework is flexible across any pre -trained diffusion models and LLMs, and can be seamlessly integrated with the re gion-guided diffusion approaches. Extensive experiments on three datasets, inclu ding our newly proposed benchmark, RareBench, containing various prompts with ra re compositions of concepts, R2F significantly surpasses existing models includi ng SD3.0 and FLUX by up to 28.1%p in T2I alignment. Code is available at https:/ /github.com/krafton-ai/Rare-to-Frequent.
****

Guangkai Xu,Yongtao Ge,Mingyu Liu,Chengxiang Fan,Kangyang Xie,Zhiyue Zhao,Hao Ch en,Chunhua Shen
What Matters When Repurposing Diffusion Models for General Dense Perception Task s?
Extensive pre-training with large data is indispensable for downstream geometry and semantic visual perception tasks. Thanks to large-scale text-to-image (T2I) pretraining, recent works show promising results by simply fine-tuning T2I diffu sion models for a few dense perception tasks. However, several crucial design de cisions in this process still lack comprehensive justification, encompassing the necessity of the multi-step diffusion mechanism, training strategy, inference e nsemble strategy, and fine-tuning data quality. In this work, we conduct a thoro ugh investigation into critical factors that affect transfer efficiency and perf ormance when using diffusion priors. Our key findings are: 1) High-quality fine- tuning data is paramount for both semantic and geometry perception tasks. 2) As a special case of the diffusion scheduler by setting its hyper-parameters, the m ulti-step generation can be simplified to a one-step fine-tuning paradigm withou t any loss of performance, while significantly speeding up inference. 3) Apart f rom fine-tuning the diffusion model with only latent space supervision, task-spe cific supervision can be beneficial to enhance fine-grained details. These obser vations culminate in the development of GenPercept, an effective deterministic o ne-step fine-tuning paradigm tailored for dense visual perception tasks exploiti ng diffusion priors. Different from the previous multi-step methods, our paradig m offers a much faster inference speed, and can be seamlessly integrated with cu stomized perception decoders and loss functions for task-specific supervision, w hich can be critical for improving the fine-grained details of predictions. Comp rehensive experiments on a diverse set of dense visual perceptual tasks, includi ng monocular depth estimation, surface normal estimation, image segmentation, an d matting, are performed to demonstrate the remarkable adaptability and effectiv eness of our proposed method. Code: https://github.com/aim-uofa/GenPercept
****

Jaehyeon Son,Soochan Lee,Gunhee Kim
Distilling Reinforcement Learning Algorithms for In-Context Model-Based Planning
Recent studies have shown that Transformers can perform in-context reinforcement learning (RL) by imitating existing RL algorithms, enabling sample-efficient ad aptation to unseen tasks without parameter updates. However, these models also i nherit the suboptimal behaviors of the RL algorithms they imitate. This issue pr

imarily arises due to the gradual update rule employed by those algorithms. Mode
l-based planning offers a promising solution to this limitation by allowing the
models to simulate potential outcomes before taking action, providing an additio
nal mechanism to deviate from the suboptimal behavior. Rather than learning a se
parate dynamics model, we propose Distillation for In-Context Planning (DICP), a
n in-context model-based RL framework where Transformers simultaneously learn en
vironment dynamics and improve policy in-context. We evaluate DICP across a rang
e of discrete and continuous environments, including Darkroom variants and Meta-
World. Our results show that DICP achieves state-of-the-art performance while re
quiring significantly fewer environment interactions than baselines, which inclu
de both model-free counterparts and existing meta-RL methods.
****
GUOJUN XIONG,Shufan Wang,Daniel Jiang,Jian Li
On the Linear Speedup of Personalized Federated Reinforcement Learning with Shar
ed Representations
Federated reinforcement learning (FedRL) enables multiple agents to collaborativ
ely learn a policy without needing to share the local trajectories collected dur
ing agent-environment interactions. However, in practice, the environments faced
 by different agents are often heterogeneous, but since existing FedRL algorithm
s learn a single policy across all agents, this may lead to poor performance. In
 this paper, we introduce a personalized FedRL framework (PFedRL) by taking adva
ntage of possibly shared common structure among agents in heterogeneous environm
ents. Specifically, we develop a class of PFedRL algorithms named PFedRL-Rep tha
t learns (1) a shared feature representation collaboratively among all agents, a
nd (2) an agent-specific weight vector personalized to its local environment. We
 analyze the convergence of PFedTD-Rep, a particular instance of the framework w
ith temporal difference (TD) learning and linear representations. To the best of
 our knowledge, we are the first to prove a linear convergence speedup with resp
ect to the number of agents in the PFedRL setting. To achieve this, we show that
 PFedTD-Rep is an example of federated two-timescale stochastic approximation wi
th Markovian noise. Experimental results demonstrate that PFedTD-Rep, along with
 an extension to the control setting based on deep Q-networks (DQN), not only im
prove learning in heterogeneous settings, but also provide better generalization
 to new environments.
****
Minsu Kim,Sanghyeok Choi,Taeyoung Yun,Emmanuel Bengio,Leo Feng,Jarrid Rector-Bro
oks,Sungsoo Ahn,Jinkyoo Park,Nikolay Malkin,Yoshua Bengio
Adaptive teachers for amortized samplers
Amortized inference is the task of training a parametric model, such as a neural
 network, to approximate a distribution with a given unnormalized density where
exact sampling is intractable. When sampling is modeled as a sequential decision
-making process, reinforcement learning (RL) methods, such as generative flow ne
tworks, can be used to train the sampling policy. Off-policy RL training facilit
ates the discovery of diverse, high-reward candidates, but existing methods stil
l face challenges in efficient exploration. We propose to use an adaptive traini
ng distribution (the Teacher) to guide the training of the primary amortized sam
pler (the Student). The Teacher, an auxiliary behavior model, is trained to sam
ple high-loss regions of the Student and can generalize across unexplored modes,
 thereby enhancing mode coverage by providing an efficient training curriculum.
We validate the effectiveness of this approach in a synthetic environment design
ed to present an exploration challenge, two diffusion-based sampling tasks, and
four biochemical discovery tasks demonstrating its ability to improve sample eff
iciency and mode coverage. Source code is available at https://github.com/alstn1
2088/adaptive-teacher.
****
Tal Amir,Nadav Dym
Fourier Sliced-Wasserstein Embedding for Multisets and Measures
We present the _Fourier Sliced Wasserstein (FSW) embedding_—a novel method to em
bed multisets and measures over $\mathbb{R}^d$ into Euclidean space.

Our proposed embedding approximately preserves the sliced Wasserstein distance o
n distributions, thereby yielding geometrically meaningful representations that
better capture the structure of the input. Moreover, it is injective on measures
 and _bi-Lipschitz_ on multisets—a significant advantage over prevalent methods
based on sum- or max-pooling, which are provably not bi-Lipschitz, and, in many
cases, not even injective.
The required output dimension for these guarantees is near-optimal: roughly $2 N
 d$, where N is the maximal input multiset size.

Furthermore, we prove that it is _impossible_ to embed distributions over $\math
bb{R}^d$ into Euclidean space in a bi-Lipschitz manner. Thus, the metric propert
ies of our embedding are, in a sense, the best possible.

Through numerical experiments, we demonstrate that our method yields superior mu
ltiset representations that improve performance in practical learning tasks. Spe
cifically, we show that (a) a simple combination of the FSW embedding with an ML
P achieves state-of-the-art performance in learning the (non-sliced) Wasserstein
 distance; and (b) replacing max-pooling with the FSW embedding makes PointNet s
ignificantly more robust to parameter reduction, with only minor performance deg
radation even after a 40-fold reduction.
****
Gang Liu,Srijit Seal,John Arevalo,Zhenwen Liang,Anne E Carpenter,Meng Jiang,Shan
tanu Singh
Learning Molecular Representation in a Cell
Predicting drug efficacy and safety in vivo requires information on biological r
esponses (e.g., cell morphology and gene expression) to small molecule perturbat
ions. However, current molecular representation learning methods do not provide
a comprehensive view of cell states under these perturbations and struggle to re
move noise, hindering model generalization. We introduce the Information Alignme
nt (InfoAlign) approach to learn molecular representations through the informati
on bottleneck method in cells. We integrate molecules and cellular response data
 as nodes into a context graph, connecting them with weighted edges based on che
mical, biological, and computational criteria. For each molecule in a training b
atch, InfoAlign optimizes the encoder's latent representation with a minimality
objective to discard redundant structural information. A sufficiency objective d
ecodes the representation to align with different feature spaces from the molecu
le's neighborhood in the context graph. We demonstrate that the proposed suffici
ency objective for alignment is tighter than existing encoder-based contrastive
methods. Empirically, we validate representations from InfoAlign in two downstre
am applications: molecular property prediction against up to 27 baseline methods
 across four datasets, plus zero-shot molecule-morphology matching. The code and
 model are available at https://github.com/liugangcode/InfoAlign.
****
Jiachen Li,Qian Long,Jian Zheng,Xiaofeng Gao,Robinson Piramuthu,Wenhu Chen,Willi
am Yang Wang
T2V-Turbo-v2: Enhancing Video Model Post-Training through Data, Reward, and Cond
itional Guidance Design
In this paper, we focus on enhancing a diffusion-based text-to-video (T2V) model
 during the post-training phase by distilling a highly capable consistency model
 from a pretrained T2V model. Our proposed method, T2V-Turbo-v2, introduces a si
gnificant advancement by integrating various supervision signals, including high
-quality training data, reward model feedback, and conditional guidance, into th
e consistency distillation process. Through comprehensive ablation studies, we h
ighlight the crucial importance of tailoring datasets to specific learning objec
tives and the effectiveness of learning from diverse reward models for enhancing
 both the visual quality and text-video alignment. Additionally, we highlight th
e vast design space of conditional guidance strategies, which centers on designi
ng an effective energy function to augment the teacher ODE solver. We demonstrat
e the potential of this approach by extracting motion guidance from the training
 datasets and incorporating it into the ODE solver, showcasing its effectiveness

in improving the motion quality of the generated videos with the improved motion-related metrics from VBench and T2V-CompBench. Empirically, our T2V-Turbo-v2 establishes a new state-of-the-art result on VBench, **with a Total score of 85.13**, surpassing proprietary systems such as Gen-3 and Kling.
****
Gyu Yeol Kim,Min-hwan Oh
ADAM Optimization with Adaptive Batch Selection
Adam is a widely used optimizer in neural network training due to its adaptive learning rate. However, because different data samples influence model updates to varying degrees, treating them equally can lead to inefficient convergence. To address this, a prior work proposed adapting the sampling distribution using a bandit framework to select samples adaptively. While promising, both the original Adam and its bandit-based variant suffer from flawed theoretical guarantees. In this paper, we introduce Adam with Combinatorial Bandit Sampling (AdamCB), which integrates combinatorial bandit techniques into Adam to resolve these issues. AdamCB is able to fully utilize feedback from multiple actions at once, enhancing both theoretical guarantees and practical performance. Our rigorous regret analysis shows that AdamCB achieves faster convergence than both the original Adam and its variants. Numerical experiments demonstrate that AdamCB consistently outperforms existing Adam-based methods, making it the first to offer both provable guarantees and practical efficiency for Adam with adaptive batch selection.
****
Xiu-Chuan Li,Tongliang Liu
Efficient and Trustworthy Causal Discovery with Latent Variables and Complex Relations
Most traditional causal discovery methods assume that all task-relevant variables are observed, an assumption often violated in practice. Although some recent works allow the presence of latent variables, they typically assume the absence of certain special causal relations to ensure a degree of simplicity, which might also be invalid in real-world scenarios. This paper tackles a challenging and important setting where latent and observed variables are interconnected through complex causal relations. Under a pure children assumption ensuring that latent variables leave adequate footprints in observed variables, we develop novel theoretical results, leading to an efficient causal discovery algorithm which is the first one capable of handling the setting with both latent variables and complex relations within polynomial time. Our algorithm first sequentially identifies latent variables from leaves to roots and then sequentially infers causal relations from roots to leaves. Moreover, we prove trustworthiness of our algorithm, meaning that when the assumption is invalid, it can raise an error signal rather than draw an incorrect causal conclusion, thus preventing potential damage to downstream tasks. We demonstrate the efficacy of our algorithm through experiments. Our work significantly enhances efficiency and reliability of causal discovery in complex systems. Our code is available at: https://github.com/XiuchuanLi/ICLR2025-ETCD
****
Zongming Li,Tianheng Cheng,Shoufa Chen,Peize Sun,Haocheng Shen,Longjin Ran,Xiaoxin Chen,Wenyu Liu,Xinggang Wang
ControlAR: Controllable Image Generation with Autoregressive Models
Autoregressive (AR) models have reformulated image generation as next-token prediction, demonstrating remarkable potential and emerging as strong competitors to diffusion models. However, control-to-image generation, akin to ControlNet, remains largely unexplored within AR models. Although a natural approach, inspired by advancements in Large Language Models, is to tokenize control images into tokens and prefill them into the autoregressive model before decoding image tokens, it still falls short in generation quality compared to ControlNet and suffers from inefficiency. To this end, we introduce ControlAR, an efficient and effective framework for integrating spatial controls into autoregressive image generation models. Firstly, we explore control encoding for AR models and propose a lightweight control encoder to transform spatial inputs (e.g., canny edges or depth maps) into control tokens. Then ControlAR exploits the conditional decoding metho

d to generate the next image token conditioned on the per-token fusion between c
ontrol and image tokens, similar to positional encodings. Compared to prefilling
 tokens, using conditional decoding significantly strengthens the control capabi
lity of AR models but also maintains the model efficiency. Furthermore, the prop
osed ControlAR surprisingly empowers AR models with arbitrary-resolution image g
eneration via conditional decoding and specific controls. Extensive experiments
can demonstrate the controllability of the proposed ControlAR for the autoregres
sive control-to-image generation across diverse inputs, including edges, depths,
 and segmentation masks. Furthermore, both quantitative and qualitative results
indicate that ControlAR surpasses previous state-of-the-art
controllable diffusion models, e.g., ControlNet++.
****

Haotong Yang,Yi Hu,Shijia Kang,Zhouchen Lin,Muhan Zhang
Number Cookbook: Number Understanding of Language Models and How to Improve It
Large language models (LLMs) can solve an increasing number of complex reasoning
 tasks while making surprising mistakes in basic numerical understanding and pro
cessing (such as $9.11 > 9.9$). The latter ability is essential for tackling com
plex arithmetic and mathematical problems and serves as a foundation for most re
asoning tasks, but previous work paid little attention to it or only discussed s
everal restricted tasks (like integer addition). In this paper, we comprehensive
ly investigate the numerical understanding and processing ability (NUPA) of LLMs
. Firstly, we introduce a benchmark covering four common numerical representatio
ns and 17 distinct numerical tasks in four major categories, resulting in 41 mea
ningful combinations in total. These tasks are derived from primary and secondar
y education curricula, encompassing nearly all everyday numerical understanding
and processing scenarios, and the rules of these tasks are very simple and clear
.
Through the benchmark, we find that current LLMs fail frequently in many of the
tasks. To study the problem, we train small models with existing and potential t
echniques for enhancing NUPA (such as tokenizers, PEs, and number formats), comp
rehensively evaluating their effectiveness using our testbed. We also finetune p
ractical-scale LLMs on our proposed NUPA tasks and find that 1) naive finetuning
 can improve NUPA a lot on many but not all tasks, and 2) surprisingly, techniqu
es designed to enhance NUPA prove ineffective for finetuning pretrained models.
We further explore the impact of chain-of-thought techniques on NUPA. Our work p
rovides a more detailed and comprehensive understanding of NUPA in LLMs.
****

Ziye Huang,Haoqi Yuan,Yuhui Fu,Zongqing Lu
Efficient Residual Learning with Mixture-of-Experts for Universal Dexterous Gras
ping
Universal dexterous grasping across diverse objects presents a fundamental yet f
ormidable challenge in robot learning. Existing approaches using reinforcement l
earning (RL) to develop policies on extensive object datasets face critical limi
tations, including complex curriculum design for multi-task learning and limited
 generalization to unseen objects.
To overcome these challenges, we introduce ResDex, a novel approach that integra
tes residual policy learning with a mixture-of-experts (MoE) framework. ResDex i
s distinguished by its use of geometry-agnostic base policies that are efficient
ly acquired on individual objects and capable of generalizing across a wide rang
e of unseen objects. Our MoE framework incorporates several base policies to fac
ilitate diverse grasping styles suitable for various objects. By learning residu
al actions alongside weights that combine these base policies, ResDex enables ef
ficient multi-task RL for universal dexterous grasping.
ResDex achieves state-of-the-art performance on the DexGraspNet dataset comprisi
ng 3,200 objects with an 88.8% success rate. It exhibits no generalization gap w
ith unseen objects and demonstrates superior training efficiency, mastering all
tasks within only 12 hours on a single GPU. For further details and videos, visi
t our project page.
****

Sheng Fang,Yongjin Liu,Wei Yao,Chengming Yu,Jin Zhang

qNBO: quasi-Newton Meets Bilevel Optimization
Bilevel optimization, which addresses challenges in hierarchical learning tasks, has gained significant interest in machine learning. Implementing gradient descent for bilevel optimization presents computational hurdles, notably the need to compute the exact lower-level solution and the inverse Hessian of the lower-level objective. While these two aspects are inherently connected, existing methods typically handle them separately by solving the lower-level problem and a linear system for the inverse Hessian-vector product. In this paper, we introduce a general framework to tackle these computational challenges in a coordinated manner. Specifically, we leverage quasi-Newton algorithms to accelerate the solution of the lower-level problem while efficiently approximating the inverse Hessian-vector product. Furthermore, by leveraging the superlinear convergence properties of BFGS, we establish a non-asymptotic convergence analysis for the BFGS adaptation within our framework. Numerical experiments demonstrate the comparable or superior performance of our proposed algorithms in real-world learning tasks, including hyperparameter optimization, data hyper-cleaning, and few-shot meta-learning.
****
Bokai Lin,Zihao Zeng,Zipeng Xiao,Siqi Kou,TianQi Hou,Xiaofeng Gao,Hao Zhang,Zhijie Deng
MatryoshkaKV: Adaptive KV Compression via Trainable Orthogonal Projection
KV cache has become a *de facto* technique for the inference of large language models (LLMs), where tensors of shape (layer number, head number, sequence length, feature dimension) are introduced to cache historical information for self-attention.
As the size of the model and data grows, the KV cache can, yet, quickly become a bottleneck within the system in both storage and memory transfer.
To address this, prior studies usually focus on the first three axes of the cache tensors for compression.
This paper supplements them, focusing on the feature dimension axis,
by utilizing low-rank projection matrices to transform the cache features into spaces with reduced dimensions.
We begin by investigating the canonical orthogonal projection method for data compression through principal component analysis (PCA).
We identify the drawback of PCA projection that model performance degrades rapidly under relatively low compression rates (less than 60%).
This phenomenon is elucidated by insights derived from the principles of attention mechanisms.
To bridge the gap, we propose to directly tune the orthogonal projection matrix on the continual pre-training or supervised fine-tuning datasets with an elaborate Matryoshka learning strategy.
Thanks to such a strategy, we can adaptively search for the optimal compression rates for various layers and heads given varying compression budgets.
Compared to Multi-head Latent Attention (MLA), our method can easily embrace pre-trained LLMs and hold a smooth tradeoff between performance and compression rate.
We witness the high data efficiency of our training procedure and find that our method can sustain over 90\% performance with an average KV cache compression rate of 60% (and up to 75% in certain extreme scenarios) for popular LLMs like LLaMA2 and Mistral.
****
Dongyoung Kim,Kimin Lee,Jinwoo Shin,Jaehyung Kim
Spread Preference Annotation: Direct Preference Judgment for Efficient LLM Alignment
Aligning large language models (LLMs) with human preferences becomes a key component to obtaining state-of-the-art performance, but it yields a huge cost to construct a large human-annotated preference dataset. To tackle this problem, we propose a new framework, Spread Preference Annotation with direct preference judgment (SPA), that boosts the alignment of LLMs using only a very small amount of human-annotated preference data.

Our key idea is leveraging the human prior knowledge within the small (seed) dat
a and progressively improving the alignment of LLM, by iteratively generating th
e responses and learning from them with the self-annotated preference data.
To be specific, we propose to derive the preference label from the logits of LLM
 to explicitly extract the model's inherent preference.
Compared to the previous approaches using external reward models or implicit in-
context learning, we observe that the proposed approach is significantly more ef
fective.
In addition, we introduce a noise-aware preference learning algorithm to mitigat
e the risk of low quality within generated preference data.
Our experimental results demonstrate that the proposed framework significantly b
oosts the alignment of LLMs.
For example, we achieve superior alignment performance on AlpacaEval 2.0 with on
ly 3.3% of the ground-truth preference labels in the Ultrafeedback data compared
 to the cases using the entire data or state-of-the-art baselines.
****
Sheryl Hsu,Omar Khattab,Chelsea Finn,Archit Sharma
Grounding by Trying: LLMs with Reinforcement Learning-Enhanced Retrieval
The hallucinations of large language models (LLMs) are increasingly mitigated by
 allowing LLMs to search for information and to ground their answers in real sou
rces. Unfortunately, LLMs often struggle with posing the right search queries, e
specially when dealing with complex or otherwise indirect topics. Observing that
 LLMs can learn to search for relevant facts by $\textit{trying}$ different quer
ies and learning to up-weight queries that successfully produce relevant results
, we introduce $\underline{Le}$arning to $\underline{Re}$trieve by $\underline{T
}$rying (LeReT), a reinforcement learning framework that explores search queries
 and uses preference-based optimization to improve their quality. LeReT can impr
ove the absolute retrieval accuracy by up to 29\% and the downstream generator e
valuations by 17\%. The simplicity and flexibility of LeReT allows it to be appl
ied to arbitrary off-the-shelf retrievers and makes it a promising technique for
 improving general LLM pipelines.
****
Shouheng Li,Floris Geerts,Dongwoo Kim,Qing Wang
Towards Bridging Generalization and Expressivity of Graph Neural Networks
Expressivity and generalization are two critical aspects of graph neural network
s (GNNs). While significant progress has been made in studying the expressivity
of GNNs, much less is known about their generalization capabilities, particularl
y when dealing with the inherent complexity of graph-structured data.
In this work, we address the intricate relationship between expressivity and gen
eralization in GNNs. Theoretical studies conjecture a trade-off between the two:
 highly expressive models risk overfitting, while those focused on generalizatio
n may sacrifice expressivity. However, empirical evidence often contradicts this
 assumption, with expressive GNNs frequently demonstrating strong generalization
. We explore this contradiction by introducing a novel framework that connects G
NN generalization to the variance in graph structures they can capture. This lea
ds us to propose a k-variance margin-based generalization bound that character
izes the structural properties of graph embeddings in terms of their upper-bound
ed expressive power. Our analysis does not rely on specific GNN architectures, m
aking it broadly applicable across GNN models. We further uncover a trade-off be
tween intra-class concentration and inter-class separation, both of which are cr
ucial for effective generalization. Through case studies and experiments on real
-world datasets, we demonstrate that our theoretical findings align with empiric
al results, offering a deeper understanding of how expressivity can enhance GNN
generalization.
****
Mateusz Pach,Koryna Lewandowska,Jacek Tabor,Bartosz Micha■ Zieli■ski,Dawid Damia
n Rymarczyk
LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable
Computer Vision
Prototypical parts networks combine the power of deep learning with the explaina

bility of case-based reasoning to make accurate, interpretable decisions. They f
ollow the this looks like that reasoning, representing each prototypical part wi
th patches from training images. However, a single image patch comprises multipl
e visual features, such as color, shape, and texture, making it difficult for us
ers to identify which feature is important to the model.

To reduce this ambiguity, we introduce the Lucid Prototypical Parts Network (Luc
idPPN), a novel prototypical parts network that separates color prototypes from
other visual features. Our method employs two reasoning branches: one for non-co
lor visual features, processing grayscale images, and another focusing solely on
 color information. This separation allows us to clarify whether the model's dec
isions are based on color, shape, or texture. Additionally, LucidPPN identifies
prototypical parts corresponding to semantic parts of classified objects, making
 comparisons between data classes more intuitive, e.g., when two bird species mi
ght differ primarily in belly color.

Our experiments demonstrate that the two branches are complementary and together
 achieve results comparable to baseline methods. More importantly, LucidPPN gene
rates less ambiguous prototypical parts, enhancing user understanding.
****
Kamel Alrashedy,Pradyumna Tambwekar,Zulfiqar Haider Zaidi,Megan Langwasser,Wei X
u,Matthew Gombolay
Generating CAD Code with Vision-Language Models for 3D Designs
Generative AI has transformed the fields of Design and Manufacturing by providin
g
efficient and automated methods for generating and modifying 3D objects. One
approach involves using Large Language Models (LLMs) to generate Computer-
Aided Design (CAD) scripting code, which can then be executed to render a 3D
object; however, the resulting 3D object may not meet the specified requirements
.
Testing the correctness of CAD generated code is challenging due to the complexi
ty
and structure of 3D objects (e.g., shapes, surfaces, and dimensions) that are no
t
feasible in code. In this paper, we introduce CADCodeVerify, a novel approach to
iteratively verify and improve 3D objects generated from CAD code. Our approach
works by producing ameliorative feedback by prompting a Vision-Language Model
(VLM) to generate and answer a set of validation questions to verify the generat
ed
object and prompt the VLM to correct deviations. To evaluate CADCodeVerify, we
introduce, CADPrompt, the first benchmark for CAD code generation, consisting of
200 natural language prompts paired with expert-annotated scripting code for 3D
objects to benchmark progress. Our findings show that CADCodeVerify improves
VLM performance by providing visual feedback, enhancing the structure of the 3D
objects, and increasing the success rate of the compiled program. When applied t
o
GPT-4, CADCodeVerify achieved a 7.30% reduction in Point Cloud distance and a
5.0% improvement in success rate compared to prior work.
****
Filippos Christianos,Georgios Papoudakis,Thomas Coste,Jianye HAO,Jun Wang,Kun Sh
ao
Lightweight Neural App Control
This paper introduces a novel mobile phone control architecture, Lightweight Mul
ti-modal App Control (LiMAC), for efficient interactions and control across vari
ous Android apps. LiMAC takes as input a textual goal and a sequence of past mo
bile observations, such as screenshots and corresponding UI trees, to generate p
recise actions. To address the computational constraints inherent to smartphones
, we introduce a small Action Transformer (AcT) integrated with a fine-tuned vis
ion-language model (VLM) for real-time decision-making and task execution. We e
valuate LiMAC on two open-source mobile control datasets, demonstrating the supe

rior performance of our small-form-factor approach against fine-tuned versions o
f open-source VLMs, such as Florence2 and Qwen2-VL. It also significantly outper
forms prompt engineering baselines utilising closed-source foundation models lik
e GPT-4o. More specifically, LiMAC increases the overall action accuracy by up t
o 19% compared to fine-tuned VLMs, and up to 42% compared to prompt-engineering
baselines.
****

Pankaj Bhagwat,Linglong Kong,Bei Jiang

CBMA: Improving Conformal Prediction through Bayesian Model Averaging

Conformal prediction has emerged as a popular technique for facilitating valid p
redictive inference across a spectrum of machine learning models, under minimal
assumption of exchangeability. Recently, Hoff (2023) showed that full conformal
Bayes provides the most efficient prediction sets (smallest by expected volume)
among all prediction sets that are valid at the $(1 - \alpha)$ level if the mod
el is correctly specified. However, a critical issue arises when the Bayesian mo
del itself may be mis-specified, resulting in prediction interval that might be
suboptimal, even though it still enjoys the frequentist coverage guarantee. To a
ddress this limitation, we propose an innovative solution that combines Bayesian
 model averaging (BMA) with conformal prediction. This hybrid not only leverages
 the strengths of Bayesian conformal prediction but also introduces a layer of r
obustness through model averaging. Theoretically, we prove that the resulting pr
ediction interval will converge to the optimal level of efficiency, if the true
model is included among the candidate models. This assurance of optimality, even
 under potential model uncertainty, provides a significant improvement over exis
ting methods, ensuring more reliable and precise uncertainty quantification.
****

Zafeirios Fountas,Martin Benfeghoul,Adnan Oomerjee,Fenia Christopoulou,Gerasimos
 Lampouras,Haitham Bou Ammar,Jun Wang

Human-inspired Episodic Memory for Infinite Context LLMs

Large language models (LLMs) have shown remarkable capabilities, but still strug
gle with processing extensive contexts, limiting their ability to maintain coher
ence and accuracy over long sequences. In contrast, the human brain excels at or
ganising and retrieving episodic experiences across vast temporal scales, spanni
ng a lifetime. In this work, we introduce EM-LLM, a novel approach that integrat
es key aspects of human episodic memory and event cognition into LLMs with no fi
ne-tuning, enabling them to handle practically infinite context lengths while ma
intaining computational efficiency. EM-LLM organises sequences of tokens into co
herent episodic events using a combination of Bayesian surprise and graph-theore
tic boundary refinement in an online fashion. When needed, these events are retr
ieved through a two-stage memory process, combining similarity-based and tempora
lly contiguous retrieval for efficient, human-inspired access to relevant inform
ation. Experiments on the LongBench and ∞-Bench benchmarks demonstrate EM
-LLM's superior performance, consistently outperforming the state-of-the-art ret
rieval model InfLLM across various baseline LLMs. In addition, EM-LLM outperform
s its popular counterpart, RAG, in a wide range of tasks, while requiring simila
r resources. Notably, EM-LLM's performance even surpasses full-context models in
 most tasks, while successfully performing retrieval across 10 million tokens --
 a scale computationally infeasible for such models. Finally, our analysis revea
ls strong correlations between EM-LLM's event segmentation and human-perceived e
vents, suggesting parallels between this artificial system and its biological co
unterpart, thereby offering a novel computational framework for exploring human
memory mechanisms.
****

Yuxin Wang,Maresa Schröder,Dennis Frauen,Jonas Schweisthal,Konstantin Hess,Stefa
n Feuerriegel

Constructing Confidence Intervals for Average Treatment Effects from Multiple Da
tasets

Constructing confidence intervals (CIs) for the average treatment effect (ATE) f
rom patient records is crucial to assess the effectiveness and safety of drugs.
However, patient records typically come from different hospitals, thus raising t

he question of how multiple observational/experimental datasets can be effective
ly combined for this purpose. In our paper, we propose a new method that estimat
es the ATE from multiple observational/experimental datasets and provides valid
CIs. Our method makes little assumptions about the observational datasets and is
 thus widely applicable in medical practice. The key idea of our method is that
we leverage prediction-powered inferences and thereby essentially `shrink' the C
Is so that we offer more precise uncertainty quantification as compared to na{\"
i}ve approaches. We further prove the unbiasedness of our method and the validit
y of our CIs. We confirm our theoretical results through various numerical exper
iments.
****
Vivek Myers,Catherine Ji,Benjamin Eysenbach
Horizon Generalization in Reinforcement Learning
We study goal-conditioned RL through the lens of generalization, but not in the
traditional sense of random augmentations and domain randomization. Rather, we a
im to learn goal-directed policies that generalize with respect to the horizon:
after training to reach nearby goals (which are easy to learn), these policies s
hould succeed in reaching distant goals (which are quite challenging to learn).
In the same way that invariance is closely linked with generalization is other a
reas of machine learning (e.g., normalization layers make a network invariant to
 scale, and therefore generalize to inputs of varying scales), we show that this
 notion of horizon generalization is closely linked with invariance to planning:
 a policy navigating towards a goal will select the same actions as if it were n
avigating to a waypoint en route to that goal. Horizon generalization and invari
ance to planning are appealing because of their potential reach: they imply that
 a policy trained to reach nearby goals would succeed at reaching goals that are
 arbitrarily more distant.Our theoretical analysis proves that both horizon gene
ralization and planning invariance are possible, under some assumptions. We pres
ent new experimental results, as well as recalling results from prior work, in s
upport of our theoretical results. Taken together, our results open the door to
studying how techniques for invariance and generalization developed in other are
as of machine learning might be adapted to achieve this alluring property.
****
Nikolaos Tsilivis,Gal Vardi,Julia Kempe
Flavors of Margin: Implicit Bias of Steepest Descent in Homogeneous Neural Netwo
rks
We study the implicit bias of the family of steepest descent algorithms with inf
initesimal learning rate, including gradient descent, sign gradient descent and
coordinate descent, in deep homogeneous neural networks. We prove that an algori
thm-dependent geometric margin increases during training and characterize the la
te-stage bias of the algorithms. In particular, we define a generalized notion o
f stationarity for optimization problems and show that the algorithms progressiv
ely reduce a (generalized) Bregman divergence, which quantifies proximity to suc
h stationary points of a margin-maximization problem. We then experimentally zoo
m into the trajectories of neural networks optimized with various steepest desce
nt algorithms, highlighting connections to the implicit bias of Adam.
****
Pengxiang Li,Lu Yin,Shiwei Liu
Mix-LN: Unleashing the Power of Deeper Layers by Combining Pre-LN and Post-LN
Large Language Models (LLMs) have achieved remarkable success, yet recent findin
gs reveal that their deeper layers often contribute minimally and can be pruned
without affecting overall performance. While some view this as an opportunity f
or model compression, we identify it as a training shortfall rooted in the wides
pread use of Pre-Layer Normalization (Pre-LN). We demonstrate that Pre-LN, commo
nly employed in models like GPT and LLaMA, leads to diminished gradient norms in
 its deeper layers, reducing their effectiveness. In contrast, Post-Layer Normal
ization (Post-LN) preserves larger gradient norms in deeper layers but suffers f
rom vanishing gradients in earlier layers. To address this, we introduce Mix-LN,
 a novel normalization technique that combines the strengths of Pre-LN and Post-
LN within the same model. Mix-LN applies Post-LN to the earlier layers and Pre-L

N to the deeper layers, ensuring more uniform gradient norms across layers. This allows all parts of the network—both shallow and deep layers—to contribute effectively to training. Extensive experiments with various model sizes demonstrate that Mix-LN consistently outperforms both Pre-LN and Post-LN, promoting more balanced, healthier gradient norms throughout the network, and enhancing the overall quality of LLM pre-training. Furthermore, we demonstrate that models pre-trained with Mix-LN learn better compared to those using Pre-LN or Post-LN during supervised fine-tuning, highlighting the critical importance of high-quality deep layers. By effectively addressing the inefficiencies of deep layers in current LLMs, Mix-LN unlocks their potential, enhancing model capacity without increasing model size. Our code is available at https://github.com/pixeli99/MixLN.
****
Xingyi Yang,Xinchao Wang
Kolmogorov-Arnold Transformer
Transformers stand as the cornerstone of mordern deep learning. Traditionally, these models rely on multi-layer
perceptron (MLP) layers to mix the information between channels. In this paper, we introduce the Kolmogorov–Arnold
Transformer (KAT), a novel architecture that replaces MLP layers with Kolmogorov-Arnold Network (KAN) layers to
enhance the expressiveness and performance of the model. Integrating KANs into transformers, however, is no easy
feat, especially when scaled up. Specifically, we identify three key challenges: (C1) Base function. The standard B-spline
function used in KANs is not optimized for parallel computing on modern hardware, resulting in slower inference speeds.
(C2) Parameter and Computation Inefficiency. KAN requires a unique function for each input-output pair, making the
computation extremely large. (C3) Weight initialization. The initialization of weights in KANs is particularly challenging
due to their learnable activation functions, which are critical for achieving convergence in deep neural networks. To
overcome the aforementioned challenges, we propose three key solutions: (S1) Rational basis. We replace B-spline functions
with rational functions to improve compatibility with modern GPUs. By implementing this in CUDA, we achieve faster
computations. (S2) Group KAN. We share the activation weights through a group of neurons, to reduce the computational
load without sacrificing performance. (S3) Variance-preserving initialization. We carefully initialize the activation weights
to make sure that the activation variance is maintained across layers. With these designs, KAT scales effectively and readily
outperforms traditional MLP-based transformers. We demonstrate the advantages of KAT across various tasks, including
image recognition, object detection, and semantic segmentation. It consistently enhances performance over the standard
transformer architectures of different model sizes.
****
Qiujie Xie,Qiming Feng,Tianqi Zhang,Qingqiu Li,Linyi Yang,Yuejie Zhang,Rui Feng,Liang He,Shang Gao,Yue Zhang
Human Simulacra: Benchmarking the Personification of Large Language Models
Large Language Models (LLMs) are recognized as systems that closely mimic aspects of human intelligence. This capability has attracted the attention of the social science community, who see the potential in leveraging LLMs to replace human participants in experiments, thereby reducing research costs and complexity. In this paper, we introduce a benchmark for LLMs personification, including a strategy for constructing virtual characters' life stories from the ground up, a Multi-Agent Cognitive Mechanism capable of simulating human cognitive processes, and a psychology-guided evaluation method to assess human simulations from both self and observational perspectives. Experimental results demonstrate that our cons

tructed simulacra can produce personified responses that align with their target characters. We hope this work will serve as a benchmark in the field of human simulation, paving the way for future research.
****
Niklas Muennighoff,Hongjin SU,Liang Wang,Nan Yang,Furu Wei,Tao Yu,Amanpreet Singh,Douwe Kiela
Generative Representational Instruction Tuning
All text-based language problems can be reduced to either generation or embedding. Current models only perform well at one or the other. We introduce generative representational instruction tuning (GRIT) whereby a large language model is trained to handle both generative and embedding tasks by distinguishing between them through instructions. Compared to other open models, our resulting GritLM-7B is among the top models on the Massive Text Embedding Benchmark (MTEB) and outperforms various models up to its size on a range of generative tasks. By scaling up further, GritLM-8x7B achieves even stronger generative performance while still being among the best embedding models. Notably, we find that GRIT matches training on only generative or embedding data, thus we can unify both at no performance loss. Among other benefits, the unification via GRIT speeds up Retrieval-Augmented Generation (RAG) by > 60% for long documents, by no longer requiring separate retrieval and generation models. Models, code, etc. are freely available at https://github.com/ContextualAI/gritlm.
****
Yang Qin,Chao Chen,Zhihang Fu,Ze Chen,Dezhong Peng,Peng Hu,Jieping Ye
ROUTE: Robust Multitask Tuning and Collaboration for Text-to-SQL
Despite the significant advancements in Text-to-SQL (Text2SQL) facilitated by large language models (LLMs), the latest state-of-the-art techniques are still trapped in the in-context learning of closed-source LLMs (e.g., GPT-4), which limits their applicability in open scenarios.
To address this challenge, we propose a novel RObust mUltitask Tuning and collaboration mEthod (ROUTE) to improve the comprehensive capabilities of open-source LLMs for Text2SQL, thereby providing a more practical solution. Our approach begins with multi-task supervised fine-tuning (SFT) using various synthetic training data related to SQL generation. Unlike existing SFT-based Text2SQL methods, we introduced several additional SFT tasks, including schema linking, noise correction, and continuation writing. Engaging in a variety of SQL generation tasks enhances the model's understanding of SQL syntax and improves its ability to generate high-quality SQL queries. Additionally, inspired by the collaborative modes of LLM agents, we introduce a Multitask Collaboration Prompting (MCP) strategy. This strategy leverages collaboration across several SQL-related tasks to reduce hallucinations during SQL generation, thereby maximizing the potential of enhancing Text2SQL performance through explicit multitask capabilities. Extensive experiments and in-depth analyses have been performed on eight open-source LLMs and five widely-used benchmarks. The results demonstrate that our proposal outperforms the latest Text2SQL methods and yields leading performance.
****
Nguyen T Thach,Patrick Habecker,Anika R. Eisenbraun,W. Alex Mason,Kimberly A. Tyler,Bilal Khan,Hau Chan
MuHBoost: Multi-Label Boosting For Practical Longitudinal Human Behavior Modeling
Longitudinal human behavior modeling has received increasing attention over the years due to its widespread applications to patient monitoring, dietary and lifestyle recommendations, and just-in-time intervention for at-risk individuals (e.g., problematic drug users and struggling students), to name a few. Using in-the-moment health data collected via ubiquitous devices (e.g., smartphones and smartwatches), this multidisciplinary field focuses on developing predictive models for certain health or well-being outcomes (e.g., depression and stress) in the short future given the time series of individual behaviors (e.g., resting heart rate, sleep quality, and current feelings). Yet, most existing models on these data, which we refer to as ubiquitous health data, do not achieve adequate accuracy. The latest works that yielded promising results have yet to consider realisti

c aspects of ubiquitous health data (e.g., containing features of different type
s and high rate of missing values) and the consumption of various resources (e.g
., computing power, time, and cost). Given these two shortcomings, it is dubious
 whether these studies could translate to realistic settings. In this paper, we
propose MuHBoost, a multi-label boosting method for addressing these shortcoming
s, by leveraging advanced methods in large language model (LLM) prompting and mu
lti-label classification (MLC) to jointly predict multiple health or well-being
outcomes. Because LLMs can hallucinate when tasked with answering multiple quest
ions simultaneously, we also develop two variants of MuHBoost that alleviate thi
s issue and thereby enhance its predictive performance. We conduct extensive exp
eriments to evaluate MuHBoost and its variants on 13 health and well-being predi
ction tasks defined from four realistic ubiquitous health datasets. Our results
show that our three developed methods outperform all considered baselines across
 three standard MLC metrics, demonstrating their effectiveness while ensuring re
source efficiency.
****
Yikun Bai,Abihith Kothapalli,Hengrong Du,Rocio Diaz Martin,Soheil Kolouri
Linear Partial Gromov-Wasserstein Embedding
The Gromov–Wasserstein (GW) problem, a variant of the classical optimal transpor
t (OT) problem, has attracted growing interest in the machine learning and data
science communities due to its ability to quantify similarity between measures i
n different metric spaces. However, like the classical OT problem, GW imposes an
 equal mass constraint between measures, which restricts its application in many
 machine learning tasks. To address this limitation, the partial Gromov-Wasserst
ein (PGW) problem has been introduced.
 It relaxes the equal mass constraint, allowing the comparison of general positi
ve Radon measures. Despite this, both GW and PGW face significant computational
challenges due to their non-convex nature. To overcome these challenges, we prop
ose the linear partial Gromov-Wasserstein (LPGW) embedding, a linearized embeddi
ng technique for the PGW problem. For K different metric measure spaces, the p
airwise computation of the PGW distance requires solving the PGW problem $\mathc
al{O}(K^2)$ times.
In contrast, the proposed linearization technique reduces this to $\mathcal{O}(K
)$ times. Similar to the linearization technique for the classical OT problem, w
e prove that LPGW defines a valid metric for metric measure spaces. Finally, we
demonstrate the effectiveness of LPGW in practical applications such as shape re
trieval and learning with transport-based embeddings, showing that LPGW preserve
s the advantages of PGW in partial matching while significantly enhancing comput
ational efficiency. The code is available at https://github.com/mint-vu/Lineariz
ed_Partial_Gromov_Wasserstein.
****
Kyeongkook Seo,Dong-Jun Han,Jaejun Yoo
PRISM: Privacy-Preserving Improved Stochastic Masking for Federated Generative M
odels
Despite recent advancements in federated learning (FL), the integration of gener
ative models into FL has been limited due to challenges such as high communicati
on costs and unstable training in heterogeneous data environments. To address th
ese issues, we propose PRISM, a FL framework tailored for generative models that
 ensures (i) stable performance in heterogeneous data distributions and (ii) res
ource efficiency in terms of communication cost and final model size. The key of
 our method is to search for an optimal stochastic binary mask for a random netw
ork rather than updating the model weights, identifying a sparse subnetwork with
 high generative performance; i.e., a ``strong lottery ticket''. By communicatin
g binary masks in a stochastic manner, PRISM minimizes communication overhead. T
his approach, combined with the utilization of maximum mean discrepancy (MMD) lo
ss and a mask-aware dynamic moving average aggregation method (MADA) on the serv
er side, facilitates stable and strong generative capabilities by mitigating loc
al divergence in FL scenarios. Moreover, thanks to its sparsifying characteristi
c, PRISM yields a lightweight model without extra pruning or quantization, makin
g it ideal for environments such as edge devices. Experiments on MNIST, FMNIST,

CelebA, and CIFAR10 demonstrate that PRISM outperforms existing methods, while m
aintaining privacy with minimal communication costs. PRISM is the first to succe
ssfully generate images under challenging non-IID and privacy-preserving FL envi
ronments on complex datasets, where previous methods have struggled.
****

Lei Shen,Zhenheng Tang,Lijun Wu,Yonggang Zhang,Xiaowen Chu,Tao Qin,Bo Han
Hot-pluggable Federated Learning: Bridging General and Personalized FL via Dynam
ic Selection
Personalized federated learning (PFL) achieves high performance by assuming clie
nts only meet test data locally, which does not meet many generic federated lear
ning (GFL) scenarios. In this work, we theoretically show that PMs can be used t
o enhance GFL with a new learning problem named Selective FL (SFL), which involv
es optimizing PFL and model selection. However, storing and selecting whole mode
ls requires impractical computation and communication costs. To practically solv
e SFL, inspired by model components that attempt to edit a sub-model for specifi
c purposes, we design an efficient and effective framework named Hot-Pluggable F
ederated Learning (HPFL). Specifically, clients individually train personalized
plug-in modules based on a shared backbone, and upload them with a plug-in marke
r on the server modular store. In inference stage, an accurate selection algorit
hm allows clients to identify and retrieve suitable plug-in modules from the mod
ular store to enhance their generalization performance on the target data distri
bution. Furthermore, we provide differential privacy protection during the selec
tion with theoretical guarantee. Our comprehensive experiments and ablation stud
ies demonstrate that HPFL significantly outperforms state-of-the-art GFL and PFL
 algorithms. Additionally, we empirically show HPFL's remarkable potential to re
solve other practical FL problems such as continual federated learning and discu
ss its possible applications in one-shot FL, anarchic FL, and FL plug-in market.
 Our work is the first attempt towards improving GFL performance through a selec
ting mechanism with personalized plug-ins.
****

Juan Luis Gastaldi,John Terilla,Luca Malagutti,Brian DuSell,Tim Vieira,Ryan Cott
erell
The Foundations of Tokenization: Statistical and Computational Concerns
Tokenization — the practice of converting strings of characters from an alphabet
 into sequences of tokens over a vocabulary — is a critical step in the NLP pipe
line. The use of token representations is widely credited with increased model p
erformance but is also the source of many undesirable behaviors, such as spuriou
s ambiguity or inconsistency. Despite its recognized importance as a standard re
presentation method in NLP, the theoretical underpinnings of tokenization are no
t yet fully understood. In particular, the impact of tokenization on language mo
del estimation has been investigated primarily through empirical means. The pres
ent paper contributes to addressing this theoretical gap by proposing a unified
formal framework for representing and analyzing tokenizer models. Based on the c
ategory of stochastic maps, this framework enables us to establish general condi
tions for a principled use of tokenizers and, most importantly, the necessary an
d sufficient conditions for a tokenizer model to preserve the consistency of sta
tistical estimators. In addition, we discuss statistical and computational conce
rns crucial for designing and implementing tokenizer models, such as inconsisten
cy, ambiguity, finiteness, and sequentiality. The framework and results advanced
 in this paper contribute to building robust theoretical foundations for represe
ntations in neural language modeling that can inform future theoretical and empi
rical research.
****

Shuo Ma,Yingwei Zhang,Yiqiang Chen,Hualei Wang,Yuan Jin,Wei Zhang,Ziyu Jia
SleepSMC: Ubiquitous Sleep Staging via Supervised Multimodal Coordination
Sleep staging is critical for assessing sleep quality and tracking health. Polys
omnography (PSG) provides comprehensive multimodal sleep-related information, bu
t its complexity and impracticality limit its practical use in daily and ubiquit
ous monitoring. Conversely, unimodal devices offer more convenience but less acc
uracy. Existing multimodal learning paradigms typically assume that the data typ

es remain consistent between the training and testing phases. This makes it chal
lenging to leverage information from other modalities in ubiquitous scenarios (e
.g., at home) where only one modality is available. To address this issue, we in
troduce a novel framework for ubiquitous Sleep staging via Supervised Multimodal
 Coordination, called SleepSMC. To capture category-related consistency and comp
lementarity across modality-level instances, we propose supervised modality-leve
l instance contrastive coordination. Specifically, modality-level instances with
in the same category are considered positive pairs, while those from different c
ategories are considered negative pairs. To explore the varying reliability of a
uxiliary modalities, we calculate uncertainty estimates based on the variance in
 confidence scores for correct predictions during multiple rounds of random mask
s. These uncertainty estimates are employed to assign adaptive weights to multip
le auxiliary modalities during contrastive learning, ensuring that the primary m
odality learns from high-quality, category-related features. Experimental result
s on four public datasets, ISRUC-S3, MASS-SS3, Sleep-EDF-78, and ISRUC-S1, show
that SleepSMC achieves state-of-the-art cross-subject performance. SleepSMC sign
ificantly improves performance when only one modality is present during testing,
 making it suitable for ubiquitous sleep monitoring.
****
Jianqiao Lu,Yingjia Wan,Yinya Huang,Jing Xiong,Zhengying Liu,Zhijiang Guo
FormalAlign: Automated Alignment Evaluation for Autoformalization
Autoformalization aims to convert informal mathematical proofs into machine-veri
fiable formats, bridging the gap between natural and formal languages. However,
ensuring semantic alignment between the informal and formalized statements remai
ns challenging. Existing approaches heavily rely on manual verification, hinderi
ng scalability. To address this, we introduce FormalAlign, a framework for autom
atically evaluating the alignment between natural and formal languages in autofo
rmalization. FormalAlign trains on both the autoformalization sequence generatio
n task and the representational alignment between input and output, employing a
dual loss that combines a pair of mutually enhancing autoformalization and align
ment tasks. Evaluated across four benchmarks augmented by our proposed misalignm
ent strategies, FormalAlign demonstrates superior performance. In our experiment
s, FormalAlign outperforms GPT-4, achieving an Alignment-Selection Score 11.58\%
 higher on \forml-Basic (99.21\% vs. 88.91\%) and 3.19\% higher on MiniF2F-Valid
 (66.39\% vs. 64.34\%). This effective alignment evaluation significantly reduce
s the need for manual verification.
****
Xiaoyu Xiong,Changyu Hu,Chunru Lin,Pingchuan Ma,Chuang Gan,Tao Du
TopoGaussian: Inferring Internal Topology Structures from Visual Clues
We present TopoGaussian, a holistic, particle-based pipeline for inferring the i
nterior structure of an opaque object from easily accessible photos and videos a
s input. Traditional mesh-based approaches require tedious and error-prone mesh
filling and fixing process, while typically output rough boundary surface. Our p
ipeline combines Gaussian Splatting with a novel, versatile particle-based diffe
rentiable simulator that simultaneously accommodates constitutive model, actuato
r, and collision, without interference with mesh. Based on the gradients from th
is simulator, we provide flexible choice of topology representation for optimiza
tion, including particle, neural implicit surface, and quadratic surface. The re
sultant pipeline takes easily accessible photos and videos as input and outputs
the topology that matches the physical characteristics of the input. We demonstr
ate the efficacy of our pipeline on a synthetic dataset and four real-world task
s with 3D-printed prototypes. Compared with existing mesh-based method, our pipe
line is 5.26x faster on average with improved shape quality. These results highl
ight the potential of our pipeline in 3D vision, soft robotics, and manufacturin
g applications.
****
Bo Yue,Shufan Wang,Ashish Gaurav,Jian Li,Pascal Poupart,Guiliang Liu
Understanding Constraint Inference in Safety-Critical Inverse Reinforcement Lear
ning
In practical applications, the underlying constraint knowledge is often unknown

and difficult to specify. To address this issue, recent advances in Inverse Constrained Reinforcement Learning (ICRL) have focused on inferring these constraints from expert demonstrations. However, the ICRL approach typically characterizes constraint learning as a tri-level optimization problem, which is inherently complex due to its interdependent variables and multiple layers of optimization.
Considering these challenges, a critical question arises: *Can we implicitly embed constraint signals into reward functions and effectively solve this problem using a classic reward inference algorithm?* The resulting method, known as Inverse Reward Correction (IRC), merits investigation. In this work, we conduct a theoretical analysis comparing the sample complexities of both solvers. Our findings confirm that the IRC solver achieves lower sample complexity than its ICRL counterpart.
Nevertheless, this reduction in complexity comes at the expense of generalizability. Specifically, in the target environment, the reward correction terms may fail to guarantee the safety of the resulting policy, whereas this issue can be effectively mitigated by transferring the constraints via the ICRL solver.
Advancing our inquiry, we investigate conditions under which the ICRL solver ensures ϵ-optimality when transferring to new environments. Empirical results across various environments validate our theoretical findings, underscoring the nuanced trade-offs between complexity reduction and generalizability in safety-critical applications.
****
Chongkai Gao,Haozhuo Zhang,Zhixuan Xu,Cai Zhehao,Lin Shao
FLIP: Flow-Centric Generative Planning as General-Purpose Manipulation World Model
We aim to develop a model-based planning framework for world models that can be scaled with increasing model and data budgets for general-purpose manipulation tasks with only language and vision inputs. To this end, we present FLow-CentrIc generative Planning (FLIP), a model-based planning algorithm on visual space that features three key modules: 1) a multi-modal flow generation model as the general-purpose action proposal module; 2) a flow-conditioned video generation model as the dynamics module; and 3) a vision-language representation learning model as the value module. Given an initial image and language instruction as the goal, FLIP can progressively search for long-horizon flow and video plans that maximize the discounted return to accomplish the task. FLIP is able to synthesize long-horizon plans across objects, robots, and tasks with image flows as the general action representation, and the dense flow information also provides rich guidance for long-horizon video generation. In addition, the synthesized flow and video plans can guide the training of low-level control policies for robot execution. Experiments on diverse benchmarks demonstrate that FLIP can improve both the success rates and quality of long-horizon video plan synthesis and has the interactive world model property, opening up wider applications for future works. Video demos are on our website: https://nus-lins-lab.github.io/flipweb/.
****
Rafael Valle,Rohan Badlani,Zhifeng Kong,Sang-gil Lee,Arushi Goel,Sungwon Kim,Joao Felipe Santos,Shuqi Dai,Siddharth Gururani,Aya Aljafari,Alexander H. Liu,Kevin J. Shih,Ryan Prenger,Wei Ping,Chao-Han Huck Yang,Bryan Catanzaro
Fugatto 1: Foundational Generative Audio Transformer Opus 1
Fugatto is a versatile audio synthesis and transformation model capable of following free-form text instructions with optional audio inputs. While large language models (LLMs) trained with text on a simple next-token prediction objective can learn to infer instructions directly from the data, models trained solely on audio data lack this capacity. This is because audio data does not inherently contain the instructions that were used to generate it. To overcome this challenge, we introduce a specialized dataset generation approach optimized for producing a wide range of audio generation and transformation tasks, ensuring the data reveals meaningful relationships between audio and language. Another challenge lies in achieving compositional abilities -- such as combining, interpolating between, or negating instructions -- using data alone. To address it, we propose ComposableART, an inference-time technique that extends classifier-free guidance to c

ompositional guidance. It enables the seamless and flexible composition of instr
uctions, leading to highly customizable audio outputs outside the training distr
ibution. Our evaluations across a diverse set of tasks demonstrate that Fugatto
performs competitively with specialized models, while ComposableART enhances its
 sonic palette and control over synthesis. Most notably, we highlight our framew
ork's ability to execute emergent sounds and tasks -- sonic phenomena that trans
cend conventional audio generation -- unlocking new creative possibilities. \hre
f{https://fugatto.github.io/}{Demo Website.}
****
Yu Bo,Weian Mao,Yanjun Shao,Weiqiang Bai,Peng Ye,Xinzhu Ma,Junbo Zhao,Hao Chen,C
hunhua Shen
Revisiting Convolution Architecture in the Realm of DNA Foundation Models
In recent years, A variety of methods based on Transformer and state space model
 (SSM) architectures have been proposed, advancing foundational DNA language mod
els.
However, there is a lack of comparison between these recent approaches and the c
lassical architecture—convolutional networks (CNNs)—on foundation model benchmar
ks.
This raises the question: are CNNs truly being surpassed by these recent approac
hes based on transformer and SSM architectures? In this paper, we develop a simp
le but well-designed CNN-based method, termed ConvNova. ConvNova identifies and
proposes three effective designs: 1) dilated convolutions, 2) gated convolutions
, and 3) a dual-branch framework for gating mechanisms.
Through extensive empirical experiments, we demonstrate that ConvNova significan
tly outperforms recent methods on more than half of the tasks across several fou
ndation model benchmarks. For example, in histone-related tasks, ConvNova exceed
s the second-best method by an average of 5.8\%, while generally utilizing fewer
 parameters and enabling faster computation. In addition, the experiments obser
ved findings that may be related to biological characteristics. This indicates t
hat CNNs are still a strong competitor compared to Transformers and SSMs. We ant
icipate that this work will spark renewed interest in CNN-based methods for DNA
foundation models.
****
Zhen Liu,Tim Z. Xiao,Weiyang Liu,Yoshua Bengio,Dinghuai Zhang
Efficient Diversity-Preserving Diffusion Alignment via Gradient-Informed GFlowNe
ts
While one commonly trains large diffusion models by collecting datasets on targe
t downstream tasks, it is often desired to align and finetune pretrained diffusi
on models with some reward functions that are either designed by experts or lear
ned from small-scale datasets. Existing post-training methods for reward finetun
ing of diffusion models typically suffer from lack of diversity in generated sam
ples, lack of prior preservation, and/or slow convergence in finetuning. Inspire
d by recent successes in generative flow networks (GFlowNets), a class of probab
ilistic models that sample with the unnormalized density of a reward function, w
e propose a novel GFlowNet method dubbed Nabla-GFlowNet (abbreviated as \nabla-G
FlowNet), the first GFlowNet method that leverages the rich signal in reward gra
dients, together with an objective called \nabla-DB plus its variant residual \n
abla-DB designed for prior-preserving diffusion finetuning. We show that our pro
posed method achieves fast yet diversity- and prior-preserving finetuning of Sta
ble Diffusion, a large-scale text-conditioned image diffusion model, on differen
t realistic reward functions.
****
Ruisi Cai,Saurav Muralidharan,Hongxu Yin,Zhangyang Wang,Jan Kautz,Pavlo Molchano
v
LLaMaFlex: Many-in-one LLMs via Generalized Pruning and Weight Sharing
Large Language Model (LLM) providers typically train a family of models, each of
 a different size targeting a specific deployment scenario. Models in the family
 are all trained from scratch, making the process extremely resource intensive.
Recent work has successfully reduced the cost of training model families through
 a combination of structured pruning and knowledge distillation; here, only the

largest model in the family is trained from scratch, and smaller models are obta
ined via pruning. We observe that while effective, this strategy must still perf
orm pruning and distillation with hundreds of billions of training tokens for ev
ery new model, keeping overall training costs high.
In this work, we introduce a novel nested weight-shared architecture named LLaMa
Flex that can be pruned across both width and depth dimensions in a zero-shot ma
nner to instantly yield a large number of highly accurate compressed models.
LLaMaFlex starts from a pretrained model, and only requires a single continued t
raining phase consisting of ~60B tokens, which trains the elastic network and an
 end-to-end Gumbel Softmax-based router; this router is able to interpolate smoo
thly across model sizes, enabling the "train once, deploy many'' paradigm.
We train LLaMaFlex on Llama 3.1 8B and use it to zero-shot generate a family of
compressed models that achieves accuracy on par with or better than state-of-the
-art pruned, elastic/flexible, and trained-from-scratch models.
****
Claudio Battiloro,Ege Karaismailoglu,Mauricio Tec,George Dasoulas,Michelle Audir
ac,Francesca Dominici
E(n) Equivariant Topological Neural Networks
Graph neural networks excel at modeling pairwise interactions, but they cannot f
lexibly accommodate higher-order interactions and features. Topological deep lea
rning (TDL) has emerged recently as a promising tool for addressing this issue.
TDL enables the principled modeling of arbitrary multi-way, hierarchical higher-
order interactions by operating on combinatorial topological spaces, such as sim
plicial or cell complexes, instead of graphs. However, little is known about how
 to leverage geometric features such as positions and velocities for TDL. This p
aper introduces E(n)-Equivariant Topological Neural Networks (ETNNs), which are
E(n)-equivariant message-passing networks operating on combinatorial complexes,
formal objects unifying graphs, hypergraphs, simplicial, path, and cell complexe
s. ETNNs incorporate geometric node features while respecting rotation, reflecti
on, and translation equivariance. Moreover, being TDL models, ETNNs are natively
 ready for settings with heterogeneous interactions. We provide a theoretical a
nalysis to show the improved expressiveness of ETNNs over architectures for geom
etric graphs. We also show how E(n)-equivariant variants of TDL models can be di
rectly derived from our framework. The broad applicability of ETNNs is demonstra
ted through two tasks of vastly different scales: i) molecular property predicti
on on the QM9 benchmark and ii) land-use regression for hyper-local estimation o
f air pollution with multi-resolution irregular geospatial data. The results in
dicate that ETNNs are an effective tool for learning from diverse types of richl
y structured data, as they match or surpass SotA equivariant TDL models with a s
ignificantly smaller computational burden, thus highlighting the benefits of a p
rincipled geometric inductive bias. Our implementation of ETNNs can be found at
https://github.com/NSAPH-Projects/topological-equivariant-networks.
****
Brandon Zhao,Aviad Levis,Liam Connor,Pratul P. Srinivasan,Katherine Bouman
Revealing the 3D Cosmic Web through Gravitationally Constrained Neural Fields
Weak gravitational lensing is the slight distortion of galaxy shapes caused prim
arily by the gravitational effects of dark matter in the universe. In our work,
we seek to invert the weak lensing signal from 2D telescope images to reconstruc
t a 3D map of the universe’s dark matter field. While inversion typically yeilds
 a 2D projection of the dark matter field, accurate 3D maps of the dark matter d
istribution are essential for localizing structures of interest and testing theo
ries of our universe. However, 3D inversion poses signficant challenges. First,
unlike standard 3D reconstruction that relies on multiple viewpoints, in this ca
se, images are only observed from a single viewpoint. This challenge can be part
ially addressed by observing how galaxy emitters throughout the volume are lense
d. However, this leads to the second challenge: the shapes and exact locations o
f unlensed galaxies are unknown, and can only be estimated with a very large deg
ree of uncertainty. This introduces an overwhelming amount of noise which nearly
 drowns out the lensing signal completely. Previous approaches tackle this by im
posing strong assumptions about the structures in the volume. We instead propos

e a methodology using a gravitationally-constrained neural field to flexibly mod
el the continuous matter distribution. We take an analysis-by-synthesis approach
, optimizing the weights of the neural network through a fully differentiable ph
ysical forward model to reproduce the lensing signal present in image measuremen
ts. We showcase our method on simulations, including realistic simulated measure
ments of dark matter distributions that mimic data from upcoming telescope surve
ys. Our results show that our method can not only outperform previous methods, b
ut importantly is also able to recover potentially surprising dark matter struct
ures.
****

Yuki Takezawa,Sebastian U Stich
Scalable Decentralized Learning with Teleportation
Decentralized SGD can run with low communication costs, but its sparse communica
tion characteristics deteriorate the convergence rate, especially when the numbe
r of nodes is large. In decentralized learning settings, communication is assume
d to occur on only a given topology, while in many practical cases, the topology
 merely represents a preferred communication pattern, and connecting to arbitrar
y nodes is still possible. Previous studies have tried to alleviate the converge
nce rate degradation in these cases by designing topologies with large spectral
gaps. However, the degradation is still significant when the number of nodes is
substantial. In this work, we propose TELEPORTATION. TELEPORTATION activates onl
y a subset of nodes, and the active nodes fetch the parameters from previous act
ive nodes. Then, the active nodes update their parameters by SGD and perform gos
sip averaging on a relatively small topology comprising only the active nodes. W
e show that by activating only a proper number of nodes, TELEPORTATION can compl
etely alleviate the convergence rate degradation. Furthermore, we propose an eff
icient hyperparameter-tuning method to search for the appropriate number of node
s to be activated. Experimentally, we showed that TELEPORTATION can train neural
 networks more stably and achieve higher accuracy than Decentralized SGD.
****

Qi Wu,Yubo Zhao,Yifan Wang,Xinhang Liu,Yu-Wing Tai,Chi-Keung Tang
Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs
While previous approaches to 3D human motion generation have achieved notable su
ccess, they often rely on extensive training and are limited to specific tasks.
To address these challenges, we introduce **Motion-Agent**, an efficient convers
ational framework designed for general human motion generation, editing, and und
erstanding.
Motion-Agent employs an open-source pre-trained language model to develop a gene
rative agent, **MotionLLM**, that bridges the gap between motion and text. This
is accomplished by encoding and quantizing motions into discrete tokens that ali
gn with the language model's vocabulary. With only 1-3% of the model's parameter
s fine-tuned using adapters, MotionLLM delivers performance on par with diffusio
n models and other transformer-based methods trained from scratch. By integratin
g MotionLLM with GPT-4 without additional training, Motion-Agent is able to gene
rate highly complex motion sequences through multi-turn conversations, a capabil
ity that previous models have struggled to achieve.
Motion-Agent supports a wide range of motion-language tasks, offering versatile
capabilities for generating and customizing human motion through interactive con
versational exchanges.
****

Hongxiang Li,Yaowei Li,Yuhang Yang,Junjie Cao,Zhihong Zhu,Xuxin Cheng,Long Chen
DisPose: Disentangling Pose Guidance for Controllable Human Image Animation
Controllable human image animation aims to generate videos from reference images
 using driving videos. Due to the limited control signals provided by sparse gui
dance (e.g., skeleton pose), recent works have attempted to introduce additional
 dense conditions (e.g., depth map) to ensure motion alignment. However, such st
rict dense guidance impairs the quality of the generated video when the body sha
pe of the reference character differs significantly from that of the driving vid
eo. In this paper, we present DisPose to mine more generalizable and effective c
ontrol signals without additional dense input, which disentangles the sparse ske

leton pose in human image animation into motion field guidance and keypoint corr
espondence. Specifically, we generate a dense motion field from a sparse motion
field and the reference image, which provides region-level dense guidance while
maintaining the generalization of the sparse pose control. We also extract diffu
sion features corresponding to pose keypoints from the reference image, and then
these point features are transferred to the target pose to provide distinct ide
ntity information. To seamlessly integrate into existing models, we propose a pl
ug-and-play hybrid ControlNet that improves the quality and consistency of gener
ated videos while freezing the existing model parameters. Extensive qualitative
and quantitative experiments demonstrate the superiority of DisPose compared to
current methods. Project page: https://github.com/lihxxx/DisPose.
****
Hongru Yang,Zhangyang Wang,Jason D. Lee,Yingbin Liang
Transformers Provably Learn Two-Mixture of Linear Classification via Gradient Fl
ow
Understanding how transformers learn and utilize hidden connections between toke
ns is crucial to understand the behavior of large language models.
To understand this mechanism, we consider the task of two-mixture of linear clas
sification which possesses a hidden correspondence structure among tokens, and s
tudy the training dynamics of a symmetric two-headed transformer with ReLU neuro
ns.
Motivated by the stage-wise learning phenomenon in our experiments, we design an
d theoretically analyze a three-stage training algorithm, which can effectively
characterize the actual gradient descent dynamics when we simultaneously train t
he neuron weights and the softmax attention.
The first stage is a neuron learning stage, where the neurons align with the und
erlying signals.
The second stage is a attention feature learning stage, where we analyze the fea
ture learning process of how the attention learns to utilize the relationship be
tween the tokens to solve certain hard samples.
In the meantime, the attention features evolve from a nearly non-separable state
(at the initialization) to a well-separated state.
The third stage is a convergence stage, where the population loss is driven towa
rds zero.
The key technique in our analysis of softmax attention is to identify a critical
sub-system inside a large dynamical system and bound the growth of the non-line
ar sub-system by a linear system.
Finally, we discuss the setting with more than two mixtures.
We empirically show the difficulty of generalizing our analysis of the gradient
flow dynamics to the case even when the number of mixtures equals three, althoug
h the transformer can still successfully learn such distribution.
On the other hand, we show by construction that there exists a transformer that
can solve mixture of linear classification given any arbitrary number of mixture
s.
****
Xiangtao Kong,Kexin Huang,Ping Li,Lei Zhang
Toward Generalizing Visual Brain Decoding to Unseen Subjects
Visual brain decoding aims to decode visual information from human brain activit
ies. Despite the great progress, one critical limitation of current brain decodi
ng research lies in the lack of generalization capability to unseen subjects. Pr
ior work typically focuses on decoding brain activity of individuals based on th
e observation that different subjects exhibit different brain activities, while
it remains unclear whether brain decoding can be generalized to unseen subjects.
This study aims to answer this question. We first consolidate an image-fMRI dat
aset consisting of stimulus-image and fMRI-response pairs, involving 177 subject
s in the movie-viewing task of the Human Connectome Project (HCP). This dataset
allows us to investigate the brain decoding performance with the increase of par
ticipants. We then present a learning paradigm that applies uniform processing a
cross all subjects, instead of employing different network heads or tokenizers f
or individuals as in previous methods, so that we can accommodate a large number

of subjects to explore the generalization capability across different subjects. A series of experiments are conducted and we have the following findings. First, the network exhibits clear generalization capabilities with the increase of training subjects. Second, the generalization capability is common to popular network architectures (MLP, CNN and Transformer). Third, the generalization performance is affected by the similarity between subjects. Our findings reveal the inherent similarities in brain activities across individuals. With the emergence of larger and more comprehensive datasets, it is possible to train a brain decoding foundation model in the future. Codes and models can be found at https://github.com/Xiangtaokong/TGBD}{https://github.com/Xiangtaokong/TGBD.
****

Priyank Pathak,Shyam Marjit,Shruti Vyas,Yogesh S Rawat
LR0.FM: LOW-RESOLUTION ZERO-SHOT CLASSIFICATION BENCHMARK FOR FOUNDATION MODELS
Visual-language foundation Models (FMs) exhibit remarkable zero-shot generalization across diverse tasks, largely attributed to extensive pre-training on large-scale datasets. However, their robustness on low-resolution/pixelated (LR) images, a common challenge in real-world scenarios, remains underexplored. We introduce LR0.FM, a comprehensive benchmark evaluating the impact of low resolution on the zero-shot classification performance of 10 FM(s) across 66 backbones and 15 datasets. We propose a novel metric, Weighted Aggregated Robustness, to address the limitations of existing metrics and better evaluate model performance across resolutions and datasets. Our key findings show that: (i) model size positively correlates with robustness to resolution degradation, (ii) pre-training dataset quality is more important than its size, and (iii) fine-tuned and higher resolution models are less robust against LR. Our analysis further reveals that the model makes semantically reasonable predictions at LR, and the lack of fine-grained details in input adversely impacts the model’s initial layers more than the deeper layers. We use these insights and introduce a simple strategy, LR-TK0, to enhance the robustness of models without compromising their pre-trained weights. We demonstrate the effectiveness of LR-TK0 for robustness against low-resolution across several datasets and its generalization capability across backbones and other approaches. Code is available at : https://github.com/shyammarjit/LR0.FM
****

Sungyoon Lee,Sokbae Lee
Prediction Risk and Estimation Risk of the Ridgeless Least Squares Estimator under General Assumptions on Regression Errors
In recent years, there has been a significant growth in research focusing on minimum ℓ_2 norm (ridgeless) interpolation least squares estimators. However, the majority of these analyses have been limited to an unrealistic regression error structure, assuming independent and identically distributed errors with zero mean and common variance. In this paper, we explore prediction risk as well as estimation risk under more general regression error assumptions, highlighting the benefits of overparameterization in a more realistic setting that allows for clustered or serial dependence. Notably, we establish that the estimation difficulties associated with the variance components of both risks can be summarized through the trace of the variance-covariance matrix of the regression errors. Our findings suggest that the benefits of overparameterization can extend to time series, panel and grouped data.
****

Ulyana Piterbarg,Lerrel Pinto,Rob Fergus
Training Language Models on Synthetic Edit Sequences Improves Code Synthesis
Software engineers mainly write code by editing existing programs. In contrast, language models (LMs) autoregressively synthesize programs in a single pass. One explanation for this is the scarcity of sequential edit data. While high-quality instruction data for code synthesis is scarce, edit data for synthesis is even scarcer. To fill this gap, we develop a synthetic data generation algorithm called LintSeq. This algorithm refactors programs into sequences of synthetic edits by using a linter to procedurally sample across interdependent lines of source code. Synthetic edits sampled with LintSeq reflect the syntax and semantics of their programming language. To test the algorithm, we use it to refactor a datase

t of instruction + program pairs into instruction + program-diff-sequence tuples . Then, we fine-tune a series of smaller LMs ranging from 2.6B to 14B parameters on both the re-factored and original versions of this dataset. We perform compr ehensive evaluations comparing edit sequence code LMs against baselines on Human Eval, MBPP(+), CodeContests, DS-1000, and BigCodeBench. We show that models fine -tuned to iteratively synthesize code match or outperform baselines on pass@1, a nd exhibit better scaling across higher pass@k as a function of total test-time FLOPs. Finally, we also pretrain our own tiny LMs for code understanding. We sho w that fine-tuning these models to synthesize code edit-by-edit results in stron g performance on HumanEval and MBPP(+) compared to existing code language models of similar scale such as CodeT5+, AlphaCode, and Codex.
****

Tianxiang Gao,Siyuan Sun,Hailiang Liu,Hongyang Gao
Global Convergence in Neural ODEs: Impact of Activation Functions
Neural Ordinary Differential Equations (ODEs) have been successful in various ap plications due to their continuous nature and parameter-sharing efficiency. Howe ver, these unique characteristics also introduce challenges in training, particu larly with respect to gradient computation accuracy and convergence analysis. In this paper, we address these challenges by investigating the impact of activati on functions. We demonstrate that the properties of activation functions—specifi cally smoothness and nonlinearity—are critical to the training dynamics. Smooth activation functions guarantee globally unique solutions for both forward and ba ckward ODEs, while sufficient nonlinearity is essential for maintaining the spec tral properties of the Neural Tangent Kernel (NTK) during training. Together, th ese properties enable us to establish the global convergence of Neural ODEs unde r gradient descent in overparameterized regimes. Our theoretical findings are va lidated by numerical experiments, which not only support our analysis but also p rovide practical guidelines for scaling Neural ODEs, potentially leading to fast er training and improved performance in real-world applications.
****

Morgan Bruce Talbot,Gabriel Kreiman,James J. DiCarlo,Guy Gaziv
L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Select ion and Enhancement
The currently leading artificial neural network models of the visual ventral str eam - which are derived from a combination of performance optimization and robus tification methods - have demonstrated a remarkable degree of behavioral alignme nt with humans on visual categorization tasks. We show that image perturbations generated by these models can enhance the ability of humans to accurately report the ground truth class. Furthermore, we find that the same models can also be u sed out-of-the-box to predict the proportion of correct human responses to indiv idual images, providing a simple, human-aligned estimator of the relative diffic ulty of each image. Motivated by these observations, we propose to augment visua l learning in humans in a way that improves human categorization accuracy at tes t time. Our learning augmentation approach consists of (i) selecting images base d on their model-estimated recognition difficulty, and (ii) applying image pertu rbations that aid recognition for novice learners. We find that combining these model-based strategies leads to categorization accuracy gains of 33-72% relative to control subjects without these interventions, on unmodified, randomly select ed held-out test images. Beyond the accuracy gain, the training time for the aug mented learning group was also shortened by 20-23%, despite both groups completi ng the same number of training trials. We demonstrate the efficacy of our approa ch in a fine-grained categorization task with natural images, as well as two tas ks in clinically relevant image domains - histology and dermoscopy - where visua l learning is notoriously challenging. To the best of our knowledge, our work is the first application of artificial neural networks to increase visual learning performance in humans by enhancing category-specific image features.
****

Yu Huang,Zixin Wen,Yuejie Chi,Yingbin Liang
A Theoretical Analysis of Self-Supervised Learning for Vision Transformers
Self-supervised learning has become a cornerstone in computer vision, primarily

divided into reconstruction-based methods like masked autoencoders (MAE) and dis
criminative methods such as contrastive learning (CL). Recent empirical observa
tions reveal that MAE and CL capture different types of representations: CL tend
s to focus on global patterns, while MAE adeptly captures **both global and sub
tle local** information simultaneously. Despite a flurry of recent empirical inv
estigations to shed light on this difference, theoretical understanding remains
limited, especially on the dominant architecture **vision transformers** (ViTs)
. In this paper, to provide rigorous insights, we model the visual data distribu
tion by considering two types of spatial features: dominant global features and
comparatively minuscule local features, and study the impact of imbalance among
these features. We analyze the training dynamics of one-layer softmax-based ViT
s on both MAE and CL objectives using gradient descent. Our analysis shows that
as the degree of feature imbalance varies, ViTs trained with the MAE objective e
ffectively learn both global and local features to achieve near-optimal reconstr
uction, while the CL-trained ViTs favor predominantly global features, even unde
r mild imbalance. These results provide a theoretical explanation for distinct b
ehaviors of MAE and CL observed in empirical studies.
****

Siqi Zeng,Sixian Du,Makoto Yamada,Han Zhao
Learning Structured Representations by Embedding Class Hierarchy with Fast Optim
al Transport
To embed structured knowledge within labels into feature representations, prior
work (Zeng et al., 2022) proposed to use the Cophenetic Correlation Coefficient
(CPCC) as a regularizer during supervised learning. This regularizer calculates
pairwise Euclidean distances of class means and aligns them with the correspondi
ng shortest path distances derived from the label hierarchy tree. However, class
means may not be good representatives of the class conditional distributions, e
specially when they are multi-mode in nature. To address this limitation, under
the CPCC framework, we propose to use the Earth Mover's Distance (EMD) to measur
e the pairwise distances among classes in the feature space. We show that our ex
act EMD method generalizes previous work, and recovers the existing algorithm wh
en class-conditional distributions are Gaussian in the feature space. To further
improve the computational efficiency of our method, we introduce the Optimal Tr
ansport-CPCC family by exploring four EMD approximation variants. Our most effic
ient OT-CPCC variant runs in linear time in the size of the dataset, while maint
aining competitive performance across datasets and tasks. The code is available
at https://github.com/uiuctml/OTCPCC.
****

Kaiyue Wen,Huaqing Zhang,Hongzhou Lin,Jingzhao Zhang
From Sparse Dependence to Sparse Attention: Unveiling How Chain-of-Thought Enhan
ces Transformer Sample Efficiency
Chain-of-thought (CoT) significantly enhances the reasoning performance of larg
e language models (LLM). While current theoretical studies often attribute this
improvement to increased expressiveness and computational capacity, we argue tha
t expressiveness is not the primary limitation in the LLM regime, as current lar
ge models will fail on simple tasks. Using a parity-learning setup, we demonstra
te that CoT can substantially improve sample efficiency even when the representa
tion power is sufficient. Specifically, with CoT, a transformer can learn the fu
nction within polynomial samples, whereas without CoT, the required sample size
is exponential. Additionally, we show that CoT simplifies the learning process b
y introducing sparse sequential dependencies among input tokens, and leads to a
sparse and interpretable attention. We validate our theoretical analysis with bo
th synthetic and real-world experiments, confirming that sparsity in attention l
ayers is a key factor of the improvement induced by CoT.
****

Sirui Li,Wenbin Ouyang,Yining Ma,Cathy Wu
Learning-Guided Rolling Horizon Optimization for Long-Horizon Flexible Job-Shop
Scheduling
Long-horizon combinatorial optimization problems (COPs), such as the Flexible Jo
b-Shop Scheduling Problem (FJSP), often involve complex, interdependent decision

s over extended time frames, posing significant challenges for existing solvers. While Rolling Horizon Optimization (RHO) addresses this by decomposing problems into overlapping shorter-horizon subproblems, such overlap often involves redundant computations. In this paper, we present L-RHO, the first learning-guided RHO framework for COPs. L-RHO employs a neural network to intelligently fix variables that in hindsight did not need to be re-optimized, resulting in smaller and thus easier-to-solve subproblems. For FJSP, this means identifying operations with unchanged machine assignments between consecutive subproblems. Applied to FJSP, L-RHO accelerates RHO by up to 54\% while significantly improving solution quality, outperforming other heuristic and learning-based baselines. We also provide in-depth discussions and verify the desirable adaptability and generalization of L-RHO across numerous FJSP variates, distributions, online scenarios and benchmark instances. Moreover, we provide a theoretical analysis to elucidate the conditions under which learning is beneficial.
****
Xinsong Feng,Zihan Yu,Yanhai Xiong,Haipeng Chen
Sequential Stochastic Combinatorial Optimization Using Hierarchal Reinforcement Learning
Reinforcement learning (RL) has emerged as a promising tool for combinatorial optimization (CO) problems due to its ability to learn fast, effective, and generalizable solutions.
Nonetheless, existing works mostly focus on one-shot deterministic CO, while sequential stochastic CO (SSCO) has rarely been studied despite its broad applications such as adaptive influence maximization (IM) and infectious disease intervention.
In this paper, we study the SSCO problem where we first decide the budget (e.g., number of seed nodes in adaptive IM) allocation for all time steps, and then select a set of nodes for each time step. The few existing studies on SSCO simplify the problems by assuming a uniformly distributed budget allocation over the time horizon, yielding suboptimal solutions. We propose a generic hierarchical RL (HRL) framework called wake-sleep option (WS-option), a two-layer option-based framework that simultaneously decides adaptive budget allocation on the higher layer and node selection on the lower layer.
WS-option starts with a coherent formulation of the two-layer Markov decision processes (MDPs), capturing the interdependencies between the two layers of decisions. Building on this, WS-option employs several innovative designs to balance the model's training stability and computational efficiency, preventing the vicious cyclic interference issue between the two layers. Empirical results show that WS-option exhibits significantly improved effectiveness and generalizability compared to traditional methods. Moreover, the learned model can be generalized to larger graphs, which significantly reduces the overhead of computational resources.
****
KISHALAY DAS,Subhojyoti Khastagir,Pawan Goyal,Seung-Cheol Lee,Satadeep Bhattacharjee,Niloy Ganguly
Periodic Materials Generation using Text-Guided Joint Diffusion Model
Equivariant diffusion models have emerged as the prevailing approach for generating novel crystal materials due to their ability to leverage the physical symmetries
of periodic material structures. However, current models do not effectively learn the
joint distribution of atom types, fractional coordinates, and lattice structure of the
crystal material in a cohesive end-to-end diffusion framework. Also, none of these
models work under realistic setups, where users specify the desired characteristics
that the generated structures must match. In this work, we introduce TGDMat, a
novel text-guided diffusion model designed for 3D periodic material generation.

Our approach integrates global structural knowledge through textual descriptions at each denoising step while jointly generating atom coordinates, types, and lattice structure using a periodic-E(3)-equivariant graph neural network (GNN). Extensive experiments using popular datasets on benchmark tasks reveal that TGDMat outperforms existing baseline methods by a good margin. Notably, for the structure prediction task, with just one generated sample, TGDMat outperforms all baseline models, highlighting the importance of text-guided diffusion. Further, in the generation task, TGDMat surpasses all baselines and their text-fusion variants, showcasing the effectiveness of the joint diffusion paradigm. Additionally, incorporating textual knowledge reduces overall training and sampling computational overhead while enhancing generative performance when utilizing real-world textual prompts from experts. Code is available at https://github.com/kdmsit/TGDMat
****
Seyed Iman Mirzadeh,Keivan Alizadeh,Hooman Shahrokhi,Oncel Tuzel,Samy Bengio,Mehrdad Farajtabar
GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models
Recent advancements in Large Language Models (LLMs) have sparked interest in their mathematical reasoning capabilities. While performance on the widely popular GSM8K benchmark has improved, questions remain about whether reported evaluation metrics are reliable, and reasoning abilities of LLMs have advanced. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models.
Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.
****
Jacob Mitchell Springer,Suhas Kotha,Daniel Fried,Graham Neubig,Aditi Raghunathan
Repetition Improves Language Model Embeddings
Bidirectional models are considered essential for strong text embeddings. Recent approaches to adapt autoregressive language models (LMs) into strong text embedding models have largely had the requirement to modify the LM architecture to be bidirectional. We challenge this premise by introducing ``echo embeddings'' which converts autoregressive LMs into high quality text embedding models \emph{without} changing the architecture or requiring fine-tuning. By repeating the input and extracting embeddings from the repeated tokens—which have access to all original tokens—echo embeddings improve over classical LM embeddings by over 5\% in zero-shot settings. Our zero-shot embeddings nearly match those obtained by bidirectionally-converted LMs that undergo additional masked-language modeling training. Echo embeddings are also compatible with supervised fine-tuning, matching or outperforming bidirectionally-converted LMs in an apples-to-apples comparison

, even with an identical compute budget during training and inference. Overall, repetition is a simple and effective strategy to circumvent the need for bidirectional attention in embedding models, paving the way towards a unified architecture for all NLP tasks.
****
Jinghan Li,Yuan Gao,Jinda Lu,Junfeng Fang,Congcong Wen,Hui Lin,Xiang Wang
DiffGAD: A Diffusion-based Unsupervised Graph Anomaly Detector
Graph Anomaly Detection (GAD) is crucial for identifying abnormal entities within networks, garnering significant attention across various fields. Traditional unsupervised methods, which decode encoded latent representations of unlabeled data with a reconstruction focus, often fail to capture critical discriminative content, leading to suboptimal anomaly detection.
To address these challenges, we present a Diffusion-based Graph Anomaly Detector (DiffGAD). At the heart of DiffGAD is a novel latent space learning paradigm, meticulously designed to enhance the model's proficiency by guiding it with discriminative content. This innovative approach leverages diffusion sampling to infuse the latent space with discriminative content and introduces a content-preservation mechanism that retains valuable information across different scales, significantly improving the model’s adeptness at identifying anomalies with limited time and space complexity.
Our comprehensive evaluation of DiffGAD, conducted on six real-world and large-scale datasets with various metrics, demonstrated its exceptional performance. Our code is available at https://github.com/fortunato-all/DiffGAD
****
Wei Hung,Shao-Hua Sun,Ping-Chun Hsieh
Efficient Action-Constrained Reinforcement Learning via Acceptance-Rejection Method and Augmented MDPs
Action-constrained reinforcement learning (ACRL) is a generic framework for learning control policies with zero action constraint violation, which is required by various safety-critical and resource-constrained applications. The existing ACRL methods can typically achieve favorable constraint satisfaction but at the cost of either high computational burden incurred by the quadratic programs (QP) or increased architectural complexity due to the use of sophisticated generative models. In this paper, we propose a generic and computationally efficient framework that can adapt a standard unconstrained RL method to ACRL through two modifications: (i) To enforce the action constraints, we leverage the classic acceptance-rejection method, where we treat the unconstrained policy as the proposal distribution and derive a modified policy with feasible actions. (ii) To improve the acceptance rate of the proposal distribution, we construct an augmented two-objective Markov decision process (MDP), which include additional self-loop state transitions and a penalty signal for the rejected actions. This augmented MDP incentives the learned policy to stay close to the feasible action sets. Through extensive experiments in both robot control and resource allocation domains, we demonstrate that the proposed framework enjoys faster training progress, better constraint satisfaction, and a lower action inference time simultaneously than the state-of-the-art ACRL methods. We have made the source code publicly available to encourage further research in this direction.
****
Ollie Liu,Deqing Fu,Dani Yogatama,Willie Neiswanger
DeLLMa: Decision Making Under Uncertainty with Large Language Models
The potential of large language models (LLMs) as decision support tools is increasingly being explored in fields such as business, engineering, and medicine, which often face challenging tasks of *decision-making under uncertainty*. In this paper, we show that directly prompting LLMs on these types of decision-making problems can yield poor results, especially as the problem complexity increases.
To aid in these tasks, we propose DeLLMa (Decision-making Large Language Model assistant), a framework designed to enhance decision-making accuracy in uncertain environments. DeLLMa involves a multi-step reasoning procedure that integrates recent best practices in scaling *inference-time reasoning*, drawing upon principles from decision theory and utility theory, to provide an accurate and human-a

uditable decision-making process. We validate our procedure on multiple realisti c decision-making environments, demonstrating that DeLLMa can consistently enhan ce the decision-making performance of leading language models, and achieve up to a 40% increase in accuracy over competing methods. Additionally, we show how pe rformance improves when scaling compute at test time, and carry out human evalua tions to benchmark components of DeLLMa.
****
Yubin Wang,Zhikang Zou,Xiaoqing Ye,Xiao Tan,Errui Ding,Cairong Zhao
Uni^{2}Det: Unified and Universal Framework for Prompt-Guided Multi-dataset 3D D etection
We present Uni^{2}Det, a brand new framework for unified and universal multi-dat aset training on 3D detection, enabling robust performance across diverse domain s and generalization to unseen domains. Due to substantial disparities in data d istribution and variations in taxonomy across diverse domains, training such a d etector by simply merging datasets poses a significant challenge. Motivated by t his observation, we introduce multi-stage prompting modules for multi-dataset 3D detection, which leverages prompts based on the characteristics of correspondin g datasets to mitigate existing differences. This elegant design facilitates sea mless plug-and-play integration within various advanced 3D detection frameworks in a unified manner, while also allowing straightforward adaptation for universa l applicability across datasets. Experiments are conducted across multiple datas et consolidation scenarios involving KITTI, Waymo, and nuScenes, demonstrating t hat our Uni^{2}Det outperforms existing methods by a large margin in multi-datas et training. Notably, results on zero-shot cross-dataset transfer validate the g eneralization capability of our proposed method. Our code is available at https: //github.com/ThomasWangY/Uni2Det.
****
Wanquan Feng,Jiawei Liu,Pengqi Tu,Tianhao Qi,Mingzhen Sun,Tianxiang Ma,Songtao Z hao,SiYu Zhou,Qian HE
I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength
Video generation technologies are developing rapidly and have broad potential ap plications. Among these technologies, camera control is crucial for generating p rofessional-quality videos that accurately meet user expectations. However, exis ting camera control methods still suffer from several limitations, including con trol precision and the neglect of the control for subject motion dynamics. In th is work, we propose I2VControl-Camera, a novel camera control method that signif icantly enhances controllability while providing adjustability over the strength of subject motion. To improve control precision, we employ point trajectory in the camera coordinate system instead of only extrinsic matrix information as our control signal. To accurately control and adjust the strength of subject motion , we explicitly model the higher-order components of the video trajectory expans ion, not merely the linear terms, and design an operator that effectively repres ents the motion strength. We use an adapter architecture that is independent of the base model structure. Experiments on static and dynamic scenes show that our framework outperformances previous methods both quantitatively and qualitativel y. Project page: https://wanquanf.github.io/I2VControlCamera.
****
Quyen Tran,Tung Lam Tran,Khanh Doan,Toan Tran,Dinh Phung,Khoat Than,Trung Le
Boosting Multiple Views for pretrained-based Continual Learning
Recent research has shown that Random Projection (RP) can effectively improve th e performance of pre-trained models in Continual learning (CL). The authors hypo thesized that using RP to map features onto a higher-dimensional space can make them more linearly separable. In this work, we theoretically analyze the role of RP and present its benefits for improving the model’s generalization ability in each task and facilitating CL overall. Additionally, we take this result to t he next level by proposing a Multi-View Random Projection scheme for a stronger ensemble classifier. In particular, we train a set of linear experts, among whic h diversity is encouraged based on the principle of AdaBoost, which was initiall y very challenging to apply to CL. Moreover, we employ a task-based adaptive bac kbone

with distinct prompts dedicated to each task for better representation learning. To properly select these task-specific components and mitigate potential feature shifts caused by misprediction, we introduce a simple yet effective technique called the self-improvement process. Experimentally, our method consistently outperforms state-of-the-art baselines across a wide range of datasets.
****
Charilaos Kanatsoulis,Evelyn Choi,Stefanie Jegelka,Jure Leskovec,Alejandro Ribeiro
Learning Efficient Positional Encodings with Graph Neural Networks
Positional encodings (PEs) are essential for effective graph representation learning because they provide position awareness in inherently position-agnostic transformer architectures and increase the expressive capacity of Graph Neural Networks (GNNs). However, designing powerful and efficient PEs for graphs poses significant challenges due to the absence of canonical node ordering and the scale of the graph. In this work, we identify four key properties that graph PEs should satisfy: stability, expressive power, scalability, and genericness. We find that existing eigenvector-based PE methods often fall short of jointly satisfying these criteria. To address this gap, we introduce PEARL, a novel framework of learnable PEs for graphs. Our primary insight is that message-passing GNNs function as nonlinear mappings of eigenvectors, enabling the design of GNN architectures for generating powerful and efficient PEs. A crucial challenge lies in initializing node features in a manner that is both expressive and permutation equivariant. We tackle this by initializing GNNs with random node inputs or standard basis vectors, thereby unlocking the expressive power of message-passing operations, while employing statistical pooling functions to maintain permutation equivariance. Our analysis demonstrates that PEARL approximates equivariant functions of eigenvectors with linear complexity, while rigorously establishing its stability and high expressive power. Experimental evaluations show that PEARL outperforms lightweight versions of eigenvector-based PEs and achieves comparable performance to full eigenvector-based PEs, but with one or two orders of magnitude lower complexity. Our code is available at https://github.com/ehejin/Pearl-PE.
****
Guanyu Zhou,Yibo Yan,Xin Zou,Kun Wang,Aiwei Liu,Xuming Hu
Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality
Multimodal Large Language Models (MLLMs) have emerged as a central focus in both industry and academia, but often suffer from biases introduced by visual and language priors, which can lead to multimodal hallucination. These biases arise from the visual encoder and the Large Language Model (LLM) backbone, affecting the attention mechanism responsible for aligning multimodal inputs. Existing decoding-based mitigation methods focus on statistical correlations and overlook the causal relationships between attention mechanisms and model output, limiting their effectiveness in addressing these biases. To tackle this issue, we propose a causal inference framework termed CausalMM that applies structural causal modeling to MLLMs, treating modality priors as a confounder between attention mechanisms and output. Specifically, by employing backdoor adjustment and counterfactual reasoning at both the visual and language attention levels, our method mitigates the negative effects of modality priors and enhances the alignment of MLLM's inputs and outputs, with a maximum score improvement of 65.3% on 6 VLind-Bench indicators and 164 points on MME Benchmark compared to conventional methods. Extensive experiments validate the effectiveness of our approach while being a plug-and-play solution. Our code is available at: https://github.com/The-Martyr/CausalMM.
****
Yorai Shaoul,Itamar Mishani,Shivam Vats,Jiaoyang Li,Maxim Likhachev
Multi-Robot Motion Planning with Diffusion Models
Diffusion models have recently been successfully applied to a wide range of robotics applications for learning complex multi-modal behaviors from data. However, prior works have mostly been confined to single-robot and small-scale environments due to the high sample complexity of learning multi-robot diffusion models.

In this paper, we propose a method for generating collision-free multi-robot tra
jectories that conform to underlying data distributions while using only single-
robot data. Our algorithm, Multi-robot Multi-model planning Diffusion (MMD), doe
s so by combining learned diffusion models with classical search-based technique
s---generating data-driven motions under collision constraints. Scaling further,
 we show how to compose multiple diffusion models to plan in large environments
where a single diffusion model fails to generalize well. We demonstrate the effe
ctiveness of our approach in planning for dozens of robots in a variety of simul
ated scenarios motivated by logistics environments.
****

Qianggang Ding,Santiago Miret,Bang Liu
MatExpert: Decomposing Materials Discovery By Mimicking Human Experts
Material discovery is a critical research area with profound implications for va
rious industries. In this work, we introduce MatExpert, a novel framework that l
everages Large Language Models (LLMs) and contrastive learning to accelerate the
 discovery and design of new solid-state materials. Inspired by the workflow of
human materials design experts, our approach integrates three key stages: retrie
val, transition, and generation. First, in the retrieval stage, MatExpert identi
fies an existing material that closely matches the desired criteria. Second, in
the transition stage, MatExpert outlines the necessary modifications to transfor
m this material formulation to meet specific requirements outlined by the initia
l user query. Third, in the generation state, MatExpert performs detailed comput
ations and structural generation to create a new material based on the provided
information. Our experimental results demonstrate that MatExpert outperforms sta
te-of-the-art methods in material generation tasks, achieving superior performan
ce across various metrics including validity, distribution, and stability. As su
ch, MatExpert represents a meaningful advancement in computational material disc
overy using langauge-based generative models.
****

Ruizhong Qiu,Zhe Xu,Wenxuan Bao,Hanghang Tong
Ask, and it shall be given: On the Turing completeness of prompting
Since the success of GPT, large language models (LLMs) have revolutionized machi
ne learning and have initiated the so-called *LLM prompting* paradigm. In the er
a of LLMs, people train a single general-purpose LLM and provide the LLM with di
fferent *prompts* to perform different tasks. However, such empirical success la
rgely lacks theoretical understanding. Here, we present the first theoretical st
udy on the LLM prompting paradigm to the best of our knowledge. In this work, we
 show that prompting is in fact Turing-complete: there exists a finite-size Tran
sformer such that for any computable function, there exists a corresponding prom
pt following which the Transformer computes the function. Furthermore, we show t
hat even though we use only a single finite-size Transformer, it can still achie
ve nearly the same complexity bounds as that of the class of all unbounded-size
Transformers. Overall, our result reveals that prompting can enable a single fin
ite-size Transformer to be efficiently universal, which establishes a theoretica
l underpinning for prompt engineering in practice.
****

Vincent Mallet,Yangyang Miao,Souhaib Attaiki,Bruno Correia,Maks Ovsjanikov
AtomSurf: Surface Representation for Learning on Protein Structures
While there has been significant progress in evaluating and comparing different
representations for learning on protein data, the role of surface-based learning
 approaches remains not well-understood. In particular, there is a lack of direc
t and fair benchmark comparison between the best available surface-based learnin
g methods against alternative representations such as graphs. Moreover, the few
existing surface-based approaches either use surface information in isolation or
, at best, perform global pooling between surface and graph-based architectures.

In this work, we fill this gap by first adapting a state-of-the-art surface enco
der for protein learning tasks. We then perform a direct and fair comparison of
the resulting method against alternative approaches within the Atom3D benchmark,

highlighting the limitations of pure surface-based learning. Finally, we propos
e an integrated approach, which allows learned feature sharing between graphs an
d surface representations on the level of nodes and vertices \textit{across all
layers}.

We demonstrate that the resulting architecture achieves state-of-the-art results
on all tasks in the Atom3D benchmark, while adhering to the strict benchmark pr
otocol, as well as more broadly on binding site identification and binding pocke
t classification. Furthermore, we use coarsened surfaces and optimize our approa
ch for efficiency, making our tool competitive in training and inference time wi
th existing techniques.

Code can be found online: https://github.com/Vincentx15/atomsurf
****
Qiang Liu,Mengyu Chu,Nils Thuerey
ConFIG: Towards Conflict-free Training of Physics Informed Neural Networks
The loss functions of many learning problems contain multiple additive terms tha
t can disagree and yield conflicting update directions. For Physics-Informed Neu
ral Networks (PINNs), loss terms on initial/boundary conditions and physics equa
tions are particularly interesting as they are well-established as highly diffic
ult tasks. To improve learning the challenging multi-objective task posed by PIN
Ns, we propose the ConFIG method, which provides conflict-free updates by ensuri
ng a positive dot product between the final update and each loss-specific gradie
nt. It also maintains consistent optimization rates for all loss terms and dynam
ically adjusts gradient magnitudes based on conflict levels. We additionally lev
erage momentum to accelerate optimizations by alternating the back-propagation o
f different loss terms. We provide a mathematical proof showing the convergence
of the ConFIG method, and it is evaluated across a range of challenging PINN sce
narios. ConFIG consistently shows superior performance and runtime compared to b
aseline methods. We also test the proposed method in a classic multi-task benchm
ark, where the ConFIG method likewise exhibits a highly promising performance. S
ource code is available at https://tum-pbs.github.io/ConFIG
****
Saket Tiwari,Omer Gottesman,George Konidaris
Geometry of Neural Reinforcement Learning in Continuous State and Action Spaces
Advances in reinforcement learning (RL) have led to its successful application i
n complex tasks with continuous state and action spaces. Despite these advances
in practice, most theoretical work pertains to finite state and action spaces. W
e propose building a theoretical understanding of continuous state and action sp
aces by employing a geometric lens to understand the locally attained set of sta
tes. The set of all parametrised policies learnt through a semi-gradient based a
pproach induce a set of attainable states in RL. We show that training dynamics
of a two layer neural policy induce a low dimensional manifold of attainable sta
tes embedded in the high-dimensional nominal state space trained using an actor-
critic algorithm. We prove that, under certain conditions, the dimensionality of
this manifold is of the order of the dimensionality of the action space. This i
s the first result of its kind, linking the geometry of the state space to the d
imensionality of the action space. We empirically corroborate this upper bound f
or four MuJoCo environments and also demonstrate the results in a toy environmen
t with varying dimensionality. We also show the applicability of this theoretica
l result by introducing a local manifold learning layer to the policy and value
function networks to improve the performance in control environments with very h
igh degrees of freedom by changing one layer of the neural network to learn spar
se representations.
****
Wenhao Zhan,Scott Fujimoto,Zheqing Zhu,Jason D. Lee,Daniel Jiang,Yonathan Efroni
Exploiting Structure in Offline Multi-Agent RL: The Benefits of Low Interaction
Rank
We study the problem of learning an approximate equilibrium in the offline multi
-agent reinforcement learning (MARL) setting. We introduce a structural assumpti

on---the interaction rank---and establish that functions with low interaction ra
nk are significantly more robust to distribution shift compared to general ones.
 Leveraging this observation, we demonstrate that utilizing function classes wit
h low interaction rank, when combined with regularization and no-regret learning
, admits decentralized, computationally and statistically efficient learning in
cooperative and competitive offline MARL. Our theoretical results are complement
ed by experiments that showcase the potential of critic architectures with low i
nteraction rank in offline MARL, contrasting with commonly used single-agent val
ue decomposition architectures.
****
Hossein Resani,Behrooz Nasihatkon
MIRACLE 3D: Memory-efficient Integrated Robust Approach for Continual Learning o
n 3D Point Clouds via Shape Model Construction
In this paper, we introduce a novel framework for memory-efficient and privacy-p
reserving continual learning in 3D object classification. Unlike conventional me
mory-based approaches in continual learning that require storing numerous exempl
ars, our method constructs a compact shape model for each class, retaining only
the mean shape along with a few key modes of variation. This strategy not only e
nables the generation of diverse training samples while drastically reducing mem
ory usage but also enhances privacy by eliminating the need to store original da
ta. To further improve model robustness against input variations—an issue common
 in 3D domains due to the absence of strong backbones and limited training data—
we incorporate Gradient Mode Regularization. This technique enhances model stabi
lity and broadens classification margins, resulting in accuracy improvements. We
 validate our approach through extensive experiments on the ModelNet40, ShapeNet
, and ScanNet datasets, where we achieve state-of-the-art performance. Notably,
our method consumes only 15% of the memory required by competing methods on the
ModelNet40 and ShapeNet, while achieving comparable performance on the challengi
ng ScanNet dataset with just 8.5% of the memory. These results underscore the sc
alability, effectiveness, and privacy-preserving strengths of our framework for
3D object classification.
****
Feng Liang,Akio Kodaira,Chenfeng Xu,Masayoshi Tomizuka,Kurt Keutzer,Diana Marcul
escu
Looking Backward: Streaming Video-to-Video Translation with Feature Banks
This paper introduces StreamV2V, a diffusion model that achieves real-time strea
ming video-to-video (V2V) translation with user prompts.
  Unlike prior V2V methods using batches to process limited frames, we opt to pr
ocess frames in a streaming fashion, to support unlimited frames.
  At the heart of StreamV2V lies a backward-looking principle that relates the p
resent to the past.
  This is realized by maintaining a feature bank, which archives information fro
m past frames.
  For incoming frames, StreamV2V extends self-attention to include banked keys a
nd values, and directly fuses similar past features into the output.
  The feature bank is continually updated by merging stored and new features, ma
king it compact yet informative.
  StreamV2V stands out for its adaptability and efficiency, seamlessly integrati
ng with image diffusion models without fine-tuning.
  It can run 20 FPS on one A100 GPU, being 15$\times$, 46$\times$, 108$\times$,
and 158$\times$ faster than FlowVid, CoDeF, Rerender, and TokenFlow, respectivel
y.
  Quantitative metrics and user studies confirm StreamV2V's exceptional ability
to maintain temporal consistency.
****
Zhuoming Chen,Ranajoy Sadhukhan,Zihao Ye,Yang Zhou,Jianyu Zhang,Niklas Nolte,Yua
ndong Tian,Matthijs Douze,Leon Bottou,Zhihao Jia,Beidi Chen
MagicPIG: LSH Sampling for Efficient LLM Generation
Large language models (LLMs) with long context windows have gained significant a
ttention. However, the KV cache, stored to avoid re-computation, becomes a bottl

eneck. Various dynamic sparse or TopK-based attention approximation methods have been proposed to leverage the common insight that attention is sparse. In this paper, we first show that TopK attention itself suffers from quality degradation in certain downstream tasks because attention is not always as sparse as expected. Rather than selecting the keys and values with the highest attention scores, sampling with theoretical guarantees can provide a better estimation for attention output. To make the sampling-based approximation practical in LLM generation, we propose MagicPIG, a heterogeneous system based on Locality Sensitive Hashing (LSH). MagicPIG significantly reduces the workload of attention computation while preserving high accuracy for diverse tasks. MagicPIG stores the LSH hash tables and runs the attention computation on the CPU, which allows it to serve longer contexts and larger batch sizes with high approximation accuracy. MagicPIG can improve decoding throughput by up to $5\times$ across various GPU hardware and achieve 54ms decoding latency on a single RTX 4090 for Llama-3.1-8B-Instruct model with a context of 96k tokens.
****
Fabio Ferreira,Ivo Rapant,Jörg K.H. Franke,Frank Hutter
Beyond Random Augmentations: Pretraining with Hard Views
Self-Supervised Learning (SSL) methods typically rely on random image augmentations, or views, to make models invariant to different transformations. We hypothesize that the efficacy of pretraining pipelines based on conventional random view sampling can be enhanced by explicitly selecting views that benefit the learning progress. A simple yet effective approach is to select hard views that yield a higher loss. In this paper, we propose Hard View Pretraining (HVP), a learning-free strategy that extends random view generation by exposing models to more challenging samples during SSL pretraining. HVP encompasses the following iterative steps: 1) randomly sample multiple views and forward each view through the pretrained model, 2) create pairs of two views and compute their loss, 3) adversarially select the pair yielding the highest loss according to the current model state, and 4) perform a backward pass with the selected pair. In contrast to existing hard view literature, we are the first to demonstrate hard view pretraining's effectiveness at scale, particularly training on the full ImageNet-1k dataset, and evaluating across multiple SSL methods, Convolutional Networks, and Vision Transformers. As a result, HVP sets a new state-of-the-art on DINO ViT-B/16, reaching 78.8% linear evaluation accuracy (a 0.6% improvement) and consistent gains of 1% for both 100 and 300 epoch pretraining, with similar improvements across transfer tasks in DINO, SimSiam, iBOT, and SimCLR.
****
Maxon Rubin-Toles,Maya Gambhir,Keshav Ramji,Aaron Roth,Surbhi Goel
Conformal Language Model Reasoning with Coherent Factuality
Language models are increasingly being used in important decision pipelines, so ensuring the correctness of their outputs is crucial. Recent work has proposed evaluating the “factuality” of claims decomposed from a language model generation and applying conformal prediction techniques to filter out those claims that are not factual. This can be effective for tasks such as information retrieval, where constituent claims may be evaluated in isolation for factuality, but is not appropriate for reasoning tasks, as steps of a logical argument can be evaluated for correctness only within the context of the claims that precede them. To capture this, we define “coherent factuality” and develop a conformal-prediction-based method to guarantee coherent factuality for language model outputs. Our approach applies split conformal prediction to subgraphs within a "deducibility" graph that represents the steps of a reasoning problem. We evaluate our method on mathematical reasoning problems from the MATH and FELM datasets and find that our algorithm consistently produces correct and substantiated orderings of claims, achieving coherent factuality across target coverage levels. Moreover, we achieve 90\% factuality on our stricter definition while retaining 80\% or more of the original claims, highlighting the utility of our deducibility-graph-guided approach.
****
Demyan Yarmoshik,Alexander Rogozin,Nikita Kiselev,Daniil Dorin,Alexander Gasniko

v,Dmitry Kovalev
Decentralized Optimization with Coupled Constraints
We consider the decentralized minimization of a separable objective $\sum_{i=1}^{n} f_i(x_i)$, where the variables are coupled through an affine constraint $\sum_{i=1}^n\left(\mathbf{A}_i x_i - b_i\right) = 0$.
We assume that the functions f_i, matrices $\mathbf{A}_i$, and vectors b_i are stored locally by the nodes of a computational network, and that the functions f_i are smooth and strongly convex.

This problem has significant applications in resource allocation and systems control and can also arise in distributed machine learning.
We propose lower complexity bounds for decentralized optimization problems with coupled constraints and a first-order algorithm achieving the lower bounds. To the best of our knowledge, our method is also the first linearly convergent first-order decentralized algorithm for problems with general affine coupled constraints.
****
Yujie Mo,Runpeng Yu,Xiaofeng Zhu,Xinchao Wang
HG-Adapter: Improving Pre-Trained Heterogeneous Graph Neural Networks with Dual Adapters
The "pre-train, prompt-tuning'' paradigm has demonstrated impressive performance for tuning pre-trained heterogeneous graph neural networks (HGNNs) by mitigating the gap between pre-trained models and downstream tasks. However, most prompt-tuning-based works may face at least two limitations: (i) the model may be insufficient to fit the graph structures well as they are generally ignored in the prompt-tuning stage, increasing the training error to decrease the generalization ability; and (ii) the model may suffer from the limited labeled data during the prompt-tuning stage, leading to a large generalization gap between the training error and the test error to further affect the model generalization. To alleviate the above limitations, we first derive the generalization error bound for existing prompt-tuning-based methods, and then propose a unified framework that combines two new adapters with potential labeled data extension to improve the generalization of pre-trained HGNN models. Specifically, we design dual structure-aware adapters to adaptively fit task-related homogeneous and heterogeneous structural information. We further design a label-propagated contrastive loss and two self-supervised losses to optimize dual adapters and incorporate unlabeled nodes as potential labeled data. Theoretical analysis indicates that the proposed method achieves a lower generalization error bound than existing methods, thus obtaining superior generalization ability. Comprehensive experiments demonstrate the effectiveness and generalization of the proposed method on different downstream tasks.
****
Antonios Antoniadis,Marek Elias,Adam Polak,Moritz Venzin
Approximation algorithms for combinatorial optimization with predictions
We initiate a systematic study of utilizing predictions to improve over approximation guarantees of classic algorithms, without increasing the running time. We propose a generic method for a wide class of optimization problems that ask to select a feasible subset of input items of minimal (or maximal) total weight. This gives simple (near-)linear-time algorithms for, e.g., Vertex Cover, Steiner Tree, Minimum Weight Perfect Matching, Knapsack, and Maximum Clique. Our algorithms produce an optimal solution when provided with perfect predictions and their approximation ratio smoothly degrades with increasing prediction error. With small enough prediction error we achieve approximation guarantees that are beyond the reach without predictions in given time bounds, as exemplified by the NP-hardness and APX-hardness of many of the above problems. Although we show our approach to be optimal for this class of problems as a whole, there is a potential for exploiting specific structural properties of individual problems to obtain improved bounds; we demonstrate this on the Steiner Tree problem. We conclude with an empirical evaluation of our approach.
****

Hanning Guo,Farah Abdellatif,Yu Fu,N. Jon Shah,Abigail Morrison,Jürgen Dammers
XAIguiFormer: explainable artificial intelligence guided transformer for brain d
isorder identification
EEG-based connectomes offer a low-cost and portable method to identify brain dis
orders using deep learning. With the growing interest in model interpretability
and transparency, explainable artificial intelligence (XAI) is widely applied to
 understand the decision of deep learning models. However, most research focuses
 solely on interpretability analysis based on the insights from XAI, overlooking
 XAI’s potential to improve model performance. To bridge this gap, we propose a
dynamical-system-inspired architecture, XAI guided transformer (XAIguiFormer), w
here XAI not only provides explanations but also contributes to enhancing the tr
ansformer by refining the originally coarse information in self-attention mechan
ism to capture more relevant dependency relationships. In order not to damage th
e connectome’s topological structure, the connectome tokenizer treats the single
-band graphs as atomic tokens to generate a sequence in the frequency domain. To
 address the limitations of conventional positional encoding in understanding th
e frequency and mitigating the individual differences, we integrate frequency an
d demographic information into tokens via a rotation matrix, resulting in a rich
ly informative representation. Our experiment demonstrates that XAIguiFormer ach
ieves superior performance over all baseline models. In addition, XAIguiFormer p
rovides valuable interpretability through visualization of the frequency band im
portance. Our code is available at https://github.com/HanningGuo/XAIguiFormer.
****
Xinyu Lin,Chaoqun Yang,Wenjie Wang,Yongqi Li,Cunxiao Du,Fuli Feng,See-Kiong Ng,T
at-Seng Chua
Efficient Inference for Large Language Model-based Generative Recommendation
Large Language Model (LLM)-based generative recommendation has achieved notable
success, yet its practical deployment is costly particularly due to excessive in
ference latency caused by autoregressive decoding. For lossless LLM decoding acc
eleration, Speculative Decoding (SD) has emerged as a promising solution. Howeve
r, applying SD to generative recommendation presents unique challenges due to th
e requirement of generating top-K items (i.e., K distinct token sequences) as a
recommendation list by beam search. This leads to more stringent verification in
 SD, where all the top-K sequences from the target LLM must be successfully draf
ted by the draft model at each decoding step. To alleviate this, we consider 1)
boosting top-K sequence alignment between the draft model and the target LLM, an
d 2) relaxing the verification strategy to reduce trivial LLM calls. To this end
, we propose an alignment framework named AtSpeed, which presents the AtSpeed-S
optimization objective for top-K alignment under the strict top-K verification.
Moreover, we introduce a relaxed sampling verification strategy that allows high
-probability non-top-K drafted sequences to be accepted, significantly reducing
LLM calls. Correspondingly, we propose AtSpeed-R for top-K alignment under this
relaxed sampling verification. Empirical results on two real-world datasets demo
nstrate that AtSpeed significantly accelerates LLM-based generative recommendati
on, e.g., near 2x speedup under strict top-K verification and up to 2.5 speedup
under relaxed sampling verification. The codes and datasets are available at~\ur
l{https://anonymous.4open.science/r/AtSpeed/}.
****
Maksym Andriushchenko,Alexandra Souly,Mateusz Dziemian,Derek Duenas,Maxwell Lin,
Justin Wang,Dan Hendrycks,Andy Zou,J Zico Kolter,Matt Fredrikson,Yarin Gal,Xande
r Davies
AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents
The robustness of LLMs to jailbreak attacks, where users design prompts to circu
mvent safety measures and misuse model capabilities, has been studied primarily
for LLMs acting as simple chatbots. Meanwhile, LLM agents---which use external t
ools and can execute multi-stage tasks---may pose a greater risk if misused, but
 their robustness remains underexplored. To facilitate research on LLM agent mis
use, we propose a new benchmark called AgentHarm. The benchmark includes a diver
se set of 110 explicitly malicious agent tasks (440 with augmentations), coverin
g 11 harm categories including fraud, cybercrime, and harassment. In addition to

measuring whether models refuse harmful agentic requests, scoring well on AgentHarm requires jailbroken agents to maintain their capabilities following an attack to complete a multi-step task. We evaluate a range of leading LLMs, and find (1) leading LLMs are surprisingly complaint with malicious agent requests without jailbreaking, (2) simple universal jailbreak strings can be adapted to effectively jailbreak agents, and (3) these jailbreaks enable coherent and malicious multi-step agent behavior and retain model capabilities. To enable simple and reliable evaluation of attacks and defenses for LLM-based agents, we publicly release AgentHarm at https://huggingface.co/datasets/ai-safety-institute/AgentHarm.
****

Yun Zhang,Wei Li,Simiao Li,Hanting Chen,Zhijun Tu,Bingyi Jing,Shaohui Lin,Jie Hu,Wenjia Wang
AugKD: Ingenious Augmentations Empower Knowledge Distillation for Image Super-Resolution
Knowledge distillation (KD) compresses deep neural networks by transferring task-related knowledge from cumbersome pre-trained teacher models to more compact student models. However, vanilla KD for image super-resolution (SR) networks yields only limited improvements due to the inherent nature of SR tasks, where the outputs of teacher models are noisy approximations of high-quality label images. In this work, we show that the potential of vanilla KD has been underestimated and demonstrate that the ingenious application of data augmentation methods can close the gap between it and more complex, well-designed methods. Unlike conventional training processes typically applying image augmentations simultaneously to both low-quality inputs and high-quality labels, we propose AugKD utilizing unpaired data augmentations to 1) generate auxiliary distillation samples and 2) impose label consistency regularization. Comprehensive experiments show that the AugKD significantly outperforms existing state-of-the-art KD methods across a range of SR tasks.
****

Giorgia Minello,Alessandro Bicciato,Luca Rossi,Andrea Torsello,Luca Cosmo
Generating Graphs via Spectral Diffusion
In this paper, we present GGSD, a novel graph generative model based on 1) the spectral decomposition of the graph Laplacian matrix and 2) a diffusion process. Specifically, we propose to use a denoising model to sample eigenvectors and eigenvalues from which we can reconstruct the graph Laplacian and adjacency matrix. Using the Laplacian spectrum allows us to naturally capture the structural characteristics of the graph and work directly in the node space while avoiding the quadratic complexity bottleneck that limits the applicability of other diffusion-based methods. This, in turn, is accomplished by truncating the spectrum, which, as we show in our experiments, results in a faster yet accurate generative process, and by designing a novel transformer-based architecture linear in the number of nodes. Our permutation invariant model can also handle node features by concatenating them to the eigenvectors of each node. An extensive set of experiments on both synthetic and real-world graphs demonstrates the strengths of our model against state-of-the-art alternatives.
****

Jiacheng Ye,Zhenyu Wu,Jiahui Gao,Zhiyong Wu,Xin Jiang,Zhenguo Li,Lingpeng Kong
Implicit Search via Discrete Diffusion: A Study on Chess
In the post-AlphaGo era, there has been a renewed interest in search techniques such as Monte Carlo Tree Search (MCTS), particularly in their application to Large Language Models (LLMs).
This renewed attention is driven by the recognition that current next-token prediction models often lack the ability for long-term planning. Is it possible to instill search-like abilities within the models to enhance their planning abilities without relying on explicit search? We propose DiffuSearch , a model that does \textit{implicit search} by looking into the future world via discrete diffusion modeling. We instantiate DiffuSearch on a classical board game, Chess, where explicit search is known to be essential. Through extensive controlled experiments, we show DiffuSearch outperforms both the searchless and explicit search-enhanced policies. Specifically, DiffuSearch outperforms the one-step policy by 19.2

\% and the MCTS-enhanced policy by 14\% on action accuracy. Furthermore, DiffuSearch demonstrates a notable 30\% enhancement in puzzle-solving abilities compared to explicit search-based policies, along with a significant 540 Elo increase in game-playing strength assessment. These results indicate that implicit search via discrete diffusion is a viable alternative to explicit search over a one-step policy. All codes are publicly available at \href{https://github.com/HKUNLP/DiffuSearch}{https://github.com/HKUNLP/DiffuSearch}.
****
Amrith Setlur,Chirag Nagpal,Adam Fisch,Xinyang Geng,Jacob Eisenstein,Rishabh Agarwal,Alekh Agarwal,Jonathan Berant,Aviral Kumar
Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning
A promising approach for improving reasoning in large language models is to use process reward models (PRMs). PRMs provide feedback at each step of a multi-step reasoning trace, improving credit assignment over outcome reward models (ORMs) that only provide feedback at the final step. However, collecting dense, per-step human labels is not scalable, and training PRMs from automatically-labeled data has thus far led to limited gains. With the goal of using PRMs to improve a *base* policy via test-time search and reinforcement learning (RL), we ask: ``How should we design process rewards?'' Our key insight is that, to be effective, the process reward for a step should measure
progress: a change in the likelihood of producing a correct response in the future, before and after taking the step, as measured under a *prover* policy distinct from the base policy. Such progress values can {distinguish} good and bad steps generated by the base policy, even though the base policy itself cannot. Theoretically, we show that even weaker provers can improve the base policy, as long as they distinguish steps without being too misaligned with the base policy. Our results show that process rewards defined as progress under such provers improve the efficiency of exploration during test-time search and online RL. We empirically validate our claims by training **process advantage verifiers (PAVs)** to measure progress under such provers and show that compared to ORM, they are >8% more accurate, and 1.5-5x more compute-efficient. Equipped with these insights, our PAVs enable **one of the first results** showing a 6x gain in sample efficiency for a policy trained using online RL with PRMs vs. ORMs.
****
Runjie Yan,Yinbo Chen,Xiaolong Wang
Consistent Flow Distillation for Text-to-3D Generation
Score Distillation Sampling (SDS) has made significant strides in distilling image-generative models for 3D generation. However, its maximum-likelihood-seeking behavior often leads to degraded visual quality and diversity, limiting its effectiveness in 3D applications. In this work, we propose Consistent Flow Distillation (CFD), which addresses these limitations. We begin by leveraging the gradient of the diffusion ODE or SDE sampling process to guide the 3D generation. From the gradient-based sampling perspective, we find that the consistency of 2D image flows across different viewpoints is important for high-quality 3D generation. To achieve this, we introduce multi-view consistent Gaussian noise on the 3D object, which can be rendered from various viewpoints to compute the flow gradient. Our experiments demonstrate that CFD, through consistent flows, significantly outperforms previous methods in text-to-3D generation.
****
David Wessels,David M Knigge,Riccardo Valperga,Samuele Papa,Sharvaree Vadgama,Efstratios Gavves,Erik J Bekkers
Grounding Continuous Representations in Geometry: Equivariant Neural Fields
Conditional Neural Fields (CNFs) are increasingly being leveraged as continuous signal representations, by associating each data-sample with a latent variable that conditions a shared backbone Neural Field (NeF) to reconstruct the sample. However, existing CNF architectures face limitations when using this latent downstream in tasks requiring fine-grained geometric reasoning, such as classification and segmentation. We posit that this results from lack of explicit modelling of geometric information (e.g. locality in the signal or the orientation of a feature) in the latent space of CNFs. As such, we propose Equivariant Neural Fields

(ENFs), a novel CNF architecture which uses a geometry-informed cross-attention to condition the NeF on a geometric variable—a latent point cloud of features—that enables an equivariant decoding from latent to field. We show that this approach induces a steerability property by which both field and latent are grounded in geometry and amenable to transformation laws: if the field transforms, the latent representation transforms accordingly—and vice versa. Crucially, this equivariance relation ensures that the latent is capable of (1) representing geometric patterns faitfhully, allowing for geometric reasoning in latent space, (2) weight-sharing over similar local patterns, allowing for efficient learning of datasets of fields. We validate these main properties in a range of tasks including classification, segmentation, forecasting, reconstruction and generative modelling, showing clear improvement over baselines with a geometry-free latent space.
****

Yifan Yang,Hao Ban,Minhui Huang,Shiqian Ma,Kaiyi Ji
Tuning-Free Bilevel Optimization: New Algorithms and Convergence Analysis
Bilevel optimization has recently attracted considerable attention due to its abundant applications in machine learning problems. However, existing methods rely on prior knowledge of problem parameters to determine stepsizes, resulting in significant effort in tuning stepsizes when these parameters are unknown. In this paper, we propose two novel tuning-free algorithms, D-TFBO and S-TFBO. D-TFBO employs a double-loop structure with stepsizes adaptively adjusted by the "inverse of cumulative gradient norms" strategy. S-TFBO features a simpler fully single-loop structure that updates three variables simultaneously with a theory-motivated joint design of adaptive stepsizes for all variables. We provide a comprehensive convergence analysis for both algorithms and show that D-TFBO and S-TFBO respectively require $\mathcal{O}(\frac{1}{\epsilon})$ and $\mathcal{O}(\frac{1}{\epsilon}\log^4(\frac{1}{\epsilon}))$ iterations to find an ϵ-accurate stationary point, (nearly) matching their well-tuned counterparts using the information of problem parameters. Experiments on various problems show that our methods achieve performance comparable to existing well-tuned approaches, while being more robust to the selection of initial stepsizes.
To the best of our knowledge, our methods are the first to completely eliminate the need for stepsize tuning, while achieving theoretical guarantees.
****

Ali Shirali,Ariel D. Procaccia,Rediet Abebe
The Hidden Cost of Waiting for Accurate Predictions
Algorithmic predictions are increasingly informing societal resource allocations by identifying individuals for targeting. Policymakers often build these systems with the assumption that by gathering more observations on individuals, they can improve predictive accuracy and, consequently, allocation efficiency. An overlooked yet consequential aspect of prediction-driven allocations is that of timing. The planner has to trade off relying on earlier and potentially noisier predictions to intervene before individuals experience undesirable outcomes, or they may wait to gather more observations to make more precise allocations. We examine this tension using a simple mathematical model, where the planner collects observations on individuals to improve predictions over time. We analyze both the ranking induced by these predictions and optimal resource allocation. We show that though individual prediction accuracy improves over time, counter-intuitively, the average ranking loss can worsen. As a result, the planner's ability to improve social welfare can decline. We identify inequality as a driving factor behind this phenomenon. Our findings provide a nuanced perspective and challenge the conventional wisdom that it is preferable to wait for more accurate predictions to ensure the most efficient allocations.
****

Zaige Fei,Fan Xu,Junyuan Mao,Yuxuan Liang,Qingsong Wen,Kun Wang,Hao Wu,Yang Wang
Open-CK: A Large Multi-Physics Fields Coupling benchmarks in Combustion Kinetics
In this paper, we use the Fire Dynamics Simulator (FDS) combined with the {\fontfamily{lmtt}\selectfont \textit{supercomputer}} support to create a \textbf{C}ombustion \textbf{K}inetics (CK) dataset for machine learning and scientific research. This dataset captures the development of fires in industrial parks with hig

h-precision Computational Fluid Dynamics (CFD) simulations. It includes various physical fields such as temperature and pressure, and covers multiple environmental combinations for exploring \underline{multi-physics} field coupling phenomena. Additionally, we evaluate several advanced machine learning architectures across our {\fontfamily{lmtt}\selectfont {Open-CK}} benchmark using a substantial computational setup of 64 NVIDIA A100 GPUs: \ding{182} vision backbone; \ding{183} spatio-temporal predictive models; \ding{184} operator learning frameworks. These architectures uniquely excel at handling complex physical field data. We also introduce three benchmarks to demonstrate their potential in enhancing the exploration of downstream tasks: (a) capturing continuous changes in combustion kinetics; (b) a neural partial differential equation solver for learning temperature fields and turbulence; (c) reconstruction of sparse physical observations. The Open-CK dataset and benchmarks aim to advance research in combustion kinetics driven by machine learning, providing a reliable baseline for developing and comparing cutting-edge technologies and models. We hope to further promote the application of deep learning in earth sciences. Our project is available at \url{https://github.com/whscience/Open-CK}.
****

Xin Dong,Yonggan Fu,Shizhe Diao,Wonmin Byeon,ZIJIA CHEN,Ameya Sunil Mahabaleshwarkar,Shih-Yang Liu,Matthijs Van keirsbilck,Min-Hung Chen,Yoshi Suhara,Yingyan Celine Lin,Jan Kautz,Pavlo Molchanov

Hymba: A Hybrid-head Architecture for Small Language Models

We propose Hymba, a family of small language models featuring a hybrid-head parallel architecture that integrates attention mechanisms and state space models (SSMs) within the same layer, offering parallel and complementary processing of the same inputs. In this hybrid-head module, attention heads provide high-resolution recall, while SSM heads facilitate efficient context summarization. Additionally, we introduce learnable meta tokens, which are prepended to prompts to store critical meta information, guiding subsequent tokens and alleviating the “forced-to-attend” burden associated with attention mechanisms. Thanks to the global context summarized by SSMs, the attention heads in our model can be further optimized through cross-layer key-value (KV) sharing and a mix of global and local attention, resulting in a compact cache size without compromising accuracy. Notably, Hymba achieves state-of-the-art performance among small LMs: Our Hymba-1.5B-Base model surpasses all sub-2B public models and even outperforms Llama-3.2-3B, achieving 1.32\% higher average accuracy, an 11.67$\times$ reduction in cache size, and 3.49$\times$ higher throughput.
****

Hyukjun Lim,Sun Kim,Sangseon Lee

CheapNet: Cross-attention on Hierarchical representations for Efficient protein-ligand binding Affinity Prediction

Accurately predicting protein-ligand binding affinity is a critical challenge in drug discovery, crucial for understanding drug efficacy. While existing models typically rely on atom-level interactions, they often fail to capture the complex, higher-order interactions, resulting in noise and computational inefficiency. Transitioning to modeling these interactions at the cluster level is challenging because it is difficult to determine which atoms form meaningful clusters that drive the protein-ligand interactions. To address this, we propose CheapNet, a novel interaction-based model that integrates atom-level representations with hierarchical cluster-level interactions through a cross-attention mechanism. By employing differentiable pooling of atom-level embeddings, CheapNet efficiently captures essential higher-order molecular representations crucial for accurate binding predictions. Extensive evaluations demonstrate that CheapNet not only achieves state-of-the-art performance across multiple binding affinity prediction tasks but also maintains prediction accuracy with reasonable computational efficiency. The code of CheapNet is available at https://github.com/hyukjunlim/CheapNet.
****

Ali Khajegili Mirabadi,Katherine Rich,Hossein Farahani,Ali Bashashati

Boltzmann Semantic Score: A Semantic Metric for Evaluating Large Vision Models Using Large Language Models

Do Large Vision Models (LVMs) extract medically and semantically relevant featur
es similar to those identified by human experts? Currently, only biased, qualita
tive approaches with limited, small-scale expert evaluations are available to an
swer this question. In this study, we propose the Boltzmann Semantic Score (BSS)
, a novel method inspired by state space modeling, to evaluate the encoding spac
e of LVMs from medical images using the encoding space of Large Language Models
(LLMs) from medical reports. Through extensive experimentation on 32 datasets fr
om The Cancer Genome Atlas collection using five state-of-the-art LLMs, we first
 establish a baseline of LLMs' performance in digital pathology and show that LL
Ms' encoding can be linked to patient outcomes. Then, we compared seven LVMs wit
h BSS and showed that LVMs suffer from poor semantic capability when compared wi
th encoded expert knowledge from pathology reports.
We also found statistically significant correlations between BSS (as a measure o
f structural similarity) and performance in two downstream tasks: information re
trieval and survival prediction tasks. Our study also investigates the consensus
 among LLMs in evaluating LVMs using BSS, indicating that LLMs generally reach s
ubstantial consensus in rating LVMs, with some variation dependant on the cancer
 type. We believe the BSS metric proposed here holds significant potential for a
pplication in other domains with similar contexts. Data and code can be found in
 \footnotesize \url{ https://github.com/AIMLab-UBC/Boltzmann}
****
Jingtong Yue,Zhiwei Lin,Xin Lin,Xiaoyu Zhou,Xiangtai Li,Lu Qi,Yongtao Wang,Ming-
Hsuan Yang
RobuRCDet: Enhancing Robustness of Radar-Camera Fusion in Bird's Eye View for 3D
 Object Detection
While recent low-cost radar-camera approaches have shown promising results in
multi-modal 3D object detection, both sensors face challenges from environmen-
tal and intrinsic disturbances. Poor lighting or adverse weather conditions de-
grade camera performance, while radar suffers from noise and positional ambigu-
ity. Achieving robust radar-camera 3D object detection requires consistent perfo
r-
mance across varying conditions, a topic that has not yet been fully explored. I
n
this work, we first conduct a systematic analysis of robustness in radar-camera
de-
tection on five kinds of noises and propose RobuRCDet, a robust object detection
model in bird’s eye view (BEV). Specifically, we design a 3D Gaussian Expan-
sion (3DGE) module to mitigate inaccuracies in radar points, including position,
Radar Cross-Section (RCS), and velocity. The 3DGE uses RCS and velocity priors
to generate a deformable kernel map and variance for kernel size adjustment and
value distribution. Additionally, we introduce a weather-adaptive fusion module,
which adaptively fuses radar and camera features based on camera signal confi-
dence. Extensive experiments on the popular benchmark, nuScenes, show that
our RobuRCDet achieves competitive results in regular and noisy conditions. The
source codes and trained models will be made available.
****
Piersilvio De Bartolomeis,Julia Kostin,Javier Abad,Yixin Wang,Fanny Yang
Doubly robust identification of treatment effects from multiple environments
Practical and ethical constraints often require the use of observational data fo
r causal inference, particularly in medicine and social sciences. Yet, observati
onal datasets are prone to confounding, potentially compromising the validity of
 causal conclusions.
 While it is possible to correct for biases if the underlying causal graph is kn
own, this is rarely a feasible ask in practical scenarios. A common strategy is
to adjust for all available covariates, yet this approach can yield biased treat
ment effect estimates, especially when post-treatment or unobserved variables ar
e present.
We propose RAMEN, an algorithm that produces unbiased treatment effect estimates
by leveraging the heterogeneity of multiple data sources without the need to kno
w or learn the underlying causal graph. Notably, RAMEN achieves *doubly robust i

dentification*: it can identify the treatment effect whenever
the causal parents of the treatment or those of the outcome are observed, and th
e node whose parents are observed satisfies an invariance assumption. Empirical
evaluations across synthetic, semi-synthetic, and real-world datasets show that
our approach significantly outperforms existing methods.
****
Maximilian Muschalik,Fabian Fumagalli,Paolo Frazzetto,Janine Strotherm,Luca Herm
es,Alessandro Sperduti,Eyke Hüllermeier,Barbara Hammer
Exact Computation of Any-Order Shapley Interactions for Graph Neural Networks
Albeit the ubiquitous use of Graph Neural Networks (GNNs) in machine learning (M
L) prediction tasks involving graph-structured data, their interpretability rema
ins challenging. In explainable artificial intelligence (XAI), the Shapley Value
 (SV) is the predominant method to quantify contributions of individual features
 to a ML model’s output. Addressing the limitations of SVs in complex prediction
 models, Shapley Interactions (SIs) extend the SV to groups of features. In this
 work, we explain single graph predictions of GNNs with SIs that quantify node c
ontributions and interactions among multiple nodes. By exploiting the GNN archit
ecture, we show that the structure of interactions in node embeddings are preser
ved for graph prediction. As a result, the exponential complexity of SIs depends
 only on the receptive fields, i.e. the message-passing ranges determined by the
 connectivity of the graph and the number of convolutional layers. Based on our
theoretical results, we introduce GraphSHAP-IQ, an efficient approach to compute
 any-order SIs exactly. GraphSHAP-IQ is applicable to popular message passing te
chniques in conjunction with a linear global pooling and output layer. We showca
se that GraphSHAP-IQ substantially reduces the exponential complexity of computi
ng exact SIs on multiple benchmark datasets. Beyond exact computation, we evalua
te GraphSHAP-IQ’s approximation of SIs on popular GNN architectures and compare
with existing baselines. Lastly, we visualize SIs of real-world water distributi
on networks and molecule structures using a SI-Graph.
****
Andrea Tirinzoni,Ahmed Touati,Jesse Farebrother,Mateusz Guzek,Anssi Kanervisto,Y
ingchen Xu,Alessandro Lazaric,Matteo Pirotta
Zero-Shot Whole-Body Humanoid Control via Behavioral Foundation Models
Unsupervised reinforcement learning (RL) aims at pre-training models that can so
lve a wide range of downstream tasks in complex environments. Despite recent adv
ancements, existing approaches suffer from several limitations: they may require
 running an RL process on each task to achieve a satisfactory performance, they
may need access to datasets with good coverage or well-curated task-specific sam
ples, or they may pre-train policies with unsupervised losses that are poorly co
rrelated with the downstream tasks of interest. In this paper, we introduce FB-C
PR, which regularizes unsupervised zero-shot RL based on the forward-backward (F
B) method towards imitating trajectories from unlabeled behaviors. The resulting
 models learn useful policies imitating the behaviors in the dataset, while ret
aining zero-shot generalization capabilities. We demonstrate the effectiveness o
f FB-CPR in a challenging humanoid control problem. Training FB-CPR online with
observation-only motion capture datasets, we obtain the first humanoid behaviora
l foundation model that can be prompted to solve a variety of whole-body tasks,
including motion tracking, goal reaching, and reward optimization. The resulting
 model is capable of expressing human-like behaviors and it achieves competitive
 performance with task-specific methods while outperforming state-of-the-art uns
upervised RL and model-based baselines.
****
Yuehan Qin,Yichi Zhang,Yi Nian,Xueying Ding,Yue Zhao
MetaOOD: Automatic Selection of OOD Detection Models
How can we automatically select an out-of-distribution (OOD) detection model for
 various underlying tasks? This is crucial for maintaining the reliability of op
en-world applications by identifying data distribution shifts, particularly in c
ritical domains such as online transactions, autonomous driving, and real-time p
atient diagnosis. Despite the availability of numerous OOD detection methods, th
e challenge of selecting an optimal model for diverse tasks remains largely unde

rexplored, especially in scenarios lacking ground truth labels. In this work, we introduce MetaOOD, the first zero-shot, unsupervised framework that utilizes meta-learning to select an OOD detection model automatically. As a meta-learning approach, MetaOOD leverages historical performance data of existing methods across various benchmark OOD detection datasets, enabling the effective selection of a suitable model for new datasets without the need for labeled data at the test time. To quantify task similarities more accurately, we introduce language model-based embeddings that capture the distinctive OOD characteristics of both datasets and detection models. Through extensive experimentation with 24 unique test dataset pairs to choose from among 11 OOD detection models, we demonstrate that MetaOOD significantly outperforms existing methods and only brings marginal time overhead. Our results, validated by Wilcoxon statistical tests, show that MetaOOD surpasses a diverse group of 11 baselines, including established OOD detectors and advanced unsupervised selection methods.
****
Xiangxin Zhou,Yi Xiao,Haowei Lin,Xinheng He,Jiaqi Guan,Yang Wang,Qiang Liu,Feng Zhou,Liang Wang,Jianzhu Ma
Integrating Protein Dynamics into Structure-Based Drug Design via Full-Atom Stochastic Flows
The dynamic nature of proteins, influenced by ligand interactions, is essential for comprehending protein function and progressing drug discovery. Traditional structure-based drug design (SBDD) approaches typically target binding sites with rigid structures, limiting their practical application in drug development. While molecular dynamics simulation can theoretically capture all the biologically relevant conformations, the transition rate is dictated by the intrinsic energy barrier between them, making the sampling process computationally expensive. To overcome the aforementioned challenges, we propose to use generative modeling for SBDD considering conformational changes of protein pockets. We curate a dataset of apo and multiple holo states of protein-ligand complexes, simulated by molecular dynamics, and propose a full-atom flow model (and a stochastic version), named DynamicFlow, that learns to transform apo pockets and noisy ligands into holo pockets and corresponding 3D ligand molecules. Our method uncovers promising ligand molecules and corresponding holo conformations of pockets. Additionally, the resultant holo-like states provide superior inputs for traditional SBDD approaches, playing a significant role in practical drug discovery.
****
Mathias Jackermeier,Alessandro Abate
DeepLTL: Learning to Efficiently Satisfy Complex LTL Specifications for Multi-Task RL
Linear temporal logic (LTL) has recently been adopted as a powerful formalism for specifying complex, temporally extended tasks in multi-task reinforcement learning (RL). However, learning policies that efficiently satisfy arbitrary specifications not observed during training remains a challenging problem. Existing approaches suffer from several shortcomings: they are often only applicable to finite-horizon fragments of LTL, are restricted to suboptimal solutions, and do not adequately handle safety constraints. In this work, we propose a novel learning approach to address these concerns. Our method leverages the structure of Büchi automata, which explicitly represent the semantics of LTL specifications, to learn policies conditioned on sequences of truth assignments that lead to satisfying the desired formulae. Experiments in a variety of discrete and continuous domains demonstrate that our approach is able to zero-shot satisfy a wide range of finite- and infinite-horizon specifications, and outperforms existing methods in terms of both satisfaction probability and efficiency. Code available at: https://deep-ltl.github.io/
****
Xinlei Chen,Zhuang Liu,Saining Xie,Kaiming He
Deconstructing Denoising Diffusion Models for Self-Supervised Learning
In this study, we examine the representation learning abilities of Denoising Diffusion Models (DDM) that were originally purposed for image generation. Our philosophy is to deconstruct a DDM, gradually transforming it into a classical Denoi

sing Autoencoder (DAE). This deconstructive process allows us to explore how var
ious components of modern DDMs influence self-supervised representation learning
. We observe that only a very few modern components are critical for learning go
od representations, while many others are nonessential. Our study ultimately arr
ives at an approach that is highly simplified and to a large extent resembles a
classical DAE. We hope our study will rekindle interest in a family of classical
 methods within the realm of modern self-supervised learning.
****

Yu-Zhe Shi,Mingchen Liu,Fanxu Meng,Qiao Xu,Zhangqian Bi,Kun He,Lecheng Ruan,Qini
ng Wang

Hierarchically Encapsulated Representation for Protocol Design in Self-Driving L
abs

Self-driving laboratories have begun to replace human experimenters in performin
g single experimental skills or predetermined experimental protocols. However, a
s the pace of idea iteration in scientific research has been intensified by Arti
ficial Intelligence, the demand for rapid design of new protocols for new discov
eries become evident. Efforts to automate protocol design have been initiated, b
ut the capabilities of knowledge-based machine designers, such as Large Language
 Models, have not been fully elicited, probably for the absence of a systematic
representation of experimental knowledge, as opposed to isolated, flatten pieces
 of information. To tackle this issue, we propose a multi-faceted, multi-scale r
epresentation, where instance actions, generalized operations, and product flow
models are hierarchically encapsulated using Domain-Specific Languages. We furth
er develop a data-driven algorithm based on non-parametric modeling that autonom
ously customizes these representations for specific domains. The proposed repres
entation is equipped with various machine designers to manage protocol design ta
sks, including planning, modification, and adjustment. The results demonstrate t
hat the proposed method could effectively complement Large Language Models in th
e protocol design process, serving as an auxiliary module in the realm of machin
e-assisted scientific exploration.
****

Rhea Sanjay Sukthanker,Arber Zela,Benedikt Staffler,Samuel Dooley,Josif Grabocka
,Frank Hutter

Multi-objective Differentiable Neural Architecture Search

Pareto front profiling in multi-objective optimization (MOO), i.e., finding a di
verse set of Pareto optimal solutions, is challenging, especially with expensive
 objectives that require training a neural network. Typically, in MOO for neural
 architecture search (NAS), we aim to balance performance and hardware metrics a
cross devices. Prior NAS approaches simplify this task by incorporating hardware
 constraints into the objective function, but profiling the Pareto front necessi
tates a computationally expensive search for each constraint. In this work, we p
ropose a novel NAS algorithm that encodes user preferences to trade-off performa
nce and hardware metrics, yielding representative and diverse architectures acro
ss multiple devices in just a single search run. To this end, we parameterize th
e joint architectural distribution across devices and multiple objectives via a
hypernetwork that can be conditioned on hardware features and preference vectors
, enabling zero-shot transferability to new devices. Extensive experiments invol
ving up to 19 hardware devices and 3 different objectives demonstrate the effect
iveness and scalability of our method. Finally, we show that, without any additi
onal costs, our method outperforms existing MOO NAS methods across a broad range
 of qualitatively different search spaces and datasets, including MobileNetV3 on
 ImageNet-1k, an encoder-decoder transformer space for machine translation and a
 decoder-only space for language modelling.
****

Kojiro Takeyama,Yimeng Liu,Misha Sra

LocoVR: Multiuser Indoor Locomotion Dataset in Virtual Reality

Understanding human locomotion is crucial for AI agents such as robots, particul
arly in complex indoor home environments. Modeling human trajectories in these s
paces requires insight into how individuals maneuver around physical obstacles a
nd manage social navigation dynamics. These dynamics include subtle behaviors in

fluenced by proxemics - the social use of space, such as stepping aside to allow others to pass or choosing longer routes to avoid collisions. Previous research has developed datasets of human motion in indoor scenes, but these are often limited in scale and lack the nuanced social navigation dynamics common in home environments.
To address this, we present LocoVR, a dataset of 7000+ two-person trajectories captured in virtual reality from over 130 different indoor home environments. LocoVR provides accurate trajectory and precise spatial information, along with rich examples of socially-motivated movement behaviors.
For example, the dataset captures instances of individuals navigating around each other in narrow spaces, adjusting paths to respect personal boundaries in living areas, and coordinating movements in high-traffic zones like entryways and kitchens. Our evaluation shows that LocoVR significantly enhances model performance in three practical indoor tasks utilizing human trajectories, and demonstrates predicting socially-aware navigation patterns in home environments.
****
Fengyuan Liu,Nikhil Kandpal,Colin Raffel
AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution
The influence of contextual input on the behavior of large language models (LLMs) has prompted the development of context attribution methods that aim to quantify each context span's effect on an LLM's generations. The leave-one-out (LOO) error, which measures the change in the likelihood of the LLM's response when a given span of the context is removed, provides a principled way to perform context attribution, but can be prohibitively expensive to compute for large models. In this work, we introduce AttriBoT, a series of novel techniques for efficiently computing an approximation of the LOO error for context attribution. Specifically, AttriBoT uses cached activations to avoid redundant operations, performs hierarchical attribution to reduce computation, and emulates the behavior of large target models with smaller proxy models. Taken together, AttriBoT can provide a 300x speedup while remaining more faithful to a target model's LOO error than prior context attribution methods. This stark increase in performance makes computing context attributions for a given response $30\times$ faster than generating the response itself, empowering real-world applications that require computing attributions at scale. We release a user-friendly and efficient implementation of AttriBoT to enable efficient LLM interpretability as well as encourage future development of efficient context attribution methods.
****
Tom George,Pierre Glaser,Kim Stachenfeld,Caswell Barry,Claudia Clopath
SIMPL: Scalable and hassle-free optimisation of neural representations from behaviour
Neural activity in the brain is known to encode low-dimensional, time-evolving, behaviour-related variables. A long-standing goal of neural data analysis has been to identify these variables and their mapping to neural activity. A productive and canonical approach has been to simply visualise neural "tuning curves" as a function of behaviour. However, significant discrepancies between behaviour and the true latent variables -- such as an agent thinking of position Y whilst located at position X -- distort and blur the tuning curves, decreasing their interpretability. To address this, latent variable models propose to learn the latent variable from data; these are typically expensive, hard to tune, or scale poorly, complicating their adoption. Here we propose SIMPL (Scalable Iterative Maximization of Population-coded Latents), an EM-style algorithm which iteratively optimises latent variables and tuning curves. SIMPL is fast, scalable and exploits behaviour as an initial condition to further improve convergence and identifiability. It can accurately recover latent variables in spatial and non-spatial tasks. When applied to a large hippocampal dataset SIMPL converges on smaller, more numerous, and more uniformly sized place fields than those based on behaviour, suggesting the brain may encode space with greater resolution than previously thought.
****

Aniket Vashishtha,Abbavaram Gowtham Reddy,Abhinav Kumar,Saketh Bachu,Vineeth N. Balasubramanian,Amit Sharma
Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference
Large Language Models (LLMs) have recently been used as experts to infer causal graphs, often by repeatedly applying a pairwise prompt that asks about the causal relationship of each variable pair. However, such experts, including human domain experts, cannot distinguish between direct and indirect effects given a pairwise prompt. Therefore, instead of the graph, we propose that causal order be used as a more stable output interface for utilizing expert knowledge. When querying a perfect expert with a pairwise prompt, we show that the inferred graph can have significant errors whereas the causal order is always correct. In practice, however, LLMs are imperfect experts and we find that pairwise prompts lead to multiple cycles and do not yield a valid order. Hence, we propose a prompting strategy that introduces an auxiliary variable for every variable pair and instructs the LLM to avoid cycles within this triplet. We show, both theoretically and empirically, that such a triplet prompt leads to fewer cycles than the pairwise prompt. Across multiple real-world graphs, the triplet prompt yields a more accurate order using both LLMs and human annotators as experts. By querying the expert with different auxiliary variables for the same variable pair, it also increases robustness---triplet method with much smaller models such as Phi-3 and Llama-3 8B outperforms a pairwise prompt with GPT-4. For practical usage, we show how the estimated causal order from the triplet method can be used to reduce error in downstream discovery and effect inference tasks.
****
Junjie Xu,Artem Moskalev,Tommaso Mansi,Mangal Prakash,Rui Liao
Beyond Sequence: Impact of Geometric Context for RNA Property Prediction
Accurate prediction of RNA properties, such as stability and interactions, is crucial for advancing our understanding of biological processes and developing RNA-based therapeutics. RNA structures can be represented as 1D sequences, 2D topological graphs, or 3D all-atom models, each offering different insights into its function. Existing works predominantly focus on 1D sequence-based models, which overlook the geometric context provided by 2D and 3D geometries. This study presents the first systematic evaluation of incorporating explicit 2D and 3D geometric information into RNA property prediction, considering not only performance but also real-world challenges such as limited data availability, partial labeling, sequencing noise, and computational efficiency. To this end, we introduce a newly curated set of RNA datasets with enhanced 2D and 3D structural annotations, providing a resource for model evaluation on RNA data. Our findings reveal that models with explicit geometry encoding generally outperform sequence-based models, with an average prediction RMSE reduction of around 12% across all various RNA tasks and excelling in low-data and partial labeling regimes, underscoring the value of explicitly incorporating geometric context. On the other hand, geometry-unaware sequence-based models are more robust under sequencing noise but often require around 2-5x training data to match the performance of geometry-aware models. Our study offers further insights into the trade-offs between different RNA representations in practical applications and addresses a significant gap in evaluating deep learning models for RNA tasks.
****
Zhenyu Sun,ziyang zhang,Zheng Xu,Gauri Joshi,Pranay Sharma,Ermin Wei
Debiasing Federated Learning with Correlated Client Participation
In cross-device federated learning (FL) with millions of mobile clients, only a small subset of clients participate in training in every communication round, and Federated Averaging (FedAvg) is the most popular algorithm in practice. Existing analyses of FedAvg usually assume the participating clients are independently sampled in each round from a uniform distribution, which does not reflect real-world scenarios. This paper introduces a theoretical framework that models client participation in FL as a Markov chain to study optimization convergence when clients have non-uniform and correlated participation across rounds.
We apply this framework to analyze a more practical pattern: every client must wait a minimum number of R rounds (minimum separation) before re-participating.

We theoretically prove and empirically observe that increasing minimum separati
on reduces the bias induced by intrinsic non-uniformity of client availability i
n cross-device FL systems.
Furthermore, we develop an effective debiasing algorithm for FedAvg that provabl
y converges to the unbiased optimal solution under arbitrary minimum separation
and unknown client availability distribution.
****
Pei Zhou,Ruizhe Liu,Qian Luo,Fan Wang,Yibing Song,Yanchao Yang
AutoCGP: Closed-Loop Concept-Guided Policies from Unlabeled Demonstrations
Training embodied agents to perform complex robotic tasks presents significant c
hallenges due to the entangled factors of task compositionality, environmental d
iversity, and dynamic changes. In this work, we introduce a novel imitation lear
ning framework to train closed-loop concept-guided policies that enhance long-ho
rizon task performance by leveraging discovered manipulation concepts. Unlike me
thods that rely on predefined skills and human-annotated labels, our approach al
lows agents to autonomously abstract manipulation concepts from their propriocep
tive states, thereby alleviating misalignment due to ambiguities in human semant
ics and environmental complexity. Our framework comprises two primary components
: an *Automatic Concept Discovery* module that identifies meaningful and consist
ent manipulation concepts, and a *Concept-Guided Policy Learning* module that ef
fectively utilizes these manipulation concepts for adaptive task execution, incl
uding a *Concept Selection Transformer* for concept-based guidance and a *Concep
t-Guided Policy* for action prediction with the selected concepts. Experiments d
emonstrate that our approach significantly outperforms baseline methods across a
 range of tasks and environments, while showcasing emergent consistency in motio
n patterns associated with the discovered manipulation concepts. Codes are avail
able at: https://github.com/PeiZhou26/AutoCGP.
****
Jiale Cheng,Xiao Liu,Cunxiang Wang,Xiaotao Gu,Yida Lu,Dan Zhang,Yuxiao Dong,Jie
Tang,Hongning Wang,Minlie Huang
SPaR: Self-Play with Tree-Search Refinement to Improve Instruction-Following in
Large Language Models
Instruction-following is a fundamental capability of language models, requiring
the model to recognize even the most subtle requirements in the instructions and
 accurately reflect them in its output.
Such an ability is well-suited for and often optimized by preference learning.
However, existing methods often directly sample multiple independent responses f
rom the model when creating preference pairs.
Such practice can introduce content variations irrelevant to whether the instruc
tion is precisely followed (e.g., different expressions about the same semantic)
, interfering with the goal of teaching models to recognize the key differences
that lead to improved instruction following.
In light of this, we introduce SPaR, a self-play framework integrating tree-sear
ch self-refinement to yield valid and comparable preference pairs free from dist
ractions.
By playing against itself, an LLM employs a tree-search strategy to refine its p
revious responses with respect to the instruction while minimizing unnecessary v
ariations.
Our experiments show that a LLaMA3-8B model, trained over three iterations guide
d by SPaR, surpasses GPT-4-Turbo on the IFEval benchmark without losing general
capabilities.
Furthermore, SPaR demonstrates promising scalability, greatly enhancing models l
ike GLM-4-9B and LLaMA3-70B.
We also identify how inference scaling in tree search would impact model perform
ance.
Our code and data are publicly available at https://github.com/thu-coai/SPaR.
****
Patrick Leask,Bart Bussmann,Michael T Pearce,Joseph Isaac Bloom,Curt Tigges,Nour
a Al Moubayed,Lee Sharkey,Neel Nanda
Sparse Autoencoders Do Not Find Canonical Units of Analysis

A common goal of mechanistic interpretability is to decompose the activations of neural networks into features: interpretable properties of the input computed by the model. Sparse autoencoders (SAEs) are a popular method for finding these features in LLMs, and it has been postulated that they can be used to find a canonical set of units: a unique and complete list of atomic features. We cast doubt on this belief using two novel techniques: SAE stitching to show they are incomplete, and meta-SAEs to show they are not atomic. SAE stitching involves inserting or swapping latents from a larger SAE into a smaller one. Latents from the larger SAE can be divided into two categories: novel latents, which improve performance when added to the smaller SAE, indicating they capture novel information, and reconstruction latents, which can replace corresponding latents in the smaller SAE that have similar behavior. The existence of novel features indicates incompleteness of smaller SAEs. Using meta-SAEs - SAEs trained on the decoder matrix of another SAE - we find that latents in SAEs often decompose into combinations of latents from a smaller SAE, showing that larger SAE latents are not atomic. The resulting decompositions are often interpretable; e.g. a latent representing "Einstein" decomposes into "scientist", "Germany", and "famous person". To train meta-SAEs we introduce BatchTopK SAEs, an improved variant of the popular TopK SAE method, that only enforces a fixed average sparsity. Even if SAEs do not find canonical units of analysis, they may still be useful tools. We suggest that future research should either pursue different approaches for identifying such units, or pragmatically choose the SAE size suited to their task. We provide an interactive dashboard to explore meta-SAEs: https://metasaes.streamlit.app/
****
Abulhair Saparov,Srushti Ajay Pawar,Shreyas Pimpalgaonkar,Nitish Joshi,Richard Yuanzhe Pang,Vishakh Padmakumar,Mehran Kazemi,Najoung Kim,He He
Transformers Struggle to Learn to Search
Search is an ability foundational in many important tasks, and recent studies have shown that large language models (LLMs) struggle to perform search robustly. It is unknown whether this inability is due to a lack of data, insufficient model parameters, or fundamental limitations of the transformer architecture. In this work, we use the foundational graph connectivity problem as a testbed to generate effectively limitless high-coverage data to train small transformers and test whether they can learn to perform search. We find that, when given the right training distribution, the transformer is able to learn to search.

We analyze the algorithm that the transformer has learned through a novel mechanistic interpretability technique that enables us to extract the computation graph from the trained model. We find that for each vertex in the input graph, transformers compute the set of vertices reachable from that vertex. Each layer then progressively expands these sets, allowing the model to search over a number of vertices exponential in the number of layers.

However, we find that as the input graph size increases, the transformer has greater difficulty in learning the task. This difficulty is not resolved even as the number of parameters is increased, suggesting that increasing model scale will not lead to robust search abilities. We also find that performing search in-context (i.e., chain-of-thought) does not resolve this inability to learn to search on larger graphs.
****
Jianting Yang,Srecko Durasinovic,Jean B. Lasserre,Victor Magron,Jun Zhao
Verifying Properties of Binary Neural Networks Using Sparse Polynomial Optimization
This paper explores methods for verifying the properties of Binary Neural Networks (BNNs), focusing on robustness against adversarial attacks. Despite their lower computational and memory needs, BNNs, like their full-precision counterparts, are also sensitive to input perturbations. Established methods for solving this problem are predominantly based on Satisfiability Modulo Theories and Mixed-Integer Linear Programming techniques, which are characterized by NP complexity and often face scalability issues.

We introduce an alternative approach using Semidefinite Programming relaxations derived from sparse Polynomial Optimization. Our approach, compatible with continuous input space, not only mitigates numerical issues associated with floating-point calculations but also enhances verification scalability through the strategic use of tighter first-order semidefinite relaxations. We demonstrate the effectiveness of our method in verifying robustness against both $\||.|\|_\infty$ and $\||.|\|_2$-based adversarial attacks.
****
Suorong Yang,Peng Ye,Wanli Ouyang,Dongzhan Zhou,Furao Shen
A CLIP-Powered Framework for Robust and Generalizable Data Selection
Large-scale datasets have been pivotal to the advancements of deep learning models in recent years, but training on such large datasets inevitably incurs substantial storage and computational overhead.
Meanwhile, real-world datasets often contain redundant and noisy data, imposing a negative impact on training efficiency and model performance.
Data selection has shown promise in identifying the most representative samples from the entire dataset, which aims to minimize the performance gap with reduced training costs.
Existing works typically rely on single-modality information to assign importance scores for individual samples, which may lead to inaccurate assessments, especially when dealing with noisy or corrupted samples.
To address this limitation, we propose a novel CLIP-powered data selection framework that leverages multimodal information for more robust and generalizable sample selection.
Specifically, our framework consists of three key modules—dataset adaptation, sample scoring, and selection optimization—that together harness extensive pre-trained multimodal knowledge to comprehensively assess sample influence and optimize the selection results through multi-objective optimization.
Extensive experiments demonstrate that our approach consistently outperforms existing state-of-the-art baselines on various benchmark datasets. Notably, our method effectively removes noisy or damaged samples from the dataset, enabling it to achieve even higher performance with less data. This indicates that it is not only a way to accelerate training but can also improve overall data quality.
The implementation is available at https://github.com/Jackbrocp/clip-powered-data-selection.
****
Haoran Chen,Micah Goldblum,Zuxuan Wu,Yu-Gang Jiang
Adaptive Retention & Correction: Test-Time Training for Continual Learning
Continual learning, also known as lifelong learning or incremental learning, refers to the process by which a model learns from a stream of incoming data over time. A common problem in continual learning is the classification layer’s bias towards the most recent task. Traditionally, methods have relied on incorporating data from past tasks during training to mitigate this issue. However, the recent shift in continual learning to memory-free environments has rendered these approaches infeasible. In this study, we propose a solution focused on the testing phase. We first introduce a simple Out-of-Task Detection method, OTD, designed to accurately identify samples from past tasks during testing. Leveraging OTD, we then propose: (1) an Adaptive Retention mechanism for dynamically tuning the classifier layer on past task data; (2) an Adaptive Correction mechanism for revising predictions when the model classifies data from previous tasks into classes from the current task. We name our approach Adaptive Retention & Correction (ARC). While designed for memory-free environments, ARC also proves effective in memorybased settings. Extensive experiments show that our proposed method can be plugged in to virtually any existing continual learning approach without requiring any modifications to its training procedure. Specifically, when integrated with state-of-the-art approaches, ARC achieves an average performance increase of 2.7% and 2.6% on the CIFAR-100 and Imagenet-R datasets, respectively
****
Muhammad Anwar Ma'sum,Mahardhika Pratama,Savitha Ramasamy,Lin Liu,H Habibullah,Ryszard Kowalczyk

Vision and Language Synergy for Rehearsal Free Continual Learning
The prompt-based approach has demonstrated its success for continual learning pr oblems. However, it still suffers from catastrophic forgetting due to inter-task vector similarity and unfitted new components of previously learned tasks. On t he other hand, the language-guided approach falls short of its full potential du e to minimum utilized knowledge and participation in the prompt tuning process. To correct this problem, we propose a novel prompt-based structure and algorithm that incorporate 4 key concepts (1) language as input for prompt generation (2) task-wise generators (3) limiting matching descriptors search space via soft ta sk-id prediction (4) generated prompt as auxiliary data. Our experimental analys is shows the superiority of our method to existing SOTAs in CIFAR100, ImageNet-R , and CUB datasets with significant margins i.e. up to 30% final average accurac y, 24% cumulative average accuracy, 8% final forgetting measure, and 7% cumulati ve forgetting measure. Our historical analysis confirms our method successfully maintains the stability-plasticity trade-off in every task. Our robustness analy sis shows the proposed method consistently achieves high performances in various prompt lengths, layer depths, and number of generators per task compared to the SOTAs. We provide a comprehensive theoretical analysis, and complete numerical results in appendix sections. The method code is available in https://github.com /anwarmaxsum/LEAPGEN for further study.
****

Yi Xu,Yun Fu
Sports-Traj: A Unified Trajectory Generation Model for Multi-Agent Movement in S ports
Understanding multi-agent movement is critical across various fields. The conven tional approaches typically focus on separate tasks such as trajectory predictio n, imputation, or spatial-temporal recovery. Considering the unique formulation and constraint of each task, most existing methods are tailored for only one, li miting the ability to handle multiple tasks simultaneously, which is a common re quirement in real-world scenarios. Another limitation is that widely used public datasets mainly focus on pedestrian movements with casual, loosely connected pa tterns, where interactions between individuals are not always present, especiall y at a long distance, making them less representative of more structured environ ments. To overcome these limitations, we propose a Unified Trajectory Generation model, UniTraj, that processes arbitrary trajectories as masked inputs, adaptab le to diverse scenarios in the domain of sports games. Specifically, we introduc e a Ghost Spatial Masking (GSM) module, embedded within a Transformer encoder, f or spatial feature extraction. We further extend recent State Space Models (SSMs), known as the Mamba model, into a Bidirectional Temporal Mamba (BTM) to better capture temporal dependencies. Additionally, we incorporate a Bidirectional Tem poral Scaled (BTS) module to thoroughly scan trajectories while preserving tempo ral missing relationships. Furthermore, we curate and benchmark three practical sports datasets, Basketball-U, Football-U, and Soccer-U, for evaluation. Extensi ve experiments demonstrate the superior performance of our model. We hope that o ur work can advance the understanding of human movement in real-world applicatio ns, particularly in sports. Our datasets, code, and model weights are available here https://github.com/colorfulfuture/UniTraj-pytorch.
****

Zhaolin Hu,Yixiao Zhou,Zhongan Wang,Xin Li,Weimin Yang,Hehe Fan,Yi Yang
OSDA Agent: Leveraging Large Language Models for De Novo Design of Organic Struc ture Directing Agents
Zeolites are crystalline porous materials that have been widely utilized in petr ochemical industries as well as sustainable chemistry areas. Synthesis of zeolit es often requires small molecules termed Organic Structure Directing Agents (OSD As), which are critical in forming the porous structure. Molecule generation mod els can aid the design of OSDAs, but they are limited by single functionality an d lack of interactivity. Meanwhile, large language models (LLMs) such as GPT-4, as general-purpose artificial intelligence systems, excel in instruction compreh ension, logical reasoning, and interactive communication. However, LLMs lack in- depth chemistry knowledge and first-principle computation capabilities, resultin

g in uncontrollable outcomes even after fine-tuning. In this paper, we propose O
SDA Agent, an interactive OSDA design framework that leverages LLMs as the brain
, coupled with computational chemistry tools. The OSDA Agent consists of three m
ain components: the Actor, responsible for generating potential OSDA structures;
 the Evaluator, which assesses and scores the generated OSDAs using computationa
l chemistry tools; and the Self-reflector, which produces reflective summaries b
ased on the Evaluator's feedback to refine the Actor's subsequent outputs. Exper
iments on representative zeolite frameworks show the generation-evaluation-refle
ction-refinement workflow can perform de novo design of OSDAs with superior gene
ration quality than the pure LLM model, generating candidates consistent with ex
perimentally validated OSDAs and optimizing known OSDAs.
****
Samy Jelassi,Clara Mohri,David Brandfonbrener,Alex Gu,Nikhil Vyas,Nikhil Anand,D
avid Alvarez-Melis,Yuanzhi Li,Sham M. Kakade,eran malach
Mixture of Parrots: Experts improve memorization more than reasoning
The Mixture-of-Experts (MoE) architecture enables a significant increase in the
total number of model parameters with minimal computational overhead.
However, it is not clear what performance tradeoffs, if any, exist between MoEs
and standard dense transformers.
In this paper,
we show that as we increase the number of experts (while fixing the number of ac
tive parameters), the memorization performance consistently increases while the
reasoning capabilities saturate.

We begin by analyzing the theoretical limitations of MoEs at reasoning. We prove
 that there exist graph problems that cannot be solved by any number of experts
 of a certain width; however, the same task can be easily solved by a dense mode
l with a slightly larger width.
On the other hand, we find that on memory-intensive tasks, MoEs can effectively
leverage a small number of active parameters with a large number of experts to m
emorize the data.
We empirically validate these findings on synthetic graph problems and memory-in
tensive closed book retrieval tasks.
Lastly, we pre-train a series of MoEs and dense transformers and evaluate them
on commonly used benchmarks in math and natural language.
We find that increasing the number of experts helps solve knowledge-intensive ta
sks, but fails to yield the same benefits for reasoning tasks.
****
Qingni Wang,Tiantian Geng,Zhiyuan Wang,Teng Wang,Bo Fu,Feng Zheng
Sample then Identify: A General Framework for Risk Control and Assessment in Mul
timodal Large Language Models
Multimodal Large Language Models (MLLMs) exhibit promising advancements across v
arious tasks, yet they still encounter significant trustworthiness issues. Prior
 studies apply Split Conformal Prediction (SCP) in language modeling to construc
t prediction sets with statistical guarantees. However, these methods typically
rely on internal model logits or are restricted to multiple-choice settings, whi
ch hampers their generalizability and adaptability in dynamic, open-ended enviro
nments. In this paper, we introduce *TRON*, a **t**wo-step framework for **r**is
k c**o**ntrol and assessme**n**t, applicable to any MLLM that supports sampling
in both open-ended and closed-ended scenarios. *TRON* comprises two main compone
nts: (1) a novel conformal score to **sample** response sets of minimum size, an
d (2) a nonconformity score to **identify** high-quality responses based on self
-consistency theory, controlling the error rates by two specific risk levels. Fu
rthermore, we investigate semantic redundancy in prediction sets within open-end
ed contexts for the first time, leading to a promising evaluation metric for MLL
Ms based on average set size. Our comprehensive experiments across four Video Qu
estion-Answering (VideoQA) datasets utilizing eight MLLMs show that *TRON* achie
ves desired error rates bounded by two user-specified risk levels. Additionally,
 deduplicated prediction sets maintain adaptiveness while being more efficient a

nd stable for risk assessment under different risk levels.
****
Zhenyu Zhang,Zechun Liu,Yuandong Tian,Harshit Khaitan,Zhangyang Wang,Steven Li
R-Sparse: Rank-Aware Activation Sparsity for Efficient LLM Inference
Large Language Models (LLMs), while demonstrating remarkable capabilities across various applications, present significant challenges during inference due to their substantial model size, especially when deployed on edge devices. Activation sparsity offers a promising solution to reduce computation and memory movement, enabling more efficient inference, particularly for small-batch on-device applications. However, current approaches face limitations with non-ReLU activation function, which are foundational to most advanced LLMs, or require heavy continual training. Additionally, the difficulty in predicting active channels and limited achievable sparsity ratios constrain the effectiveness of activation sparsity-based methods. In this paper, we introduce R-Sparse, a training-free activation sparsity approach capable of achieving high sparsity levels in advanced LLMs. We conducted two preliminary investigations into how different components contribute to the output within a single linear layer and found two key observations: (i) the non-sparse components of the input function can be regarded as a few bias terms, and (ii) The full computation can be effectively approximated by an appropriate combination of input channels and weight singular values. Building on this, we replace the linear layers in LLMs with a rank-aware sparse inference method that leverages the sparsity of input channels and singular value components, eliminating the need for active channel prediction like the output sparsity based approaches. Experiments on Llama-2/3 and Mistral models across ten diverse tasks demonstrate that R-Sparse achieves comparable performance at 50\% model-level sparsity, resulting in a significant 43\% end-to-end efficient improvements with customized kernels.
****
Alexandros Hollender,Gilbert Maystre,Sai Ganesh Nagarajan
The Complexity of Two-Team Polymatrix Games with Independent Adversaries
Adversarial multiplayer games are an important object of study in multiagent learning. In particular, polymatrix zero-sum games are a multiplayer setting where Nash equilibria are known to be efficiently computable. Towards understanding the limits of tractability in polymatrix games, we study the computation of Nash equilibria in such games where each pair of players plays either a zero-sum or a coordination game. We are particularly interested in the setting where players can be grouped into a small number of teams of identical interest. While the three-team version of the problem is known to be PPAD-complete, the complexity for two teams has remained open. Our main contribution is to prove that the two-team version remains hard, namely it is CLS-hard. Furthermore, we show that this lower bound is tight for the setting where one of the teams consists of multiple independent adversaries. On the way to obtaining our main result, we prove hardness of finding any stationary point in the simplest type of non-convex-concave min-max constrained optimization problem, namely for a class of bilinear polynomial objective functions.
****
Hyuna Cho,Ziquan Wei,Seungjoo Lee,Tingting Dan,Guorong Wu,Won Hwa Kim
Conditional Diffusion with Ordinal Regression: Longitudinal Data Generation for Neurodegenerative Disease Studies
Modeling the progression of neurodegenerative diseases such as Alzheimer’s disease (AD) is crucial for early detection and prevention given their irreversible nature. However, the scarcity of longitudinal data and complex disease dynamics make the analysis highly challenging. Moreover, longitudinal samples often contain irregular and large intervals between subject visits, which underscore the necessity for advanced data generation techniques that can accurately simulate disease progression over time. In this regime, we propose a novel conditional generative model for synthesizing longitudinal sequences and present its application to neurodegenerative disease data generation conditioned on multiple time-dependent ordinal factors, such as age and disease severity. Our method sequentially generates continuous data by bridging gaps between sparse data points with a diffu

sion model, ensuring a realistic representation of disease progression. The synt
hetic data are curated to integrate both cohort-level and individual-specific ch
aracteristics, where the cohort-level representations are modeled with an ordina
l regression to capture longitudinally monotonic behavior. Extensive experiments
 on four AD biomarkers validate the superiority of our method over nine baseline
 approaches, highlighting its potential to be applied to a variety of longitudin
al data generation.
****
Thomas Hehn,Markus Peschl,Tribhuvanesh Orekondy,Arash Behboodi,Johann Brehmer
Differentiable and Learnable Wireless Simulation with Geometric Transformers
Modelling the propagation of electromagnetic wireless signals is critical for de
signing modern communication systems. Wireless ray tracing simulators model sign
al propagation based on the 3D geometry and other scene parameters, but their ac
curacy is fundamentally limited by underlying modelling assumptions and correctn
ess of parameters. In this work, we introduce Wi-GATr, a fully-learnable neural
simulation surrogate designed to predict the channel observations based on scene
 primitives (e. g., surface mesh, antenna position and orientation). Recognizing
 the inherently geometric nature of these primitives, Wi-GATr leverages an equiv
ariant Geometric Algebra Transformer that operates on a tokenizer specifically t
ailored for wireless simulation. We evaluate our approach on a range of tasks (i
. e., signal strength and delay spread prediction, receiver localization, and ge
ometry reconstruction) and find that Wi-GATr is accurate, fast, sample-efficient
, and robust to symmetry-induced transformations. Remarkably, we find our result
s also translate well to the real world: Wi-GATr demonstrates more than 35% lowe
r error than hybrid techniques, and 70% lower error than a calibrated wireless t
racer.
****
Lazar Atanackovic,Xi Zhang,Brandon Amos,Mathieu Blanchette,Leo J Lee,Yoshua Beng
io,Alexander Tong,Kirill Neklyudov
Meta Flow Matching: Integrating Vector Fields on the Wasserstein Manifold
Numerous biological and physical processes can be modeled as systems of interact
ing entities evolving continuously over time, e.g. the dynamics of communicating
 cells or physical particles. Learning the dynamics of such systems is essential
 for predicting the temporal evolution of populations across novel samples and u
nseen environments. Flow-based models allow for learning these dynamics at the p
opulation level - they model the evolution of the entire distribution of samples
. However, current flow-based models are limited to a single initial population
and a set of predefined conditions which describe different dynamics. We argue t
hat multiple processes in natural sciences have to be represented as vector fiel
ds on the Wasserstein manifold of probability densities. That is, the change of
the population at any moment in time depends on the population itself due to the
 interactions between samples. In particular, this is crucial for personalized m
edicine where the development of diseases and their respective treatment respons
e depend on the microenvironment of cells specific to each patient. We propose *
Meta Flow Matching* (MFM), a practical approach to integrate along these vector
fields on the Wasserstein manifold by amortizing the flow model over the initial
 populations. Namely, we embed the population of samples using a Graph Neural Ne
twork (GNN) and use these embeddings to train a Flow Matching model. This gives
MFM the ability to generalize over the initial distributions, unlike previously
proposed methods. We demonstrate the ability of MFM to improve the prediction of
 individual treatment responses on a large-scale multi-patient single-cell drug
screen dataset.
****
Xin Gu,Ming Li,Libo Zhang,Fan Chen,Longyin Wen,Tiejian Luo,Sijie Zhu
Multi-Reward as Condition for Instruction-based Image Editing
High-quality training triplets (instruction, original image, edited image) are e
ssential for instruction-based image editing. Predominant training datasets (e.g
., InsPix2Pix) are created using text-to-image generative models (e.g., Stable D
iffusion, DALL-E) which are not trained for image editing. Accordingly, these da
tasets suffer from inaccurate instruction following, poor detail preserving, and

generation artifacts. In this paper, we propose to address the training data quality issue with multi-perspective reward data instead of refining the ground-truth image quality. 1) we first design a quantitative metric system based on best-in-class LVLM (Large Vision Language Model), i.e., GPT-4o in our case, to evaluate the generation quality from 3 perspectives, namely, instruction following, detail preserving, and generation quality. For each perspective, we collected quantitative score in $0\sim 5$ and text descriptive feedback on the specific failure points in ground-truth edited images, resulting in a high-quality editing reward dataset, i.e., RewardEdit20K. 2) We further proposed a novel training framework to seamlessly integrate the metric output, regarded as multi-reward, into editing models to learn from the imperfect training triplets. During training, the reward scores and text descriptions are encoded as embeddings and fed into both the latent space and the U-Net of the editing models as auxiliary conditions. During inference, we set these additional conditions to the highest score with no text description for failure points, to aim at the best generation outcome. 3) We also build a challenging evaluation benchmark with real-world images/photos and diverse editing instructions, named as Real-Edit. Experiments indicate that our multi-reward conditioned model outperforms its no-reward counterpart on two popular editing pipelines, i.e., InsPix2Pix and SmartEdit. Code is released at https://github.com/bytedance/Multi-Reward-Editing.
****

Xiaochuan Li,Zichun Yu,Chenyan Xiong

Montessori-Instruct: Generate Influential Training Data Tailored for Student Learning

Synthetic data has been widely used to train large language models, but their generative nature inevitably introduces noisy, non-informative, and misleading learning signals. In this paper, we propose Montessori-Instruct, a novel data synthesis framework that tailors the data synthesis ability of the teacher language model toward the student language model's learning process. Specifically, we utilize local data influence of synthetic training data points on students to characterize students' learning preferences. Then, we train the teacher model with Direct Preference Optimization (DPO) to generate synthetic data tailored toward student learning preferences. Experiments with Llama3-8B-Instruct (teacher) and Llama3-8B (student) on Alpaca Eval and MT-Bench demonstrate that Montessori-Instruct significantly outperforms standard synthesis methods by 18.35\% and 46.24\% relatively. Our method also beats data synthesized by a stronger teacher model, GPT-4o. Further analysis confirms the benefits of teacher's learning to generate more influential training data in the student's improved learning, the advantages of local data influence in accurately measuring student preferences, and the robustness of Montessori-Instruct across different student models. Our code and data are open-sourced at https://github.com/cxcscmu/Montessori-Instruct.
****

Negin Raoof,Litu Rout,Giannis Daras,Sujay Sanghavi,Constantine Caramanis,Sanjay Shakkottai,Alex Dimakis

Infilling Score: A Pretraining Data Detection Algorithm for Large Language Models

In pretraining data detection, the goal is to detect whether a given sentence is in the dataset used for training a Large Language Model LLM). Recent methods (such as Min-K % and Min-K%++) reveal that most training corpora are likely contaminated with both sensitive content and evaluation benchmarks, leading to inflated test set performance. These methods sometimes fail to detect samples from the pretraining data, primarily because they depend on statistics composed of causal token likelihoods. We introduce Infilling Score, a new test-statistic based on non-causal token likelihoods. Infilling Score can be computed for autoregressive models without re-training using Bayes rule. A naive application of Bayes rule scales linearly with the vocabulary size. However, we propose a ratio test-statistic whose computation is invariant to vocabulary size. Empirically, our method achieves a significant accuracy gain over state-of-the-art methods including Min-K%, and Min-K%++ on the WikiMIA benchmark across seven models with different parameter sizes. Further, we achieve higher AUC compared to reference-free methods

on the challenging MIMIR benchmark. Finally, we create a benchmark dataset consisting of recent data sources published after the release of Llama-3; this benchmark provides a statistical baseline to indicate potential corpora used for Llama-3 training.
****
June Suk Choi,Kyungmin Lee,Jongheon Jeong,Saining Xie,Jinwoo Shin,Kimin Lee
DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing
Recent advances in diffusion models have introduced a new era of text-guided image manipulation, enabling users to create realistic edited images with simple textual prompts. However, there is significant concern about the potential misuse of these methods, especially in creating misleading or harmful content. Although recent defense strategies, which introduce imperceptible adversarial noise to induce model failure, have shown promise, they remain ineffective against more sophisticated manipulations, such as editing with a mask. In this work, we propose DiffusionGuard, a robust and effective defense method against unauthorized edits by diffusion-based image editing models, even in challenging setups. Through a detailed analysis of these models, we introduce a novel objective that generates adversarial noise targeting the early stage of the diffusion process. This approach significantly improves the efficiency and effectiveness of adversarial noises. We also introduce a mask-augmentation technique to enhance robustness against various masks during test time. Finally, we introduce a comprehensive benchmark designed to evaluate the effectiveness and robustness of methods in protecting against privacy threats in realistic scenarios. Through extensive experiments, we show that our method achieves stronger protection and improved mask robustness with lower computational costs compared to the strongest baseline. Additionally, our method exhibits superior transferability and better resilience to noise removal techniques compared to all baseline methods. Our source code is publicly available at https://choi403.github.io/diffusionguard.
****
Caigao JIANG,Xiang Shu,Hong Qian,Xingyu Lu,JUN ZHOU,Aimin Zhou,Yang Yu
LLMOPT: Learning to Define and Solve General Optimization Problems from Scratch
Optimization problems are prevalent across various scenarios. Formulating and then solving optimization problems described by natural language often requires highly specialized human expertise, which could block the widespread application of optimization-based decision making. To automate problem formulation and solving, leveraging large language models (LLMs) has emerged as a potential way. However, this kind of approach suffers from the issue of optimization generalization. Namely, the accuracy of most current LLM-based methods and the generality of optimization problem types that they can model are still limited. In this paper, we propose a unified learning-based framework called LLMOPT to boost optimization generalization. Starting from the natural language descriptions of optimization problems and a pre-trained LLM, LLMOPT constructs the introduced five-element formulation as a universal model for learning to define diverse optimization problem types. Then, LLMOPT employs the multi-instruction tuning to enhance both problem formalization and solver code generation accuracy and generality. After that, to prevent hallucinations in LLMs, such as sacrificing solving accuracy to avoid execution errors, the model alignment and self-correction mechanism are adopted in LLMOPT. We evaluate the optimization generalization ability of LLMOPT and compared methods across six real-world datasets covering roughly 20 fields such as health, environment, energy and manufacturing, etc. Extensive experiment results show that LLMOPT is able to model various optimization problem types such as linear/nonlinear programming, mixed integer programming, and combinatorial optimization, and achieves a notable 11.08% average solving accuracy improvement compared with the state-of-the-art methods. The code is available at https://github.com/caigaojiang/LLMOPT.
****
XiaoHua Feng,Yuyuan Li,Chaochao Chen,Li Zhang,Longfei Li,JUN ZHOU,Xiaolin Zheng
Controllable Unlearning for Image-to-Image Generative Models via ϵ-Constrained Optimization
While generative models have made significant advancements in recent years, they

also raise concerns such as privacy breaches and biases. Machine unlearning has emerged as a viable solution, aiming to remove specific training data, e.g., co ntaining private information and bias, from models. In this paper, we study the machine unlearning problem in Image-to-Image (I2I) generative models. Previous s tudies mainly treat it as a single objective optimization problem, offering a so litary solution, thereby neglecting the varied user expectations towards the tra de-off between complete unlearning and model utility. To address this issue, we propose a controllable unlearning framework that uses a control coefficient $\ep silon$ to control the trade-off. We reformulate the I2I generative model unlearn ing problem into a ϵ-constrained optimization problem and solve it with a gradient-based method to find optimal solutions for unlearning boundaries. Th ese boundaries define the valid range for the control coefficient. Within this r ange, every yielded solution is theoretically guaranteed with Pareto optimality. We also analyze the convergence rate of our framework under various control fun ctions. Extensive experiments on two benchmark datasets across three mainstream I2I models demonstrate the effectiveness of our controllable unlearning framewor k.
****
Yutong Chen,Marko Mihajlovic,Xiyi Chen,Yiming Wang,Sergey Prokudin,Siyu Tang
SplatFormer: Point Transformer for Robust 3D Gaussian Splatting
3D Gaussian Splatting (3DGS) has recently transformed photorealistic reconstruct ion, achieving high visual fidelity and real-time performance. However, renderin g quality significantly deteriorates when test views deviate from the camera ang les used during training, posing a major challenge for applications in immersive free-viewpoint rendering and navigation. In this work, we conduct a comprehensi ve evaluation of 3DGS and related novel view synthesis methods under out-of-dist ribution (OOD) test camera scenarios. By creating diverse test cases with synthe tic and real-world datasets, we demonstrate that most existing methods, includin g those incorporating various regularization techniques and data-driven priors, struggle to generalize effectively to OOD views. To address this limitation, we introduce SplatFormer, the first point transformer model specifically designed t o operate on Gaussian splats. SplatFormer takes as input an initial 3DGS set opt imized under limited training views and refines it in a single forward pass, eff ectively removing potential artifacts in OOD test views. To our knowledge, this is the first successful application of point transformers directly on 3DGS sets, surpassing the limitations of previous multi-scene training methods, which coul d handle only a restricted number of input views during inference. Our model sig nificantly improves rendering quality under extreme novel views, achieving state -of-the-art performance in these challenging scenarios and outperforming various 3DGS regularization techniques, multi-scene models tailored for sparse view syn thesis, and diffusion-based frameworks. The project url is https://sergeyprokudi n.github.io/splatformer.
****
Ziyang Li,Saikat Dutta,Mayur Naik
IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities
Software is prone to security vulnerabilities. Program analysis tools to detect them have limited effectiveness in practice due to their reliance on human label ed specifications. Large language models (or LLMs) have shown impressive code ge neration capabilities but they cannot do complex reasoning over code to detect s uch vulnerabilities especially since this task requires whole-repository analysi s. We propose IRIS, a neuro-symbolic approach that systematically combines LLMs with static analysis to perform whole-repository reasoning for security vulnerab ility detection. Specifically, IRIS leverages LLMs to infer taint specifications and perform contextual analysis, alleviating needs for human specifications and inspection. For evaluation, we curate a new dataset, CWE-Bench-Java, comprising 120 manually validated security vulnerabilities in real-world Java projects. A state-of-the-art static analysis tool CodeQL detects only 27 of these vulnerabil ities whereas IRIS with GPT-4 detects 55 (+28) and improves upon CodeQL's averag e false discovery rate by 5% points.
Furthermore, IRIS identifies 4 previously unknown vulnerabilities which cannot b

e found by existing tools. IRIS is available publicly at https://github.com/iris-sast/iris.
****
Zhengmian Hu,Tong Zheng,Vignesh Viswanathan,Ziyi Chen,Ryan A. Rossi,Yihan Wu,Dinesh Manocha,Heng Huang
Towards Optimal Multi-draft Speculative Decoding
Large Language Models (LLMs) have become an indispensable part of natural language processing tasks. However, autoregressive sampling has become an efficiency bottleneck. Multi-Draft Speculative Decoding (MDSD) is a recent approach where, when generating each token, a small draft model generates multiple drafts, and the target LLM verifies them in parallel, ensuring that the final output conforms to the target model distribution. The two main design choices in MDSD are the draft sampling method and the verification algorithm. For a fixed draft sampling method, the optimal acceptance rate is a solution to an optimal transport problem, but the complexity of this problem makes it difficult to solve for the optimal acceptance rate and measure the gap between existing verification algorithms and the theoretical upper bound. This paper discusses the dual of the optimal transport problem, providing a way to efficiently compute the optimal acceptance rate. For the first time, we measure the theoretical upper bound of MDSD efficiency for vocabulary sizes in the thousands and quantify the gap between existing verification algorithms and this bound. We also compare different draft sampling methods based on their optimal acceptance rates. Our results show that the draft sampling method strongly influences the optimal acceptance rate, with sampling without replacement outperforming sampling with replacement. Additionally, existing verification algorithms do not reach the theoretical upper bound for both without replacement and with replacement sampling. Our findings suggest that carefully designed draft sampling methods can potentially improve the optimal acceptance rate and enable the development of verification algorithms that closely match the theoretical upper bound.
****
Joseph Fioresi,Ishan Rajendrakumar Dave,Mubarak Shah
ALBAR: Adversarial Learning approach to mitigate Biases in Action Recognition
Bias in machine learning models can lead to unfair decision making, and while it has been well-studied in the image and text domains, it remains underexplored in action recognition. Action recognition models often suffer from background bias (i.e., inferring actions based on background cues) and foreground bias (i.e., relying on subject appearance), which can be detrimental to real-life applications such as autonomous vehicles or assisted living monitoring. While prior approaches have mainly focused on mitigating background bias using specialized augmentations, we thoroughly study both foreground and background bias. We propose ALBAR, a novel adversarial training method that mitigates foreground and background biases without requiring specialized knowledge of the bias attributes. Our framework applies an adversarial cross-entropy loss to the sampled static clip (where all the frames are the same) and aims to make its class probabilities uniform using a proposed entropy maximization loss. Additionally, we introduce a gradient penalty loss for regularization against the debiasing process. We evaluate our method on established background and foreground bias protocols, setting a new state-of-the-art and strongly improving combined debiasing performance by over 12% absolute on HMDB51.
Furthermore, we identify an issue of background leakage in the existing UCF101 protocol for bias evaluation which provides a shortcut to predict actions and does not provide an accurate measure of the debiasing capability of a model. We address this issue by proposing more fine-grained segmentation boundaries for the actor, where our method also outperforms existing approaches.
****
Haoran Xu,Shuozhe Li,Harshit Sikchi,Scott Niekum,Amy Zhang
An Optimal Discriminator Weighted Imitation Perspective for Reinforcement Learning
We introduce Iterative Dual Reinforcement Learning (IDRL), a new method that takes an optimal discriminator-weighted imitation view of solving RL. Our method is

motivated by a simple experiment in which we find training a discriminator usin g the offline dataset plus an additional expert dataset and then performing disc riminator-weighted behavior cloning gives strong results on various types of dat asets. That optimal discriminator weight is quite similar to the learned visitat ion distribution ratio in Dual-RL, however, we find that current Dual-RL methods do not correctly estimate that ratio. In IDRL, we propose a correction method t o iteratively approach the optimal visitation distribution ratio in the offline dataset given no addtional expert dataset. During each iteration, IDRL removes z ero-weight suboptimal transitions using the learned ratio from the previous iter ation and runs Dual-RL on the remaining subdataset. This can be seen as replacin g the behavior visitation distribution with the optimized visitation distributio n from the previous iteration, which theoretically gives a curriculum of improve d visitation distribution ratios that are closer to the optimal discriminator we ight. We verify the effectiveness of IDRL on various kinds of offline datasets, including D4RL datasets and more realistic corrupted demonstrations. IDRL beats strong Primal-RL and Dual-RL baselines in terms of both performance and stabilit y, on all datasets.
****

Tsung-Han Wu,Giscard Biamby,Jerome Quenum,Ritwik Gupta,Joseph E. Gonzalez,Trevor Darrell,David Chan

Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Benchmark

Large Multimodal Models (LMMs) have made significant strides in visual question-answering for single images. Recent advancements like long-context LMMs have all owed them to ingest larger, or even multiple, images. However, the ability to pr ocess a large number of visual tokens does not guarantee effective retrieval and reasoning for multi-image question answering (MIQA), especially in real-world a pplications like photo album searches or satellite imagery analysis. In this wor k, we first assess the limitations of current benchmarks for long-context LMMs. We address these limitations by introducing a new vision-centric, long-context b enchmark, "Visual Haystacks (VHs)". We comprehensively evaluate both open-source and proprietary models on VHs, and demonstrate that these models struggle when reasoning across potentially unrelated images, perform poorly on cross-image rea soning, as well as exhibit biases based on the placement of key information with in the context window. Towards a solution, we introduce MIRAGE (Multi-Image Retr ieval Augmented Generation), an open-source, lightweight visual-RAG framework th at processes up to 10k images on a single 40G A100 GPU—far surpassing the 1k-ima ge limit of contemporary models. MIRAGE demonstrates up to 13% performance impro vement over existing open-source LMMs on VHs, sets a new state-of-the-art on the RetVQA multi-image QA benchmark, and achieves competitive performance on single -image QA with state-of-the-art LMMs. Our dataset, model, and code are available at: https://visual-haystacks.github.io.
****

Yannick Metz,Andras Geiszl,Raphaël Baur,Mennatallah El-Assady

Reward Learning from Multiple Feedback Types

Learning rewards from preference feedback has become an important tool in the al ignment of agentic models. Preference-based feedback, often implemented as a bin ary comparison between multiple completions, is an established method to acquire large-scale human feedback. However, human feedback in other contexts is often much more diverse. Such diverse feedback can better support the goals of a human annotator, and the simultaneous use of multiple sources might be mutually infor mative for the learning process or carry type-dependent biases for the reward le arning process.

Despite these potential benefits, learning from different feedback types has yet to be explored extensively.

In this paper, we bridge this gap by enabling experimentation and evaluating mul ti-type feedback in a wide set of environments. We present a process to generate high-quality simulated feedback of six different types. Then, we implement rewa rd models and downstream RL training for all six feedback types.

Based on the simulated feedback, we investigate the use of types of feedback acr oss ten RL environments and compare them to pure preference-based baselines. We

show empirically that diverse types of feedback can be utilized and lead to stro
ng reward modeling performance. This work is the first strong indicator of the p
otential of multi-type feedback for RLHF.
****

Yuzi Yan,Yibo Miao,Jialian Li,YipinZhang,Jian Xie,Zhijie Deng,Dong Yan
3D-Properties: Identifying Challenges in DPO and Charting a Path Forward
Aligning large language models (LLMs) with human preferences has gained signific
ant attention, with Proximal Policy Optimization (PPO) as a standard yet computa
tionally expensive method and Direct Preference Optimization (DPO) as a more eff
icient alternative. While DPO offers simplicity, it remains underutilized in sta
te-of-the-art LLMs, suggesting potential limitations. In this work, we revisit D
PO, analyzing its theoretical foundations and empirical performance to bridge th
is gap. We identify three key properties—termed \textbf{3D}-properties—that emer
ge from DPO’s learning process: \textbf{D}rastic drop in rejected response likel
ihood, \textbf{D}egradation into response suppression, and \textbf{D}ispersion e
ffect on unseen responses. We show that these issues arise from DPO’s optimizati
on dynamics, where the interaction between chosen and rejected response gradient
s leads to instability. Our findings are supported by experiments on both a cont
rolled toy model and real-world LLM tasks, including mathematical problem-solvin
g and instruction following. To address these challenges, we propose simple regu
larization techniques that improve training stability and performance. Additiona
lly, we examine how preference data distribution impacts DPO’s effectiveness, of
fering insights into how alignment models handle out-of-domain (OOD) data. Our w
ork connects these observations to broader research and provides a theoretical e
xplanation for DPO’s limitations. We hope these insights will guide future advan
cements in reward-model-free preference learning, bringing it closer to reward-m
odel-based approaches.
****

Kangrui Du,Yuhang Wu,Shikuang Deng,Shi Gu
Temporal Flexibility in Spiking Neural Networks: Towards Generalization Across T
ime Steps and Deployment Friendliness
Spiking Neural Networks (SNNs), models inspired by neural mechanisms in the brai
n, allow for energy-efficient implementation on neuromorphic hardware. However,
SNNs trained with current direct training approaches are constrained to a specif
ic time step. This "temporal inflexibility" 1) hinders SNNs' deployment on time-
step-free fully event-driven chips and 2) prevents energy-performance balance ba
sed on dynamic inference time steps. In this study, we first explore the feasibi
lity of training SNNs that generalize across different time steps. We then intro
duce Mixed Time-step Training (MTT), a novel method that improves the temporal f
lexibility of SNNs, making SNNs adaptive to diverse temporal structures. During
each iteration of MTT, random time steps are assigned to different SNN stages, w
ith spikes transmitted between stages via communication modules. After training,
 the weights are deployed and evaluated on both time-stepped and fully event-dri
ven platforms. Experimental results show that models trained by MTT gain remarka
ble temporal flexibility, friendliness for both event-driven and clock-driven de
ployment (nearly lossless on N-MNIST and 10.1\% higher than standard methods on
CIFAR10-DVS), enhanced network generalization, and near SOTA performance. To the
 best of our knowledge, this is the first work to report the results of large-sc
ale SNN deployment on fully event-driven scenarios.
****

Yuchen Lin,Chenguo Lin,Jianjin Xu,Yadong MU
OmniPhysGS: 3D Constitutive Gaussians for General Physics-Based Dynamics Generat
ion
Recently, significant advancements have been made in the reconstruction and gene
ration of 3D assets, including static cases and those with physical interactions
. To recover the physical properties of 3D assets, existing methods typically as
sume that all materials belong to a specific predefined category (e.g., elastici
ty). However, such assumptions ignore the complex composition of multiple hetero
geneous objects in real scenarios and tend to render less physically plausible a
nimation given a wider range of objects. We propose OmniPhysGS for synthesizing

a physics-based 3D dynamic scene composed of more general objects. A key design of OmniPhysGS is treating each 3D asset as a collection of constitutive 3D Gaussians. For each Gaussian, its physical material is represented by an ensemble of 12 physical domain-expert sub-models (rubber, metal, honey, water, etc.), which greatly enhances the flexibility of the proposed model. In the implementation, we define a scene by user-specified prompts and supervise the estimation of material weighting factors via a pretrained video diffusion model. Comprehensive experiments demonstrate that OmniPhysGS achieves more general and realistic physical dynamics across a broader spectrum of materials, including elastic, viscoelastic, plastic, and fluid substances, as well as interactions between different materials. Our method surpasses existing methods by approximately 3% to 16% in metrics of visual quality and text alignment.
****
Zihao Wang,Bin CUI,Shaoduo Gan
SqueezeAttention: 2D Management of KV-Cache in LLM Inference via Layer-wise Optimal Budget
Optimizing the Key-Value (KV) cache of the Large Language Model (LLM) has been considered critical to saving the cost of inference. Most of the existing KV-cache compression algorithms attempted to sparsify the sequence of tokens by taking advantage of the different importance of tokens. However, most of these methods treat all layers equally, allocating the same KV budget to each layer. This approach is suboptimal, as some layers may be less sensitive to input tokens yet still receive the same budget as others. In this work, we found that by identifying the importance of attention layers, we could optimize the KV-cache jointly from two dimensions, i.e., sequence-wise and layer-wise. Based on our observations regarding layer-wise importance in inference, we propose \sys to precisely optimize the allocation of KV-cache budget among layers on-the-fly and then incorporate three representative sequence-wise algorithms to compress the KV-cache for each layer with its very own budget. Specifically, we first measure each layer's importance by calculating the cosine similarity of the input prompt differences before and after the self-attention layers. Based on this similarity, we then categorize the layers into two groups and adjust their KV budgets accordingly. By optimizing the KV-cache from both sequence's and layer's dimensions, \sys achieves around 30\% to 70\% of the memory reductions and up to 2.2 $\times$ of throughput improvements in a wide range of LLMs and benchmarks. The code is available at https://github.com/hetailang/SqueezeAttention.
****
Tiago Silva,Rodrigo Barreto Alves,Eliezer de Souza da Silva,Amauri H Souza,Vikas Garg,Samuel Kaski,Diego Mesquita
When do GFlowNets learn the right distribution?
Generative Flow Networks (GFlowNets) are an emerging class of sampling methods for distributions over discrete and compositional objects, e.g., graphs. In spite of their remarkable success in problems such as drug discovery and phylogenetic inference, the question of when and whether GFlowNets learn to sample from the target distribution remains underexplored. To tackle this issue, we first assess the extent to which a violation of the detailed balance of the underlying flow network might hamper the correctness of GFlowNet's sampling distribution. In particular, we demonstrate that the impact of an imbalanced edge on the model's accuracy is influenced by the total amount of flow passing through it and, as a consequence, is unevenly distributed across the network. We also argue that, depending on the parameterization, imbalance may be inevitable. In this regard, we consider the problem of sampling from distributions over graphs with GFlowNets parameterized by graph neural networks (GNNs) and show that the representation limits of GNNs delineate which distributions these GFlowNets can approximate. Lastly, we address these limitations by proposing a theoretically sound and computationally tractable metric for assessing GFlowNets, experimentally showing it is a better proxy for correctness than popular evaluation protocols.
****
Defa Zhu,Hongzhi Huang,Zihao Huang,Yutao Zeng,Yunyao Mao,Banggu Wu,Qiyang Min,Xun Zhou

Hyper-Connections
We present hyper-connections, a simple yet effective method that can serve as an alternative to residual connections. This approach specifically addresses common drawbacks observed in residual connection variants, such as the seesaw effect between gradient vanishing and representation collapse. Theoretically, hyper-connections allow the network to adjust the strength of connections between features at different depths and dynamically rearrange layers. We conduct experiments focusing on the pre-training of large language models, including dense and sparse models, where hyper-connections show significant performance improvements over residual connections. Additional experiments conducted on vision tasks also demonstrate similar improvements. We anticipate that this method will be broadly applicable and beneficial across a wide range of AI problems.
****
Kanishk Bhatia,Felix Koehler,Nils Thuerey
PRDP: Progressively Refined Differentiable Physics
The physics solvers employed for neural network training are primarily iterative, and hence, differentiating through them introduces a severe computational burden as iterations grow large. Inspired by works in bilevel optimization, we show that full accuracy of the network is achievable through physics significantly coarser than fully converged solvers. We propose *progressively refined differentiable physics* (PRDP), an approach that identifies the level of physics refinement sufficient for full training accuracy. By beginning with coarse physics, adaptively refining it during training, and stopping refinement at the level adequate for training, it enables significant compute savings without sacrificing network accuracy. Our focus is on differentiating iterative linear solvers for sparsely discretized differential operators, which are fundamental to scientific computing. PRDP is applicable to both unrolled and implicit differentiation. We validate its performance on a variety of learning scenarios involving differentiable physics solvers such as inverse problems, autoregressive neural emulators, and correction-based neural-hybrid solvers. In the challenging example of emulating the Navier-Stokes equations, we reduce training time by 62%.
****
Arduin Findeis,Timo Kaufmann,Eyke Hüllermeier,Samuel Albanie,Robert D. Mullins
Inverse Constitutional AI: Compressing Preferences into Principles
Feedback data is widely used for fine-tuning and evaluating state-of-the-art AI models. Pairwise text preferences, where human or AI annotators select the “better” of two options, are particularly common. Such preferences are used to train (reward) models or to rank models with aggregate statistics. For many applications it is desirable to understand annotator preferences in addition to modelling them■■– not least because extensive prior work has shown various unintended biases in preference datasets. Yet, preference datasets remain challenging to interpret. Neither black-box reward models nor statistics can answer why one text is preferred over another. Manual interpretation of the numerous (long) response pairs is usually equally infeasible. In this paper, we introduce the Inverse Constitutional AI (ICAI) problem, formulating the interpretation of pairwise text preference data as a compression task. In constitutional AI, a set of principles (a constitution) is used to provide feedback and fine-tune AI models. ICAI inverts this process: given a feedback dataset, we aim to extract a constitution that best enables a large language model (LLM) to reconstruct the original annotations. We propose a corresponding ICAI algorithm and validate its generated constitutions quantitatively based on annotation reconstruction accuracy on several datasets: (a) synthetic feedback data with known principles; (b) AlpacaEval cross-annotated human feedback data; (c) crowdsourced Chatbot Arena data; and (d) PRISM data from diverse demographic groups. As an example application, we further demonstrate the detection of biases in human feedback data. As a short and interpretable representation of the original dataset, generated constitutions have many potential use cases: they may help identify undesirable annotator biases, better understand model performance, scale feedback to unseen data, or assist with adapting AI models to individual user or group preferences. We release the source code for our algorithm and experiments at https://github.com/rdnfn/icai.

****
MingYu Lu,Chris Lin,Chanwoo Kim,Su-In Lee
An Efficient Framework for Crediting Data Contributors of Diffusion Models
As diffusion models are deployed in real-world settings and their performance dr
iven by training data, appraising the contribution of data contributors is cruci
al to creating incentives for sharing quality data and to implementing policies
for data compensation. Depending on the use case, model performance corresponds
to various global properties of the distribution learned by a diffusion model (e
.g., overall aesthetic quality). Hence, here we address the problem of attributi
ng global properties of diffusion models to data contributors. The Shapley value
 provides a principled approach to valuation by uniquely satisfying game-theoret
ic axioms of fairness. However, estimating Shapley values for diffusion models i
s computationally impractical because it requires retraining and rerunning infer
ence on many subsets of data contributors. We introduce a method to efficiently
retrain and rerun inference for Shapley value estimation, by leveraging model pr
uning and fine-tuning. We evaluate the utility of our method with three use case
s: (i) image quality for a DDPM trained on a CIFAR dataset, (ii) demographic div
ersity for an LDM trained on CelebA-HQ, and (iii) aesthetic quality for a Stable
 Diffusion model LoRA-finetuned on Post-Impressionist artworks. Our results empi
rically demonstrate that our framework can identify important data contributors
across global properties, outperforming existing attribution methods for diffusi
on models.
****
Yuma Ichikawa,Yamato Arai
Optimization by Parallel Quasi-Quantum Annealing with Gradient-Based Sampling
Learning-based methods have gained attention as general-purpose solvers due to t
heir ability to automatically learn problem-specific heuristics, reducing the ne
ed for manually crafted heuristics. However, these methods often face scalabilit
y challenges. To address these issues, the improved Sampling algorithm for Combi
natorial Optimization (iSCO), using discrete Langevin dynamics, has been propose
d, demonstrating better performance than several learning-based solvers. This st
udy proposes a different approach that integrates gradient-based update through
continuous relaxation, combined with Quasi-Quantum Annealing (QQA). QQA smoothly
 transitions the objective function, starting from a simple convex function, min
imized at half-integral values, to the original objective function, where the re
laxed variables are minimized only in the discrete space. Furthermore, we incorp
orate parallel run communication leveraging GPUs to enhance exploration capabili
ties and accelerate convergence. Numerical experiments demonstrate that our meth
od is a competitive general-purpose solver, achieving performance comparable to
iSCO and learning-based solvers across various benchmark problems. Notably, our
method exhibits superior speed-quality trade-offs for large-scale instances comp
ared to iSCO, learning-based solvers, commercial solvers, and specialized algori
thms.
****
Hritik Bansal,Zongyu Lin,Tianyi Xie,Zeshun Zong,Michal Yarom,Yonatan Bitton,Chen
fanfu Jiang,Yizhou Sun,Kai-Wei Chang,Aditya Grover
VideoPhy: Evaluating Physical Commonsense for Video Generation
Recent advances in internet-scale video data pretraining have led to the develop
ment of text-to-video generative models that can create high-quality videos acro
ss a broad range of visual concepts, synthesize realistic motions and render com
plex objects. Hence, these generative models have the potential to become genera
l-purpose simulators of the physical world. However, it is unclear how far we ar
e from this goal with the existing text-to-video generative models. To this end,
 we present VideoPhy, a benchmark designed to assess whether the generated video
s follow physical commonsense for real-world activities (e.g. marbles will roll
down when placed on a slanted surface). Specifically, we curate diverse prompts
that involve interactions between various material types in the physical world (
e.g., solid-solid, solid-fluid, fluid-fluid). We then generate videos conditione
d on these captions from diverse state-of-the-art text-to-video generative model
s, including open models (e.g., CogVideoX) and closed models (e.g., Lumiere, Dre

am Machine). Our human evaluation reveals that the existing models severely lack the ability to generate videos adhering to the given text prompts, while also lack physical commonsense. Specifically, the best performing model, CogVideoX-5B, generates videos that adhere to the caption and physical laws for 39.6% of the instances. VideoPhy thus highlights that the video generative models are far from accurately simulating the physical world. Finally, we propose an auto-evaluator, VideoCon-Physics, to assess the performance reliably for the newly released models. The code is available here: https://github.com/Hritikbansal/videophy.
****

Tongtian Yue,Longteng Guo,Jie Cheng,Xuange Gao,Hua Huang,Jing Liu
Ada-K Routing: Boosting the Efficiency of MoE-based LLMs
In the era of Large Language Models (LLMs), Mixture-of-Experts (MoE) architectures offer a promising approach to managing computational costs while scaling up model parameters. Conventional MoE-based LLMs typically employ static Top-K routing, which activates a fixed and equal number of experts for each token regardless of their significance within the context. In this paper, we propose a novel Ada-K routing strategy that dynamically adjusts the number of activated experts for each token, thereby improving the balance between computational efficiency and model performance. Specifically, our strategy incorporates learnable and lightweight allocator modules that decide customized expert resource allocation tailored to the contextual needs for each token. These allocators are designed to be fully pluggable, making it broadly applicable across all mainstream MoE-based LLMs. We leverage the Proximal Policy Optimization (PPO) algorithm to facilitate an end-to-end learning process for this non-differentiable decision-making framework. Extensive evaluations on four popular baseline models demonstrate that our Ada-K routing method significantly outperforms conventional Top-K routing. Compared to Top-K, our method achieves over 25% reduction in FLOPs and more than 20% inference speedup while still improving performance across various benchmarks. Moreover, the training of Ada-K is highly efficient. Even for Mixtral-8x22B, a MoE-based LLM with more than 140B parameters, the training time is limited to 8 hours. Detailed analysis shows that harder tasks, middle layers, and content words tend to activate more experts, providing valuable insights for future adaptive MoE system designs. Both the training code and model checkpoints will be publicly available.
****

Yiming Chen,yuan zhang,Liyuan Cao,Kun Yuan,Zaiwen Wen
Enhancing Zeroth-order Fine-tuning for Language Models with Low-rank Structures
Parameter-efficient fine-tuning (PEFT) significantly reduces memory costs when adapting large language models (LLMs) for downstream applications. However, traditional first-order (FO) fine-tuning algorithms incur substantial memory overhead due to the need to store activation values for back-propagation during gradient computation, particularly in long-context fine-tuning tasks. Zeroth-order (ZO) algorithms offer a promising alternative by approximating gradients using finite differences of function values, thus eliminating the need for activation storage. Nevertheless, existing ZO methods struggle to capture the low-rank gradient structure common in LLM fine-tuning, leading to suboptimal performance. This paper proposes a low-rank ZO gradient estimator and introduces a novel **lo**w-rank **ZO** algorithm (LOZO) that effectively captures this structure in LLMs. We provide convergence guarantees for LOZO by framing it as a subspace optimization method. Additionally, its low-rank nature enables LOZO to integrate with momentum techniques while incurring negligible extra memory costs. Extensive experiments across various model sizes and downstream tasks demonstrate that LOZO and its momentum-based variant outperform existing ZO methods and closely approach the performance of FO algorithms.
****

Yuanpei Liu,Kai Han
DebGCD: Debiased Learning with Distribution Guidance for Generalized Category Discovery
In this paper, we tackle the problem of Generalized Category Discovery (GCD). Given a dataset containing both labelled and unlabelled images, the objective is

to categorize all images in the unlabelled subset, irrespective of whether they are from known or unknown classes.
In GCD, an inherent label bias exists between known and unknown classes due to the lack of ground-truth labels for the latter.
State-of-the-art methods in GCD leverage parametric classifiers trained through self-distillation with soft labels, leaving the bias issue unattended.
Besides, they treat all unlabelled samples uniformly, neglecting variations in certainty levels and resulting in suboptimal learning.
Moreover, the explicit identification of semantic distribution shifts between known and unknown classes, a vital aspect for effective GCD, has been neglected.
To address these challenges, we introduce DebGCD, a Debiased learning with distribution guidance framework for GCD.
Initially, DebGCD co-trains an auxiliary debiased classifier in the same feature space as the GCD classifier, progressively enhancing the GCD features. Moreover, we introduce a semantic distribution detector in a separate feature space to implicitly boost the learning efficacy of GCD. Additionally, we employ a curriculum learning strategy based on semantic distribution certainty to steer the debiased learning at an optimized pace.
Thorough evaluations on GCD benchmarks demonstrate the consistent state-of-the-art performance of our framework, highlighting its superiority. Project page: https://visual-ai.github.io/debgcd/
****
Doohyuk Jang,Sihwan Park,June Yong Yang,Yeonsung Jung,Jihun Yun,Souvik Kundu,Sung-Yub Kim,Eunho Yang
LANTERN: Accelerating Visual Autoregressive Models with Relaxed Speculative Decoding
Auto-Regressive (AR) models have recently gained prominence in image generation, often matching or even surpassing the performance of diffusion models. However, one major limitation of AR models is their sequential nature, which processes tokens one at a time, slowing down generation compared to models like GANs or diffusion-based methods that operate more efficiently. While speculative decoding has proven effective for accelerating LLMs by generating multiple tokens in a single forward, its application in visual AR models remains largely unexplored. In this work, we identify a challenge in this setting, which we term \textit{token selection ambiguity}, wherein visual AR models frequently assign uniformly low probabilities to tokens, hampering the performance of speculative decoding. To overcome this challenge, we propose a relaxed acceptance condition referred to as LANTERN that leverages the interchangeability of tokens in latent space. This relaxation restores the effectiveness of speculative decoding in visual AR models by enabling more flexible use of candidate tokens that would otherwise be prematurely rejected. Furthermore, by incorporating a total variation distance bound, we ensure that these speed gains are achieved without significantly compromising image quality or semantic coherence. Experimental results demonstrate the efficacy of our method in providing a substantial speed-up over speculative decoding. In specific, compared to a na\"ive application of the state-of-the-art speculative decoding, LANTERN increases speed-ups by $\mathbf{1.75}\times$ and $\mathbf{1.82}\times$, as compared to greedy decoding and random sampling, respectively, when applied to LlamaGen, a contemporary visual AR model. The code is publicly available at \url{https://github.com/jadohu/LANTERN}.
****
Bingrui Li,Wei Huang,Andi Han,Zhanpeng Zhou,Taiji Suzuki,Jun Zhu,Jianfei Chen
On the Optimization and Generalization of Two-layer Transformers with Sign Gradient Descent
The Adam optimizer is widely used for transformer optimization in practice, which makes understanding the underlying optimization mechanisms an important problem.
However, due to the Adam's complexity, theoretical analysis of how it optimizes transformers remains a challenging task.
Fortunately, Sign Gradient Descent (SignGD) serves as an effective surrogate for Adam.

Despite its simplicity, theoretical understanding of how SignGD optimizes transformers still lags behind.
In this work, we study how SignGD optimizes a two-layer transformer -- consisting of a softmax attention layer with trainable query-key parameterization followed by a linear layer -- on
a linearly separable noisy dataset.
We identify four stages in the training dynamics, each exhibiting intriguing behaviors.
Based on the training dynamics, we prove the fast convergence but poor generalization of the learned transformer on the noisy dataset.
We also show that Adam behaves similarly to SignGD in terms of both optimization and generalization in this setting.
Additionally, we find that the poor generalization of SignGD is not solely due to data noise,
suggesting that both SignGD and Adam requires high-quality data for real-world tasks.
Finally, experiments on synthetic and real-world datasets empirically support our theoretical results.
****
Kai Gan,Bo Ye,Min-Ling Zhang,Tong Wei
Semi-Supervised CLIP Adaptation by Enforcing Semantic and Trapezoidal Consistency
Vision-language pre-training models, such as CLIP, have demonstrated strong capability in rapidly adapting to downstream tasks through fine-tuning, and have been widely applied across various tasks. However, when the downstream tasks are constrained by limited image-text paired data, CLIP struggles to effectively address the domain gap between the pre-training and the target tasks. To address this limitation, we propose a novel semi-supervised CLIP training method coined SemiCLIP that leverages a small amount of image-text pairs alongside a large volume of images without text descriptions to enhance CLIP’s cross-modal alignment. To effectively utilize unlabeled images, we introduce semantic concept mining to improve task-specific visual representations by matching images with relevant concepts mined from labeled data. Leveraging matched semantic concepts, we construct learnable surrogate captions for unlabeled images and optimize a trapezoidal consistency to regulate the geometric structure of image-text pairs in the representation space. Experimental results demonstrate that our approach significantly improves the adaptability of CLIP in target tasks with limited labeled data, achieving gains ranging from 1.72\% -- 6.58\% for zero-shot classification accuracy and 2.32\% -- 3.23\% for image-text retrieval performance on standard benchmarks. The source code is available at https://github.com/Gank0078/SemiCLIP.
****
Nie Lin,Takehiko Ohkawa,Yifei Huang,Mingfang Zhang,Minjie Cai,Ming Li,Ryosuke Furuta,Yoichi Sato
SiMHand: Mining Similar Hands for Large-Scale 3D Hand Pose Pre-training
We present a framework for pre-training of 3D hand pose estimation from in-the-wild hand images sharing with similar hand characteristics, dubbed SiMHand. Pre-training with large-scale images achieves promising results in various tasks, but prior methods for 3D hand pose pre-training have not fully utilized the potential of diverse hand images accessible from in-the-wild videos. To facilitate scalable pre-training, we first prepare an extensive pool of hand images from in-the-wild videos and design our pre-training method with contrastive learning. Specifically, we collect over 2.0M hand images from recent human-centric videos, such as 100DOH and Ego4D. To extract discriminative information from these images, we focus on the similarity of hands: pairs of non-identical samples with similar hand poses. We then propose a novel contrastive learning method that embeds similar hand pairs closer in the feature space. Our method not only learns from similar samples but also adaptively weights the contrastive learning loss based on inter-sample distance, leading to additional performance gains. Our experiments demonstrate that our method outperforms conventional contrastive learning approaches that produce positive pairs solely from a single image with data augmentatio

n. We achieve significant improvements over the state-of-the-art method (PeCLR) in various datasets, with gains of 15% on FreiHand, 10% on DexYCB, and 4% on AssemblyHands. Our code is available at https://github.com/ut-vision/SiMHand.
****

Andrew Ying
Causal Identification for Complex Functional Longitudinal Studies
Real-time monitoring in modern medical research introduces functional longitudinal data, characterized by continuous-time measurements of outcomes, treatments, and confounders. This complexity leads to uncountably infinite treatment-confounder feedbacks, which traditional causal inference methodologies cannot handle. Inspired by the coarsened data framework, we adopt stochastic process theory, measure theory, and net convergence to propose a nonparametric causal identification framework. This framework generalizes classical g-computation, inverse probability weighting, and doubly robust formulas, accommodating time-varying outcomes subject to mortality and censoring for functional longitudinal data. We examine our framework through Monte Carlo simulations. Our approach addresses significant gaps in current methodologies, providing a solution for functional longitudinal data and paving the way for future estimation work in this domain.
****

Jin-Young Kim,Hyojun Go,Soonwoo Kwon,Hyun-Gyoon Kim
Denoising Task Difficulty-based Curriculum for Training Diffusion Models
Diffusion-based generative models have emerged as powerful tools in the realm of generative modeling. Despite extensive research on denoising across various timesteps and noise levels, a conflict persists regarding the relative difficulties of the denoising tasks. While various studies argue that lower timesteps present more challenging tasks, others contend that higher timesteps are more difficult. To address this conflict, our study undertakes a comprehensive examination of task difficulties, focusing on convergence behavior and changes in relative entropy between consecutive probability distributions across timesteps. Our observational study reveals that denoising at earlier timesteps poses challenges characterized by slower convergence and higher relative entropy, indicating increased task difficulty at these lower timesteps. Building on these observations, we introduce an easy-to-hard learning scheme, drawing from curriculum learning, to enhance the training process of diffusion models. By organizing timesteps or noise levels into clusters and training models with ascending orders of difficulty, we facilitate an order-aware training regime, progressing from easier to harder denoising tasks, thereby deviating from the conventional approach of training diffusion models simultaneously across all timesteps. Our approach leads to improved performance and faster convergence by leveraging benefits of curriculum learning, while maintaining orthogonality with existing improvements in diffusion training techniques. We validate these advantages through comprehensive experiments in image generation tasks, including unconditional, class-conditional, and text-to-image generation.
****

Nicholas Jiang,Anish Kachinthaya,Suzanne Petryk,Yossi Gandelsman
Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations
We investigate the internal representations of vision-language models (VLMs) to address hallucinations, a persistent challenge despite advances in model size and training. We project VLMs’ internal image representations to their language vocabulary and observe more confident output probabilities on real objects than hallucinated objects. We additionally use these output probabilities to spatially localize real objects. Building on this approach, we introduce a knowledge erasure algorithm that removes hallucinations by linearly orthogonalizing image features with respect to hallucinated object features. We show that targeted edits to a model’s latent representations can reduce hallucinations by up to 25.7% on the COCO2014 dataset while preserving performance. Our findings demonstrate how a deeper understanding of VLMs’ latent representations can enhance reliability and enable novel capabilities, such as zero-shot segmentation.
****

Haoru Tan,Sitong Wu,Wei Huang,Shizhen Zhao,XIAOJUAN QI
Data Pruning by Information Maximization
In this paper, we present InfoMax, a novel data pruning method, also known as co reset selection, designed to maximize the information content of selected sample s while minimizing redundancy. By doing so, InfoMax enhances the overall informa tiveness of the coreset. The information of individual samples is measured by im portance scores, which capture their influence or difficulty in model learning.
To quantify redundancy, we use pairwise sample similarities, based on the premis e that similar samples contribute similarly to the learning process.
We formalize the coreset selection problem as a discrete quadratic programming (DQP) task, with the objective of maximizing the total information content, repre sented as the sum of individual sample contributions minus the redundancies intr oduced by similar samples within the coreset.
To ensure practical scalability, we introduce an efficient gradient-based solver , complemented by sparsification techniques applied to the similarity matrix and dataset partitioning strategies.
This enables InfoMax to seamlessly scale to datasets with millions of samples.
Extensive experiments demonstrate the superior performance of InfoMax in various data pruning tasks, including image classification, vision-language pre-trainin g, and instruction tuning for large language models.
****
Xingbo Fu,Yinhan He,Jundong Li
Edge Prompt Tuning for Graph Neural Networks
Pre-training powerful Graph Neural Networks (GNNs) with unlabeled graph data in a self-supervised manner has emerged as a prominent technique in recent years. H owever, inevitable objective gaps often exist between pre-training and downstrea m tasks. To bridge this gap, graph prompt tuning techniques design and learn gra ph prompts by manipulating input graphs or reframing downstream tasks as pre-tra ining tasks without fine-tuning the pre-trained GNN models. While recent graph p rompt tuning methods have proven effective in adapting pre-trained GNN models fo r downstream tasks, they overlook the crucial role of edges in graph prompt desi gn, which can significantly affect the quality of graph representations for down stream tasks.
In this study, we propose EdgePrompt, a simple yet effective graph prompt tuning method from the perspective of edges. Unlike previous studies that design promp t vectors on node features, EdgePrompt manipulates input graphs by learning addi tional prompt vectors for edges and incorporates the edge prompts through messag e passing in the pre-trained GNN models to better embed graph structural informa tion for downstream tasks.
Our method is compatible with prevalent GNN architectures pre-trained under vari ous pre-training strategies and is universal for different downstream tasks.
We provide comprehensive theoretical analyses of our method regarding its capabi lity of handling node classification and graph classification as downstream task s.
Extensive experiments on ten graph datasets under four pre-training strategies d emonstrate the superiority of our proposed method against six baselines. Our cod e is available at https://github.com/xbfu/EdgePrompt.
****
Mutian He,Philip N. Garner
Joint Fine-tuning and Conversion of Pretrained Speech and Language Models toward s Linear Complexity
Architectures such as Linformer and Mamba have recently emerged as competitive l inear time replacements for transformers. However, corresponding large pretraine d models are often unavailable, especially in non-text domains. To remedy this, we present a Cross-Architecture Layerwise Distillation (CALD) approach that join tly converts a transformer model to a linear time substitute and fine-tunes it t o a target task. We also compare several means to guide the fine-tuning to optim ally retain the desired inference capability from the original model. The method s differ in their use of the target model and the trajectory of the parameters.
In a series of empirical studies on language processing, language modeling, and

speech processing, we show that CALD can effectively recover the result of the o
riginal model, and that the guiding strategy contributes to the result. Some rea
sons for the variation are suggested.
****

Ravindran Kannan,Chiranjib Bhattacharyya,Praneeth Kacham,David Woodruff
LevAttention: Time, Space and Streaming Efficient Algorithm for Heavy Attentions
A central problem related to transformers can be stated as follows: given two $n
 \times d$ matrices Q and K, and a non-negative function f, define the mat
rix A as follows: (1) apply the function f to each entry of the $n \times n$
 matrix $Q K^T$, and then (2) normalize each of the row sums of A to be equal
to 1. The matrix A can be computed in $O(n^2 d)$ time assuming f can be ap
plied to a number in constant time, but the quadratic dependence on n is prohi
bitive in applications where it corresponds to long context lengths. For a large
 class of functions f, we show how to find all the "large attention scores", i
.e., entries of A which are at least a positive value ε, in time w
ith linear dependence on n (i.e., $n \cdot \textrm{poly}(d/\varepsilon)$) for
a positive parameter $\varepsilon > 0$. Our class of functions include all funct
ions f of the form $f(x) = |x|^p$, as explored recently in transformer models.
 Using recently developed tools from randomized numerical linear algebra, we pro
ve that for any K, there is a "universal set" $U \subset [n]$ of size independ
ent of n, such that for any Q and any row i, the large attention scores $A
_{i,j}$ in row i of A all have $j \in U$. We also find U in $n \cdot \text
rm{poly}(d/\varepsilon)$ time. Notably, we
(1) make no assumptions on the data, (2) our workspace does not grow with n, a
nd (3) our algorithms can be computed in streaming and parallel settings. We emp
irically show the benefits of our scheme for vision transformers, showing how to
 train new models that use our universal set while training as well, showing tha
t our model is able to consistently select "important keys'" during training. We
 also provide theoretical motivation by formulating a planted model in which our
 efficient algorithms provably identify relevant keys for
each query.
****

Naama Rozen,Liat Bezalel,Gal Elidan,Amir Globerson,Ella Daniel
Do LLMs have Consistent Values?
Large Language Models (LLM) technology is rapidly advancing towards human- like
dialogue. Values are fundamental drivers of human behavior, yet research on the
values expressed in LLM-generated text remains limited. While prior work has beg
un to explore value ranking in LLMs, the crucial aspect of value correlation – t
he interrelationship and consistency between different values – has been largely
 un-examined. Drawing on established psychological theories of human value struc
ture, this paper investigates whether LLMs exhibit human-like value correlations
 within a single session, reflecting a coherent “persona”. Our findings reveal t
hat standard prompting methods fail to produce human-consistent value correlatio
ns. However, we demonstrate that a novel prompting strategy (referred to as "Val
ue Anchoring"), significantly improves the alignment of LLM value correlations w
ith human data. Furthermore, we analyze the mechanism by which Value Anchoring a
chieves this effect. These results not only deepen our understanding of value re
presentation in LLMs but also introduce new methodologies for evaluating consist
ency and human-likeness in LLM responses, highlighting the importance of explici
t value prompting for generating human-aligned outputs.
****

Byoungwoo Park,Hyungi Lee,Juho Lee
Amortized Control of Continuous State Space Feynman-Kac Model for Irregular Time
 Series
Many real-world datasets, such as healthcare, climate, and economics, are often
collected as irregular time series, which poses challenges for accurate modeling
. In this paper, we propose the Amortized Control of continuous State Space Mode
l (ACSSM) for continuous dynamical modeling of time series for irregular and dis
crete observations. We first present a multi-marginal Doob's h-transform to co
nstruct a continuous dynamical system conditioned on these irregular observation

s. Following this, we introduce a variational inference algorithm with a tight e
vidence lower bound (ELBO), leveraging stochastic optimal control (SOC) theory t
o approximate the intractable Doob's h-transform and simulate the conditioned
dynamics. To improve efficiency and scalability during both training and inferen
ce, ACSSM leverages auxiliary variable to flexibly parameterize the latent dynam
ics and amortized control. Additionally, it incorporates a simulation-free laten
t dynamics framework and a transformer-based data assimilation scheme, facilitat
ing parallel inference of the latent states and ELBO computation. Through empiri
cal evaluations across a variety of real-world datasets, ACSSM demonstrates supe
rior performance in tasks such as classification, regression, interpolation, and
 extrapolation, while maintaining computational efficiency.
****

Philipp Scholl,Katharina Bieker,Hillary Hauger,Gitta Kutyniok
ParFam -- (Neural Guided) Symbolic Regression via Continuous Global Optimization
The problem of symbolic regression (SR) arises in many different applications, s
uch as identifying physical laws or deriving mathematical equations describing t
he behavior of financial markets from given data. Various methods exist to addre
ss the problem of SR, often based on genetic programming. However, these methods
 are usually complicated and involve various hyperparameters. In this paper, we
present our new approach ParFam that utilizes parametric families of suitable sy
mbolic functions to translate the discrete symbolic regression problem into a co
ntinuous one, resulting in a more straightforward setup compared to current stat
e-of-the-art methods. In combination with a global optimizer, this approach resu
lts in a highly effective method to tackle the problem of SR. We theoretically a
nalyze the expressivity of ParFam and demonstrate its performance with extensive
 numerical experiments based on the common SR benchmark suit SRBench, showing th
at we achieve state-of-the-art results. Moreover, we present an extension incorp
orating a pre-trained transformer network (DL-ParFam) to guide ParFam, accelerat
ing the optimization process by up to two magnitudes. Our code and results can b
e found at https://anonymous.4open.science/r/parfam-D402.
****

Wilson Wu,Louis Jaburi,jacob drori,Jason Gross
Towards a Unified and Verified Understanding of Group-Operation Networks
A recent line of work in mechanistic interpretability has focused on reverse-eng
ineering the computation performed by neural networks trained on the binary oper
ation of finite groups. We investigate the internals of one-hidden-layer neural
networks trained on this task, revealing previously unidentified structure and p
roducing a more complete description of such models in a step towards unifying t
he explanations of previous works (Chughtai et al., 2023; Stander et al., 2024).
 Notably, these models approximate equivariance in each input argument. We verif
y that our explanation applies to a large fraction of networks trained on this t
ask by translating it into a compact proof of model performance, a quantitative
evaluation of the extent to which we faithfully and concisely explain model inte
rnals. In the main text, we focus on the symmetric group S5. For models trained
on this group, our explanation yields a guarantee of model accuracy that runs 3x
 faster than brute force and gives a >=95% accuracy bound for 45% of the models
we trained. We were unable to obtain nontrivial non-vacuous accuracy bounds usin
g only explanations from previous works.
****

Yunlong Yuan,Yuanfan Guo,Chunwei Wang,Wei Zhang,Hang Xu,Li Zhang
FreqPrior: Improving Video Diffusion Models with Frequency Filtering Gaussian No
ise
Text-driven video generation has advanced significantly due to developments in d
iffusion models. Beyond the training and sampling phases, recent studies have in
vestigated noise priors of diffusion models, as improved noise priors yield bett
er generation results. One recent approach employs the Fourier transform to mani
pulate noise, marking the initial exploration of frequency operations in this co
ntext. However, it often generates videos that lack motion dynamics and imaging
details. In this work, we provide a comprehensive theoretical analysis of the va
riance decay issue present in existing methods, contributing to the loss of deta

ils and motion dynamics. Recognizing the critical impact of noise distribution o
n generation quality, we introduce FreqPrior, a novel noise initialization strat
egy that refines noise in the frequency domain. Our method features a novel filt
ering technique designed to address different frequency signals while maintainin
g the noise prior distribution that closely approximates a standard Gaussian dis
tribution. Additionally, we propose a partial sampling process by perturbing the
 latent at an intermediate timestep while finding the noise prior, significantly
 reducing inference time without compromising quality. Extensive experiments on
VBench demonstrate that our method achieves the highest scores in both quality a
nd semantic assessments, resulting in the best overall total score. These result
s highlight the superiority of our proposed noise prior.
****
Lijia Yu,Yibo Miao,Yifan Zhu,Xiao-Shan Gao,Lijun Zhang
Generalizability of Neural Networks Minimizing Empirical Risk Based on Expressiv
e Power
The primary objective of learning methods is generalization. Classic generalizat
ion bounds, based on VC-dimension or Rademacher complexity, are uniformly applic
able to all networks in the hypothesis space. On the other hand, algorithm-depen
dent generalization bounds, like stability bounds, address more practical scenar
ios and provide generalization conditions for neural networks trained using SGD.
 However, these bounds often rely on strict assumptions, such as the NTK hypothe
sis or convexity of the empirical loss, which are typically not met by neural ne
tworks. In order to establish generalizability under less stringent assumptions,
 this paper investigates generalizability of neural networks that minimize the e
mpirical risk. A lower bound for population accuracy is established based on the
 expressiveness of these networks, which indicates that with adequately large tr
aining sample and network sizes, these networks can generalize effectively. Addi
tionally, we provide a lower bound necessary for generalization, demonstrating t
hat, for certain data distributions, the quantity of data required to ensure gen
eralization exceeds the network size needed to represent that distribution. Fina
lly, we provide theoretical insights into several phenomena in deep learning, in
cluding robust overfitting, importance of over-parameterization networks, and ef
fects of loss functions.
****
Roussel Desmond Nzoyem,David A.W. Barton,Tom Deakin
Neural Context Flows for Meta-Learning of Dynamical Systems
Neural Ordinary Differential Equations (NODEs) often struggle to adapt to new dy
namic behaviors caused by parameter changes in the underlying physical system, e
ven when these dynamics are similar to previously observed behaviors. This probl
em becomes more challenging when the changing parameters are unobserved, meaning
 their value or influence cannot be directly measured when collecting data. To a
ddress this issue, we introduce Neural Context Flow (NCF), a robust and interpre
table Meta-Learning framework that includes uncertainty estimation. NCF uses Tay
lor expansion to enable contextual self-modulation, allowing context vectors to
influence dynamics from other domains while also modulating themselves. After es
tablishing theoretical guarantees, we empirically test NCF and compare it to rel
ated adaptation methods. Our results show that NCF achieves state-of-the-art Out
-of-Distribution performance on 5 out of 6 linear and non-linear benchmark probl
ems. Through extensive experiments, we explore the flexible model architecture o
f NCF and the encoded representations within the learned context vectors. Our fi
ndings highlight the potential implications of NCF for foundational models in th
e physical sciences, offering a promising approach to improving the adaptability
 and generalization of NODEs in various scientific applications.
****
Jihye Choi,Jayaram Raghuram,Yixuan Li,Somesh Jha
CONDA: Adaptive Concept Bottleneck for Foundation Models Under Distribution Shif
ts
Advancements in foundation models (FMs) have led to a paradigm shift in machine
learning. The rich, expressive feature representations from these pre-trained, l
arge-

scale FMs are leveraged for multiple downstream tasks, usually via lightweight fine-tuning of a shallow fully-connected network following the representation. However, the non-interpretable, black-box nature of this prediction pipeline can be a challenge, especially in critical domains, such as healthcare, finance, and security.
In this paper, we explore the potential of Concept Bottleneck Models (CBMs) for transforming complex, non-interpretable foundation models into interpretable decision-making pipelines using high-level concept vectors. Specifically, we focus on the test-time deployment of such an interpretable CBM pipeline “in the wild”, where the distribution of inputs often shifts from the original training distribution.
We first identify the potential failure modes of such pipelines under different types of distribution shifts. Then we propose an adaptive concept bottleneck framework to address these failure modes, that dynamically adapts the concept-vector bank and the prediction layer based solely on unlabeled data from the target domain, without access to the source dataset. Empirical evaluations with various real-world distribution shifts show our framework produces concept-based interpretations better aligned with the test data and boosts post-deployment accuracy by up to 28%, aligning CBM performance with that of non-interpretable classification.
****
Isaac Ong,Amjad Almahairi,Vincent Wu,Wei-Lin Chiang,Tianhao Wu,Joseph E. Gonzalez,M Waleed Kadous,Ion Stoica
RouteLLM: Learning to Route LLMs from Preference Data
Large language models (LLMs) excel at a wide range of tasks, but choosing the right model often involves balancing performance and cost. Powerful models offer better results but are expensive, while smaller models are more cost-effective but less capable. To address this trade-off, we introduce a training framework for learning efficient router models that dynamically select between a stronger and weaker LLM during inference. Our framework leverages human preference data and employs data augmentation techniques to enhance performance. Evaluations on public benchmarks show that our approach can reduce costs by over 2 times without sacrificing response quality. Moreover, our routers exhibit strong generalization capabilities, maintaining performance even when routing between LLMs not included in training. This highlights the potential of our framework to deliver cost-effective, high-performance LLM solutions.
****
Dimitris Oikonomou,Nicolas Loizou
Sharpness-Aware Minimization: General Analysis and Improved Rates
Sharpness-Aware Minimization (SAM) has emerged as a powerful method for improving generalization in machine learning models by minimizing the sharpness of the loss landscape. However, despite its success, several important questions regarding the convergence properties of SAM in non-convex settings are still open, including the benefits of using normalization in the update rule, the dependence of the analysis on the restrictive bounded variance assumption, and the convergence guarantees under different sampling strategies. To address these questions, in this paper, we provide a unified analysis of SAM and its unnormalized variant (USAM) under one single flexible update rule (Unified SAM), and we present convergence results of the new algorithm under a relaxed and more natural assumption on the stochastic noise. Our analysis provides convergence guarantees for SAM under different step size selections for non-convex problems and functions that satisfy the Polyak-Lojasiewicz (PL) condition (a non-convex generalization of strongly convex functions). The proposed theory holds under the arbitrary sampling paradigm, which includes importance sampling as special case, allowing us to analyze variants of SAM that were never explicitly considered in the literature. Experiments validate the theoretical findings and further demonstrate the practical effectiveness of Unified SAM in training deep neural networks for image classific

ation tasks.
****
Minbeom Kim,Thibaut Thonet,Jos Rozen,Hwaran Lee,Kyomin Jung,Marc Dymetman
Guaranteed Generation from Large Language Models
As large language models (LLMs) are increasingly used across various application s, there is a growing need to control text generation to satisfy specific constr aints or requirements. This raises a crucial question: Is it possible to guarant ee strict constraint satisfaction in generated outputs while preserving the dist ribution of the original model as much as possible? We first define the ideal di stribution — the one closest to the original model, which also always satisfies the expressed constraint — as the ultimate goal of guaranteed generation. We the n state a fundamental limitation, namely that it is impossible to reach that goa l through autoregressive training alone. This motivates the necessity of combini ng training-time and inference-time methods to enforce such guarantees. Based on this insight, we propose GUARD, a simple yet effective approach that combines a n autoregressive proposal distribution with rejection sampling. Through GUARD's theoretical properties, we show how controlling the KL divergence between a spec ific proposal and the target ideal distribution simultaneously optimizes inferen ce speed and distributional closeness. To validate these theoretical concepts, w e conduct extensive experiments on two text generation settings with hard-to-sat isfy constraints: a lexical constraint scenario and a sentiment reversal scenari o. These experiments show that GUARD achieves perfect constraint satisfaction wh ile almost preserving the ideal distribution with highly improved inference effi ciency. GUARD provides a principled approach to enforcing strict guarantees for LLMs without compromising their generative capabilities.
****
Soyeon Kim,Yuji Roh,Geon Heo,Steven Euijong Whang
PFGuard: A Generative Framework with Privacy and Fairness Safeguards
Generative models must ensure both privacy and fairness for Trustworthy AI. Whil e these goals have been pursued separately, recent studies propose to combine ex isting privacy and fairness techniques to achieve both goals. However, naively c ombining these techniques can be insufficient due to privacy-fairness conflicts, where a sample in a minority group may be represented in ways that support fair ness, only to be suppressed for privacy. We demonstrate how these conflicts lead to adverse effects, such as privacy violations and unexpected fairness-utility tradeoffs. To mitigate these risks, we propose PFGuard, a generative framework w ith privacy and fairness safeguards, which simultaneously addresses privacy, fai rness, and utility. By using an ensemble of multiple teacher models, PFGuard bal ances privacy-fairness conflicts between fair and private training stages and ac hieves high utility based on ensemble learning. Extensive experiments show that PFGuard successfully generates synthetic data on high-dimensional data while pro viding both DP guarantees and convergence in fair generative modeling.
****
Haochen Wang,Anlin Zheng,Yucheng Zhao,Tiancai Wang,Zheng Ge,Xiangyu Zhang,Zhaoxi ang Zhang
Reconstructive Visual Instruction Tuning
This paper introduces reconstructive visual instruction tuning (ROSS), a family of Large Multimodal Models (LMMs) that exploit vision-centric supervision signal s. In contrast to conventional visual instruction tuning approaches that exclusi vely supervise text outputs, ROSS prompts LMMs to supervise visual outputs via r econstructing input images. By doing so, it capitalizes on the inherent richness and detail present within input images themselves, which are often lost in pure text supervision. However, producing meaningful feedback from natural images is challenging due to the heavy spatial redundancy of visual signals. To address t his issue, ROSS employs a denoising objective to reconstruct latent representati ons of input images, avoiding directly regressing exact raw RGB values. This int rinsic activation design inherently encourages LMMs to maintain image detail, th ereby enhancing their fine-grained comprehension capabilities and reducing hallu cinations. Empirically, ROSS consistently brings significant improvements across different visual encoders and language models. In comparison with extrinsic ass

istance state-of-the-art alternatives that aggregate multiple visual experts, ROSS delivers competitive performance with a single SigLIP visual encoder, demonstrating the efficacy of our vision-centric supervision tailored for visual outputs. The code will be made publicly available upon acceptance.
****
Zongyi Li,Shujie HU,Shujie LIU,Long Zhou,Jeongsoo Choi,Lingwei Meng,Xun Guo,Jinyu Li,Hefei Ling,Furu Wei
ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation
Text-to-video (T2V) models have recently undergone rapid and substantial advancements. Nevertheless, due to limitations in data and computational resources, achieving efficient generation of long videos with rich motion dynamics remains a significant challenge.
To generate high-quality, dynamic, and temporally consistent long videos, this paper presents ARLON, a novel framework that boosts diffusion Transformers with autoregressive (\textbf{AR}) models for long (\textbf{LON}) video generation, by integrating the coarse spatial and long-range temporal information provided by the AR model to guide the DiT model effectively.
Specifically, ARLON incorporates several key innovations:
1) A latent Vector Quantized Variational Autoencoder (VQ-VAE) compresses the input latent space of the DiT model into compact and highly quantized visual tokens, bridging the AR and DiT models and balancing the learning complexity and information density;
2) An adaptive norm-based semantic injection module integrates the coarse discrete visual units from the AR model into the DiT model, ensuring effective guidance during video generation;
3) To enhance the tolerance capability of noise introduced from the AR inference, the DiT model is trained with coarser visual latent tokens incorporated with an uncertainty sampling module.
Experimental results demonstrate that ARLON significantly outperforms the baseline OpenSora-V1.2 on eight out of eleven metrics selected from VBench, with notable improvements in dynamic degree and aesthetic quality, while delivering competitive results on the remaining three and simultaneously accelerating the generation process. In addition, ARLON achieves state-of-the-art performance in long video generation, outperforming other open-source models in this domain.
Detailed analyses of the improvements in inference efficiency are presented, alongside a practical application that demonstrates the generation of long videos using progressive text prompts. Project page: \url{http://aka.ms/arlon}.
****
Junzhe Zhu,Yuanchen Ju,Junyi Zhang,Muhan Wang,Zhecheng Yuan,Kaizhe Hu,Huazhe Xu
DenseMatcher: Learning 3D Semantic Correspondence for Category-Level Manipulation from a Single Demo
Dense 3D correspondence can enhance robotic manipulation by enabling the generalization of spatial, functional, and dynamic information from one object to an unseen counterpart. Compared to shape correspondence, semantic correspondence is more effective in generalizing across different object categories. To this end, we present DenseMatcher, a method capable of computing 3D correspondences between in-the-wild objects that share similar structures. DenseMatcher first computes vertex features by projecting multiview 2D features onto meshes and refining them with a 3D network, and subsequently finds dense correspondences with the obtained features using functional map. In addition, we craft the first 3D matching dataset that contains colored object meshes across diverse categories. We demonstrate the downstream effectiveness of DenseMatcher in (i) robotic manipulation, where it achieves cross-instance and cross-category generalization on long-horizon complex manipulation tasks from observing only one demo; (ii) zero-shot color mapping between digital assets, where appearance can be transferred between different objects with relatable geometry. More details and demonstrations can be found at https://tea-lab.github.io/DenseMatcher/.
****
Ghada Sokar,Johan Samir Obando Ceron,Aaron Courville,Hugo Larochelle,Pablo Samue

l Castro
Don't flatten, tokenize! Unlocking the key to SoftMoE's efficacy in deep RL
The use of deep neural networks in reinforcement learning (RL) often suffers fro
m performance degradation as model size increases. While soft mixtures of expert
s (SoftMoEs) have recently shown promise in mitigating this issue for online RL,
 the reasons behind their effectiveness remain largely unknown. In this work we
provide an in-depth analysis identifying the key factors driving this performanc
e gain. We discover the surprising result that tokenizing the encoder output, ra
ther than the use of multiple experts, is what is behind the efficacy of SoftMoE
s. Indeed, we demonstrate that even with an appropriately scaled single expert,
we are able to maintain the performance gains, largely thanks to tokenization.
****
Shahaf Bassan,Ron Eliav,Shlomit Gur
Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficie
nt Reasons
Minimal sufficient reasons represent a prevalent form of explanation - the sma
llest subset of input features which, when held constant at their corresponding
values, ensure that the prediction remains unchanged. Previous *post-hoc* method
s attempt to obtain such explanations but face two main limitations: (1) Obtaini
ng these subsets poses a computational challenge, leading most scalable methods
to converge towards suboptimal, less meaningful subsets; (2) These methods heavi
ly rely on sampling out-of-distribution input assignments, potentially resulting
 in counterintuitive behaviors. To tackle these limitations, we propose in this
work a self-supervised training approach, which we term *sufficient subset train
ing* (SST). Using SST, we train models to generate concise sufficient reasons fo
r their predictions as an integral part of their output. Our results indicate th
at our framework produces succinct and faithful subsets substantially more effic
iently than competing post-hoc methods while maintaining comparable predictive p
erformance.
****
Bao Duong,Sunil Gupta,Thin Nguyen
Causal Discovery via Bayesian Optimization
Existing score-based methods for directed acyclic graph (DAG) learning from obse
rvational data struggle to recover the causal graph accurately and sample-effici
ently. To overcome this, in this study, we propose DrBO (DAG recovery via Bayesi
an Optimization)—a novel DAG learning framework leveraging Bayesian optimization
 (BO) to find high-scoring DAGs. We show that, by sophisticatedly choosing the p
romising DAGs to explore, we can find higher-scoring ones much more efficiently.
 To address the scalability issues of conventional BO in DAG learning, we replac
e Gaussian Processes commonly employed in BO with dropout neural networks, train
ed in a continual manner, which allows for (i) flexibly modeling the DAG scores
without overfitting, (ii) incorporation of uncertainty into the estimated scores
, and (iii) scaling with the number of evaluations. As a result, DrBO is computa
tionally efficient and can find the accurate DAG in fewer trials and less time t
han existing state-of-the-art methods. This is demonstrated through an extensive
 set of empirical evaluations on many challenging settings with both synthetic a
nd real data. Our implementation is available at https://github.com/baosws/DrBO.
****
Zheyang Xiong,Vasilis Papageorgiou,Kangwook Lee,Dimitris Papailiopoulos
From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in L
LMs by Finetuning on Synthetic Data
Recent studies have shown that Large Language Models (LLMs) struggle to accurate
ly retrieve information and maintain reasoning capabilities when processing long
-context inputs. To address these limitations, we propose a finetuning approach
utilizing a carefully designed synthetic dataset comprising numerical key-value
retrieval tasks. Our experiments on models like GPT-3.5 Turbo and Mistral 7B dem
onstrate that finetuning LLMs on this dataset significantly improves LLMs' infor
mation retrieval and reasoning capabilities in longer-context settings. We prese
nt an analysis of the finetuned models, illustrating the transfer of skills from
 synthetic to real task evaluations (e.g., 10.5% improvement on 20 document

s MDQA at position 10 for GPT-3.5 Turbo). We also find that finetuned LLMs' pe
rformance on general benchmarks remains almost constant while LLMs finetuned on
other baseline long-context augmentation data can encourage hallucination (e.g.,
 on TriviaQA, Mistral 7B finetuned on our synthetic data cause no performance dr
op while other baseline data can cause a drop that ranges from 2.33% to $6.19
\%$). Our study highlights the potential of finetuning on synthetic data for imp
roving the performance of LLMs on longer-context tasks.
****
Yongjin Yang,Sihyeon Kim,Hojung Jung,Sangmin Bae,SangMook Kim,Se-Young Yun,Kimin
 Lee
Automated Filtering of Human Feedback Data for Aligning Text-to-Image Diffusion
Models
Fine-tuning text-to-image diffusion models with human feedback is an effective m
ethod for aligning model behavior with human intentions. However, this alignment
 process often suffers from slow convergence due to the large size and noise pre
sent in human feedback datasets. In this work, we propose FiFA, a novel automate
d data filtering algorithm designed to enhance the fine-tuning of diffusion mode
ls using human feedback datasets with direct preference optimization (DPO). Spec
ifically, our approach selects data by solving an optimization problem to maximi
ze three components: preference margin, text quality, and text diversity. The co
ncept of preference margin is used to identify samples that are highly informati
ve in addressing the noisy nature of feedback dataset, which is calculated using
 a proxy reward model. Additionally, we incorporate text quality, assessed by la
rge language models to prevent harmful contents, and consider text diversity thr
ough a k-nearest neighbor entropy estimator to improve generalization. Finally,
we integrate all these components into an optimization process, with approximati
ng the solution by assigning importance score to each data pair and selecting th
e most important ones. As a result, our method efficiently filters data automati
cally, without the need for manual intervention, and can be applied to any large
-scale dataset. Experimental results show that FiFA significantly enhances train
ing stability and achieves better performance, being preferred by humans 17% mor
e, while using less than 0.5% of the full data and thus 1% of the GPU hours comp
ared to utilizing full human feedback datasets.
****
Bo Liu,Rui Wang,Lemeng Wu,Yihao Feng,Peter Stone,qiang liu
Longhorn: State Space Models are Amortized Online Learners
The most fundamental capability of modern AI methods such as Large Language Mode
ls (LLMs) is the ability to predict the next token in a long sequence of tokens,
 known as “sequence modeling.” Although the Transformers model is the current d
ominant approach to sequence modeling, its quadratic computational cost with res
pect to sequence length is a significant drawback. State-space models (SSMs) off
er a promising alternative due to their linear decoding efficiency and high para
llelizability during training. However, existing SSMs often rely on seemingly ad
 hoc linear recurrence designs.
In this work, we explore SSM design through the lens of online learning, concept
ualizing SSMs as meta-modules for specific online learning problems. This approa
ch links SSM design to formulating precise online learning objectives, with stat
e transition rules derived from optimizing these objectives.
Based on this insight, we introduce a novel deep SSM architecture based on the i
mplicit update for optimizing an online regression objective. Our experimental r
esults show that our models outperform state-of-the-art SSMs, including the Mamb
a model, on standard sequence modeling benchmarks and language modeling tasks.
****
Shangzhe Di,Zhelun Yu,Guanghao Zhang,Haoyuan Li,TaoZhong,Hao Cheng,Bolin Li,Wang
gui He,Fangxun Shu,Hao Jiang
Streaming Video Question-Answering with In-context Video KV-Cache Retrieval
We propose ReKV, a novel training-free approach that enables efficient streaming
 video question-answering (StreamingVQA), by seamlessly integrating with existin
g Video Large Language Models (Video-LLMs). Traditional VideoQA systems struggle
 with long videos, as they must process entire videos before responding to queri

es, and repeat this process for each new question. In contrast, our approach ana
lyzes long videos in a streaming manner, allowing for prompt responses as soon a
s user queries are received. Building on a common Video-LLM, we first incorporat
e a sliding-window attention mechanism, ensuring that input frames attend to a l
imited number of preceding frames, thereby reducing computational overhead. To p
revent information loss, we store processed video key-value caches (KV-Caches) i
n RAM and disk, reloading them into GPU memory as needed. Additionally, we intro
duce a retrieval method that leverages an external retriever or the parameters w
ithin Video-LLMs to retrieve only query-relevant KV-Caches, ensuring both effici
ency and accuracy in question answering. ReKV enables the separation of video an
alyzing and question-answering across different processes and GPUs, significantl
y enhancing the efficiency of StreamingVQA. Through comprehensive experimentatio
n, we validate the efficacy and practicality of our approach, which significantl
y boosts efficiency and enhances applicability over existing VideoQA models.
****
Kareem Ahmed,Kai-Wei Chang,Guy Van den Broeck
Controllable Generation via Locally Constrained Resampling
Autoregressive models have demonstrated an unprecedented ability at modeling the
 intricacies
of natural language. However, they continue to struggle with generating complex
outputs that adhere to
logical constraints. Sampling from a fully-independent distribution subject to a
 constraint is hard. Sampling from an autoregressive distribution subject to a c
onstraint is doubly hard: We have to contend not only with the hardness of the c
onstraint but also the distribution's lack of structure. We propose a tractable
 probabilistic approach that performs Bayesian conditioning to draw samples subj
ect to a constraint. By factoring in information about the entire sequence, our
approach offers better contextual awareness during constrained generation compar
ed to current greedy approaches. Starting from a model sample, we induce a local
, factorized distribution which we can
tractably condition on the constraint. To generate samples that satisfy the cons
traint, we sample from the conditional distribution,
correct for biases in the sample weights, and resample. The resulting samples cl
osely approximate the target distribution and are guaranteed to satisfy the cons
traints. We evaluate our approach on several tasks, including LLM detoxification
 and solving Sudoku puzzles. We show that by disallowing a list of toxic express
ions our approach is able to steer the model's outputs away from toxic generatio
ns, outperforming similar approaches to detoxification. We also show that our ap
proach achieves a perfect accuracy on Sudoku, compared to less than 50% for G
PT4-o and Gemini 1.5.
****
Pengcheng Jiang,Cao Xiao,Minhao Jiang,Parminder Bhatia,Taha Kass-Hout,Jimeng Sun
,Jiawei Han
Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retriev
al
Large language models (LLMs) have demonstrated significant potential in clinical
 decision support. Yet LLMs still suffer from hallucinations and lack fine-grain
ed contextual medical knowledge, limiting their high-stake healthcare applicatio
ns such as clinical diagnosis. Traditional retrieval-augmented generation (RAG)
methods attempt to address these limitations but frequently retrieve sparse or i
rrelevant information, undermining prediction accuracy. We introduce KARE, a nov
el framework that integrates knowledge graph (KG) community-level retrieval with
 LLM reasoning to enhance healthcare predictions. KARE constructs a comprehensiv
e multi-source KG by integrating biomedical databases, clinical literature, and
LLM-generated insights, and organizes it using hierarchical graph community dete
ction and summarization for precise and contextually relevant information retrie
val. Our key innovations include: (1) a dense medical knowledge structuring appr
oach enabling accurate retrieval of relevant information; (2) a dynamic knowledg
e retrieval mechanism that enriches patient contexts with focused, multi-faceted
 medical insights; and (3) a reasoning-enhanced prediction framework that levera

ges these enriched contexts to produce both accurate and interpretable clinical predictions. Extensive experiments demonstrate that KARE outperforms leading models by up to 10.8-15.0\% on MIMIC-III and 12.6-12.7\% on MIMIC-IV for mortality and readmission predictions. In addition to its impressive prediction accuracy, our framework leverages the reasoning capabilities of LLMs, enhancing the trustworthiness of clinical predictions.
****

Minghao Guo,Bohan Wang,Kaiming He,Wojciech Matusik
TetSphere Splatting: Representing High-Quality Geometry with Lagrangian Volumetric Meshes
We introduce TetSphere Splatting, a Lagrangian geometry representation designed for high-quality 3D shape modeling. TetSphere splatting leverages an underused yet powerful geometric primitive -- volumetric tetrahedral meshes. It represents 3D shapes by deforming a collection of tetrahedral spheres, with geometric regularizations and constraints that effectively resolve common mesh issues such as irregular triangles, non-manifoldness, and floating artifacts. Experimental results on multi-view and single-view reconstruction highlight TetSphere splatting's superior mesh quality while maintaining competitive reconstruction accuracy compared to state-of-the-art methods. Additionally, TetSphere splatting demonstrates versatility by seamlessly integrating into generative modeling tasks, such as image-to-3D and text-to-3D generation.
****

Jinwoo Choi,Seung-Woo Seo
Dynamic Contrastive Skill Learning with State-Transition Based Skill Clustering and Dynamic Length Adjustment
Reinforcement learning (RL) has made significant progress in various domains, but scaling it to long-horizon tasks with complex decision-making remains challenging. Skill learning attempts to address this by abstracting actions into higher-level behaviors. However, current approaches often fail to recognize semantically similar behaviors as the same skill and use fixed skill lengths, limiting flexibility and generalization. To address this, we propose Dynamic Contrastive Skill Learning (DCSL), a novel framework that redefines skill representation and learning. DCSL introduces three key ideas: state-transition based skill definition, skill similarity function learning, and dynamic skill length adjustment. By focusing on state transitions and leveraging contrastive learning, DCSL effectively captures the semantic context of behaviors and adapts skill lengths to match the appropriate temporal extent of behaviors. Our approach enables more flexible and adaptive skill extraction, particularly in complex or noisy datasets, and demonstrates competitive performance compared to existing methods in task completion and efficiency.
****

Francesco Emanuele Stradi,Matteo Castiglioni,Alberto Marchesi,Nicola Gatti
Optimal Strong Regret and Violation in Constrained MDPs via Policy Optimization
We study online learning in constrained MDPs (CMDPs), focusing on the goal of attaining sublinear strong regret and strong cumulative constraint violation. Differently from their standard (weak) counterparts, these metrics do not allow negative terms to compensate positive ones, raising considerable additional challenges. Efroni et al. (2020) were the first to propose an algorithm with sublinear strong regret and strong violation, by exploiting linear programming. Thus, their algorithm is highly inefficient, leaving as an open problem achieving sublinear bounds by means of policy optimization methods, which are much more efficient in practice. Very recently, Muller et al. (2024) have partially addressed this problem by proposing a policy optimization method that allows to attain $\widetilde{\mathcal{O}}(T^{0.93})$ strong regret/violation. This still leaves open the question of whether optimal bounds are achievable by using an approach of this kind. We answer such a question affirmatively, by providing an efficient policy optimization algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ strong regret/violation. Our algorithm implements a primal-dual scheme that employs a state-of-the-art policy optimization approach for adversarial (unconstrained) MDPs as primal algorithm, and a UCB-like update for dual variables.

****
Lu Yi,Jie Peng,Yanping Zheng,Fengran Mo,Zhewei Wei,Yuhang Ye,Yue Zixuan,Zengfeng Huang
TGB-Seq Benchmark: Challenging Temporal GNNs with Complex Sequential Dynamics
Future link prediction is a fundamental challenge in various real-world dynamic systems. To address this, numerous temporal graph neural networks (temporal GNNs) and benchmark datasets have been developed. However, these datasets often feature excessive repeated edges and lack complex sequential dynamics, a key characteristic inherent in many real-world applications such as recommender systems and "Who-To-Follow" on social networks. This oversight has led existing methods to inadvertently downplay the importance of learning sequential dynamics, focusing primarily on predicting repeated edges.

In this study, we demonstrate that existing methods, such as GraphMixer and DyGFormer, are inherently incapable of learning simple sequential dynamics, such as "a user who has followed OpenAI and Anthropic is more likely to follow AI at Meta next." Motivated by this issue, we introduce the Temporal Graph Benchmark with Sequential Dynamics (TGB-Seq), a new benchmark carefully curated to minimize repeated edges, challenging models to learn sequential dynamics and generalize to unseen edges. TGB-Seq comprises large real-world datasets spanning diverse domains, including e-commerce interactions, movie ratings, business reviews, social networks, citation networks and web link networks. Benchmarking experiments reveal that current methods usually suffer significant performance degradation and incur substantial training costs on TGB-Seq, posing new challenges and opportunities for future research. TGB-Seq datasets, leaderboards, and example codes are available at https://tgb-seq.github.io/.
****
Sihang Li,Jin Huang,Jiaxi Zhuang,Yaorui Shi,Xiaochen Cai,Mingjun Xu,Xiang Wang,Linfeng Zhang,Guolin Ke,Hengxing Cai
SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding
Scientific literature understanding is crucial for extracting targeted information and garnering insights, thereby significantly advancing scientific discovery. Despite the remarkable success of Large Language Models (LLMs), they face challenges in scientific literature understanding, primarily due to (1) a lack of scientific knowledge and (2) unfamiliarity with specialized scientific tasks.
To develop an LLM specialized in scientific literature understanding, we propose a hybrid strategy that integrates continual pre-training (CPT) and supervised fine-tuning (SFT), to simultaneously infuse scientific domain knowledge and enhance instruction-following capabilities for domain-specific tasks.
In this process, we identify two key challenges: (1) constructing high-quality CPT corpora, and (2) generating diverse SFT instructions.
We address these challenges through a meticulous pipeline, including PDF text extraction, parsing content error correction, quality filtering, and synthetic instruction creation.
Applying this strategy, we present a suite of LLMs: SciLitLLM, specialized in scientific literature understanding.
These models demonstrate promising performance on scientific literature understanding benchmarks.
(1) We present an effective framework that integrates CPT and SFT to adapt LLMs to scientific literature understanding, which can also be easily adapted to other domains.
(2) We propose an LLM-based synthesis method to generate diverse and high-quality scientific instructions, resulting in a new instruction set -- SciLitIns -- for less-represented scientific domains.
(3) SciLitLLM achieves promising performance in scientific literature understanding benchmarks.
****
Ryien Hosseini,Filippo Simini,Venkatram Vishwanath,Rebecca Willett,Henry Hoffmann
Quality Measures for Dynamic Graph Generative Models

Deep generative models have recently achieved significant success in modeling gr
aph data, including dynamic graphs, where topology and features evolve over time
. However, unlike in vision and natural language domains, evaluating generative
models for dynamic graphs is challenging due to the difficulty of visualizing th
eir output, making quantitative metrics essential. In this work, we develop a ne
w quality metric for evaluating generative models of dynamic graphs. Current met
rics for dynamic graphs typically involve discretizing the continuous-evolution
of graphs into static snapshots and then applying conventional graph similarity
measures. This approach has several limitations: (a) it models temporally relate
d events as i.i.d. samples, failing to capture the non-uniform evolution of dyna
mic graphs; (b) it lacks a unified measure that is sensitive to both features an
d topology; (c) it fails to provide a scalar metric, requiring multiple metrics
without clear superiority; and (d) it requires explicitly instantiating each sta
tic snapshot, leading to impractical runtime demands that hinder evaluation at s
cale. We propose a novel metric based on the Johnson-Lindenstrauss lemma, applyi
ng random projections directly to dynamic graph data. This results in an express
ive, scalar, and application-agnostic measure of dynamic graph similarity that o
vercomes the limitations of traditional methods. We also provide a comprehensive
 empirical evaluation of metrics for continuous-time dynamic graphs, demonstrati
ng the effectiveness of our approach compared to existing methods. Our implement
ation is available at https://github.com/ryienh/jl-metric.
****

Jaeyeon Kim,Heeseung Yun,Gunhee Kim
ViSAGe: Video-to-Spatial Audio Generation
Spatial audio is essential for enhancing the immersiveness of audio-visual exper
iences, yet its production typically demands complex recording systems and speci
alized expertise. In this work, we address a novel problem of generating first-o
rder ambisonics, a widely used spatial audio format, directly from silent videos
. To support this task, we introduce YT-Ambigen, a dataset comprising 102K 5-sec
ond YouTube video clips paired with corresponding first-order ambisonics. We als
o propose new evaluation metrics to assess the spatial aspect of generated audio
 based on audio energy maps and saliency metrics. Furthermore, we present Video-
to-Spatial Audio Generation (ViSAGe), an end-to-end framework that generates fir
st-order ambisonics from silent video frames by leveraging CLIP visual features,
 autoregressive neural audio codec modeling with both directional and visual gui
dance. Experimental results demonstrate that ViSAGe produces plausible and coher
ent first-order ambisonics, outperforming two-stage approaches consisting of vid
eo-to-audio generation and audio spatialization. Qualitative examples further il
lustrate that ViSAGe generates temporally aligned high-quality spatial audio tha
t adapts to viewpoint changes.
****

Bin Wang,Anwesa Choudhuri,Meng Zheng,Zhongpai Gao,Benjamin Planche,Andong Deng,Q
in Liu,Terrence Chen,Ulas Bagci,Ziyan Wu
Order-aware Interactive Segmentation
Interactive segmentation aims to accurately segment target objects with minimal
user interactions. However, current methods often fail to accurately separate ta
rget objects from the background, due to a limited understanding of order, the r
elative depth between objects in a scene. To address this issue, we propose OIS:
 order-aware interactive segmentation, where we explicitly encode the relative d
epth between objects into order maps. We introduce a novel order-aware attention
, where the order maps seamlessly guide the user interactions (in the form of cl
icks) to attend to the image features. We further present an object-aware attent
ion module to incorporate a strong object-level understanding to better differen
tiate objects with similar order. Our approach allows both dense and sparse int
egration of user clicks, enhancing both accuracy and efficiency as compared to p
rior works. Experimental results demonstrate that OIS achieves state-of-the-art
 performance, improving mIoU after one click by 7.61 on the HQSeg44K dataset and
 1.32 on the DAVIS dataset as compared to the previous state-of-the-art SegNext,
 while also doubling inference speed compared to current leading methods.
****

Hongliang Chi,Qiong Wu,Zhengyi Zhou,Yao Ma
Shapley-Guided Utility Learning for Effective Graph Inference Data Valuation
Graph Neural Networks (GNNs) have demonstrated remarkable performance in various graph-based machine learning tasks, yet evaluating the importance of neighbors of testing nodes remains largely unexplored due to the challenge of assessing data importance without test labels. To address this gap, we propose Shapley-Guided Utility Learning (SGUL), a novel framework for graph inference data valuation. SGUL innovatively combines transferable data-specific and modelspecific features to approximate test accuracy without relying on ground truth labels. By incorporating Shapley values as a preprocessing step and using feature Shapley values as input, our method enables direct optimization of Shapley value prediction while reducing computational demands. SGUL overcomes key limitations of existing methods, including poor generalization to unseen test-time structures and indirect optimization. Experiments on diverse graph datasets demonstrate that SGUL consistently outperforms existing baselines in both inductive and transductive settings. SGUL offers an effective, efficient, and interpretable approach for quantifying the value of test-time neighbors.
****
Weixuan Wang,JINGYUAN YANG,Wei Peng
Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors
Large language models (LLMs) have achieved remarkable performance across many tasks, yet aligning them with desired behaviors remains challenging. Activation intervention has emerged as an effective and economical method to modify the behavior of LLMs. Despite considerable interest in this area, current intervention methods exclusively employ a fixed steering vector to modify model activations, lacking adaptability to diverse input semantics. To address this limitation, we propose Semantics-Adaptive Dynamic Intervention (SADI), a novel method that constructs a dynamic steering vector to intervene model activations at inference time. More specifically, SADI utilizes activation differences in contrastive pairs to precisely identify critical elements of an LLM (i.e., attention heads, hidden states, and neurons) for targeted intervention. During inference, SADI dynamically steers model behavior by scaling element-wise activations based on the directions of input semantics. Experimental results show that SADI outperforms established baselines by substantial margins, improving task performance without training. SADI's cost-effectiveness and generalizability across various LLM backbones and tasks highlight its potential as a versatile alignment technique. We will release the code to foster research in this area.
****
Michael Zhang,Simran Arora,Rahul Chalamala,Benjamin Frederick Spector,Alan Wu,Krithik Ramesh,Aaryan Singhal,Christopher Re
LoLCATs: On Low-Rank Linearizing of Large Language Models
Recent works show we can linearize large language models (LLMs)—swapping the quadratic attentions of popular Transformer-based LLMs with subquadratic analogs, such as linear attention—avoiding the expensive pretraining costs. However, linearizing LLMs often significantly degrades model quality, still requires training over billions of tokens, and remains limited to smaller 1.3B to 7B LLMs. We thus propose Low-rank Linear Conversion via Attention Transfer (LoLCATs), a simple two-step method that improves LLM linearizing quality with orders of magnitudes less memory and compute. We base these steps on two findings. First, we can replace an LLM's softmax attentions with closely-approximating linear attentions, simply by *training* the linear attentions to match their softmax counterparts with an output MSE loss (“attention transfer”). Then, this enables adjusting for approximation errors and recovering LLM quality simply with *low-rank* adaptation (LoRA). LoLCATs significantly improves linearizing quality, training efficiency, and scalability. We significantly reduce the linearizing quality gap and produce state-of-the-art subquadratic LLMs from Llama 3 8B and Mistral 7B v0.1, leading to 20+ points of improvement on 5-shot MMLU. Furthermore, LoLCATs does so with only 0.2% of past methods' model parameters and 0.04-0.2% of their training tokens. Finally, we apply LoLCATs to create the first linearized 70B and 405B LLMs (50$\times$ that of prior work). When compared with prior approaches under the sa

me compute budgets, LoLCATs significantly improves linearizing quality, closing the gap between linearized and original Llama 3.1 70B and 405B LLMs by 77.8\% and 78.1\% on 5-shot MMLU.
****
Haoxuan Che,Xuanhua He,Quande Liu,Cheng Jin,Hao Chen
GameGen-X: Interactive Open-world Game Video Generation
We introduce GameGen-$\mathbb{X}$, the first diffusion transformer model specifically designed for both generating and interactively controlling open-world game videos.
This model facilitates high-quality, open-domain generation by approximating various game elements, such as innovative characters, dynamic environments, complex actions, and diverse events.
Additionally, it provides interactive controllability, predicting and altering future content based on the current clip, thus allowing for gameplay simulation.
To realize this vision, we first collected and built an Open-World Video Game Dataset (OGameData) from scratch.
It is the first and largest dataset for open-world game video generation and control, which comprises over one million diverse gameplay video clips with informative captions.
GameGen-$\mathbb{X}$ undergoes a two-stage training process, consisting of pre-training and instruction tuning.
Firstly, the model was pre-trained via text-to-video generation and video continuation, enabling long-sequence open-domain game video generation with improved fidelity and coherence.
Further, to achieve interactive controllability, we designed InstructNet to incorporate game-related multi-modal control signal experts.
This allows the model to adjust latent representations based on user inputs, advancing the integration of character interaction and scene content control in video generation.
During instruction tuning, only the InstructNet is updated while the pre-trained foundation model is frozen, enabling the integration of interactive controllability without loss of diversity and quality of generated content.
GameGen-$\mathbb{X}$ contributes to advancements in open-world game design using generative models.
It demonstrates the potential of generative models to serve as auxiliary tools to traditional rendering techniques, demonstrating the potential for merging creative generation with interactive capabilities.
The project will be available at https://github.com/GameGen-X/GameGen-X.
****
Thomas Tian,Kratarth Goel
Direct Post-Training Preference Alignment for Multi-Agent Motion Generation Model Using Implicit Feedback from Pre-training Demonstrations
Recent advancements in Large Language Models (LLMs) have revolutionized motion generation models in embodied applications such as autonomous driving and robotic manipulation. While LLM-type auto-regressive motion generation models benefit from training scalability, there remains a discrepancy between their token prediction objectives and human preferences. As a result, models pre-trained solely with token-prediction objectives often generate behaviors that deviate from what humans would prefer, making post-training preference alignment crucial for producing human-preferred motions. Unfortunately, post-training alignment requires extensive preference rankings of motions generated by the pre-trained model, which are costly and time-consuming to annotate, especially in multi-agent motion generation settings. Recently, there has been growing interest in leveraging expert demonstrations previously used during pre-training to scalably generate preference data for post-training alignment. However, these methods often adopt an adversarial assumption, treating all pre-trained model-generated samples as unpreferred examples and relying solely on pre-training expert demonstrations to construct preferred examples. This adversarial approach overlooks the valuable signal provided by preference rankings among the model's own generations, ultimately redu

cing alignment effectiveness and potentially leading to misaligned behaviors. In this work, instead of treating all generated samples as equally bad, we propose a principled approach that leverages implicit preferences encoded in pre-training expert demonstrations to construct preference rankings among the pre-trained model's generations, offering more nuanced preference alignment guidance with zero human cost. We apply our approach to large-scale traffic simulation (more than 100 agents) and demonstrate its effectiveness in improving the realism of pre-trained model's generated behaviors, making a lightweight 1M motion generation model comparable to state-of-the-art large imitation-based models by relying solely on implicit feedback from pre-training demonstrations, without requiring additional post-training human preference annotations or incurring high computational costs. Furthermore, we provide an in-depth analysis of preference data scaling laws and their effects on over-optimization, offering valuable insights for future studies.
****

Trong Binh Nguyen,Duong Minh Nguyen,Jinsun Park,Viet Quoc Pham,Won-Joo Hwang
Federated Domain Generalization with Data-free On-server Matching Gradient
Domain Generalization (DG) aims to learn from multiple known source domains a model that can generalize well to unknown target domains. One of the key approaches in DG is training an encoder which generates domain-invariant representations. However, this approach is not applicable in Federated Domain Generalization (FDG), where data from various domains are distributed across different clients. In this paper, we introduce a novel approach, dubbed Federated Learning via On-server Matching Gradient (FedOMG), which can efficiently leverage domain information from distributed domains. Specifically, we utilize the local gradients as information about the distributed models to find an invariant gradient direction across all domains through gradient inner product maximization. The advantages are two-fold: 1) FedOMG can aggregate the characteristics of distributed models on the centralized server without incurring any additional communication cost, and 2) FedOMG is orthogonal to many existing FL/FDG methods, allowing for additional performance improvements by being seamlessly integrated with them. Extensive experimental evaluations on various settings demonstrate the robustness of FedOMG compared to other FL/FDG baselines. Our method outperforms recent SOTA baselines on four FL benchmark datasets (MNIST, EMNIST, CIFAR-10, and CIFAR-100), and three FDG benchmark datasets (PACS, VLCS, and OfficeHome). The reproducible code is publicly available~\footnote[1]{\url{https://github.com/skydvn/fedomg}}.
****

Ke-Yuan Shen
Learn hybrid prototypes for multivariate time series anomaly detection
In multivariate time series anomaly detection (MTSAD), reconstruction-based models reconstruct testing series with learned knowledge of only normal series and identify anomalies with higher reconstruction errors. In practice, over-generalization often occurs with unexpectedly well reconstruction of anomalies. Although memory banks are employed by reconstruction-based models to fight against over-generalization, these models are only efficient to detect point anomalies since they learn normal prototypes from time points, leaving interval anomalies and periodical anomalies to be discovered. To settle this problem, this paper propose a hybrid prototypes learning model for MTSAD based on reconstruction, named as H-PAD. First, normal prototypes are learned from different sizes of the patches for time series to discover interval anomalies. These prototypes in different sizes are integrated together to reconstruct query series so that any anomalies would be smoothed off and high reconstruction errors are produced. Furthermore, period prototypes are learned to discover periodical anomalies. One period prototype is memorized for one variable of the query series. Finally, extensive experiments on five benchmark datasets show the effectiveness of H-PAD.
****

David Glukhov,Ziwen Han,Ilia Shumailov,Vardan Papyan,Nicolas Papernot
Breach By A Thousand Leaks: Unsafe Information Leakage in 'Safe' AI Responses
Vulnerability of Frontier language models to misuse has prompted the development of safety measures like filters and alignment training seeking to ensure safety

through robustness to adversarially crafted prompts. We assert that robustness is fundamentally insufficient for ensuring safety goals due to inferential threats from dual-intent queries, with current defenses and evaluations failing to account for these risks. To quantify these risks, we introduce a new safety evaluation framework based on $\textit{impermissible information leakage}$ of model outputs and demonstrate how our proposed question-decomposition attack can extract dangerous knowledge from a censored LLM more effectively than traditional jailbreaking. Underlying our proposed evaluation method is a novel information-theoretic threat model of $\textit{inferential adversaries}$, distinguished from $\textit{security adversaries}$, such as jailbreaks, in that success involves inferring impermissible knowledge from victim outputs as opposed to forcing explicitly impermissible victim outputs. Through our information-theoretic framework, we show that ensuring safety against inferential adversaries requires defenses which bound impermissible information leakage, and, such defenses inevitably incur safety-utility trade-offs.
****
Eric J Bigelow,Ari Holtzman,Hidenori Tanaka,Tomer Ullman
Forking Paths in Neural Text Generation
Estimating uncertainty in Large Language Models (LLMs) is important for properly evaluating LLMs, and ensuring safety for users. However, prior approaches to uncertainty estimation focus on the final answer in generated text, ignoring intermediate steps that might dramatically impact the outcome. We hypothesize that there exist key forking tokens, such that re-sampling the system at those specific tokens, but not others, leads to very different outcomes. To test this empirically, we develop a novel approach to representing uncertainty dynamics across individual tokens of text generation, and applying statistical models to test our hypothesis. Our approach is highly flexible: it can be applied to any dataset and any LLM, without fine tuning or accessing model weights. We use our method to analyze LLM responses on 7 different tasks across 4 domains, spanning a wide range of typical use cases. We find many examples of forking tokens, including surprising ones such as a space character instead of a colon, suggesting that LLMs are often just a single token away from saying something very different.
****
TaiMing Lu,Tianmin Shu,Alan Yuille,Daniel Khashabi,Jieneng Chen
GenEx: Generating an Explorable World
Planning with partial observation is a central challenge in embodied AI. A majority of prior works have tackled this challenge by developing agents that physically explore their environment to update their beliefs about the world state. However, humans can imagine unseen parts of the world through a mental exploration and revise their beliefs with imagined observations. Such updated beliefs can allow them to make more informed decisions at the current step, without having to physically explore the world first. To achieve this human-like ability, we introduce the **Generative World Explorer (Genex)**, a video generation model that allows an agent to mentally explore a large-scale 3D world (e.g., urban scenes) and acquire imagined observations to update its belief. This updated belief will then help the agent to make a more informed decision at the current step. To train Genex, we create a synthetic urban scene dataset, Genex-DB. Our experimental results demonstrate that (1) Genex can generate high-quality and consistent observations during long-horizon mental exploration of large 3D scenes and (2) the beliefs updated with the generated observations can inform an existing decision-making model (e.g., an LLM agent) to make better plans.
****
Xiangbo Gao,Runsheng Xu,Jiachen Li,Ziran Wang,Zhiwen Fan,Zhengzhong Tu
STAMP: Scalable Task- And Model-agnostic Collaborative Perception
Perception is a crucial component of autonomous driving systems. However, single-agent setups often face limitations due to sensor constraints, especially under challenging conditions like severe occlusion, adverse weather, and long-range object detection. Multi-agent collaborative perception (CP) offers a promising solution that enables communication and information sharing between connected vehicles. Yet, the heterogeneity among agents—in terms of sensors, models, and tasks

—significantly hinders effective and efficient cross-agent collaboration. To address these challenges, we propose STAMP, a scalable task- and model-agnostic collaborative perception framework tailored for heterogeneous agents. STAMP utilizes lightweight adapter-reverter pairs to transform Bird's Eye View (BEV) features between agent-specific domains and a shared protocol domain, facilitating efficient feature sharing and fusion while minimizing computational overhead. Moreover, our approach enhances scalability, preserves model security, and accommodates a diverse range of agents. Extensive experiments on both simulated (OPV2V) and real-world (V2V4Real) datasets demonstrate that STAMP achieves comparable or superior accuracy to state-of-the-art models with significantly reduced computational costs. As the first-of-its-kind task- and model-agnostic collaborative perception framework, STAMP aims to advance research in scalable and secure mobility systems, bringing us closer to Level 5 autonomy. Our project page is at https://xiangbogaobarry.github.io/STAMP and the code is available at https://github.com/taco-group/STAMP.
****
Mingjie Liu,Yun-Da Tsai,Wenfei Zhou,Haoxing Ren
CraftRTL: High-quality Synthetic Data Generation for Verilog Code Models with Correct-by-Construction Non-Textual Representations and Targeted Code Repair
Despite the significant progress made in code generation with large language models, challenges persist, especially with hardware description languages such as Verilog. This paper first presents an analysis of fine-tuned LLMs on Verilog coding, with synthetic data from prior methods. We identify two main issues: difficulties in handling non-textual representations (Karnaugh maps, state-transition diagrams and waveforms) and significant variability during training with models randomly making ''minor'' mistakes. To address these limitations, we enhance data curation by creating correct-by-construction data targeting non-textual representations. Additionally, we introduce an automated framework that generates error reports from various model checkpoints and injects these errors into open-source code to create targeted code repair data. Our fine-tuned Starcoder2-15B outperforms prior state-of-the-art results by 3.8\%, 10.9\%, 6.6\% for pass@1 on VerilogEval-Machine, VerilogEval-Human, and RTLLM.
****
Jilan Xu,Yifei Huang,Baoqi Pei,Junlin Hou,Qingqiu Li,Guo Chen,Yuejie Zhang,Rui Feng,Weidi Xie
X-Gen: Ego-centric Video Prediction by Watching Exo-centric Videos
Generating videos in the first-person perspective has broad application prospects in the field of augmented reality and embodied intelligence.
In this work, we explore the cross-view video prediction task, where given an exo-centric video, the first frame of the corresponding ego-centric video, and textual instructions, the goal is to generate future frames of the ego-centric video.
Inspired by the notion that hand-object interactions (HOI) in ego-centric videos represent the primary intentions and actions of the current actor, we present X-Gen that explicitly models the hand-object dynamics for cross-view video prediction.
X-Gen consists of two stages. First, we design a cross-view HOI mask prediction model that anticipates the HOI masks in future ego-frames by modeling the spatio-temporal ego-exo correspondence.
Next, we employ a video diffusion model to predict future ego-frames using the first ego-frame and textual instructions, while incorporating the HOI masks as structural guidance to enhance prediction quality.
To facilitate training, we develop a fully automated pipeline to generate pseudo HOI masks for both ego- and exo-videos by exploiting vision foundation models.
Extensive experiments demonstrate that our proposed X-Gen achieves better prediction performance compared to previous video prediction models on the public Ego-Exo4D and H2O benchmark datasets, with the HOI masks significantly improving the generation of hands and interactive objects in the ego-centric videos.
****
Hanzhen Zhao,Xingyu Xie,Cong Fang,Zhouchen Lin

SEPARATE: A Simple Low-rank Projection for Gradient Compression in Modern Large-scale Model Training Process

Training Large Language Models (LLMs) presents a significant communication bottleneck, predominantly due to the growing scale of the gradient to communicate across multi-device clusters. However, how to mitigate communication overhead in practice remains a formidable challenge due to the weakness of the methodology of the existing compression methods, especially the neglect of the characteristics of the gradient. In this paper, we consider and demonstrate the low-rank properties of gradient and Hessian observed in LLMs training dynamic, and take advantage of such natural properties to design SEPARATE, a simple low-rank projection for gradient compression in modern large-scale model training processes. SEPARATE realizes dimensional reduction by common random Gaussian variables and an improved moving average error-feedback technique. We theoretically demonstrate that SEPARATE-based optimizers maintain the original convergence rate for SGD and Adam-Type optimizers for general non-convex objectives. Experimental results show that SEPARATE accelerates training speed by up to 2× for GPT-2-Medium pre-training, and improves performance on various benchmarks for LLAMA2-7B fine-tuning.

****

Xiaojian Lin,Hanhui Li,Yuhao Cheng,Yiqiang Yan,Xiaodan Liang

GDrag:Towards General-Purpose Interactive Editing with Anti-ambiguity Point Diffusion

Recent interactive point-based image manipulation methods have gained considerable attention for being user-friendly. However, these methods still face two types of ambiguity issues that can lead to unsatisfactory outcomes, namely, intention ambiguity which misinterprets the purposes of users, and content ambiguity where target image areas are distorted by distracting elements. To address these issues and achieve general-purpose manipulations, we propose a novel task-aware, training-free framework called GDrag. Specifically, GDrag defines a taxonomy of atomic manipulations, which can be parameterized and combined unitedly to represent complex manipulations, thereby reducing intention ambiguity. Furthermore, GDrag introduces two strategies to mitigate content ambiguity, including an anti-ambiguity dense trajectory calculation method (ADT) and a self-adaptive motion supervision method (SMS). Given an atomic manipulation, ADT converts the sparse user-defined handle points into a dense point set by selecting their semantic and geometric neighbors, and calculates the trajectory of the point set. Unlike previous motion supervision methods relying on a single global scale for low-rank adaption, SMS jointly optimizes point-wise adaption scales and latent feature biases. These two methods allow us to model fine-grained target contexts and generate precise trajectories. As a result, GDrag consistently produces precise and appealing results in different editing tasks. Extensive experiments on the challenging DragBench dataset demonstrate that GDrag outperforms state-of-the-art methods significantly. The code of GDrag will be released upon acceptance.

****

Egor Zverev,Sahar Abdelnabi,Soroush Tabesh,Mario Fritz,Christoph H. Lampert

Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?

Large Language Models (LLMs) show impressive results in numerous practical applications, but they lack essential safety features that are common in other areas of computer science, particularly an explicit separation of instructions and data. This makes them vulnerable to manipulations such as indirect prompt injections and generally unsuitable for safety-critical tasks. Surprisingly, there is currently no established definition or benchmark to quantify this phenomenon. In this work, we close this gap by introducing a formal measure for instruction-data separation for single-turn language models and an empirical variant that is calculable from a model’s outputs. We also present a new dataset, SEP, that allows estimating the measure for real-world models. Our results on various LLMs show that the problem of instruction-data separation is real: all models fail to achieve high separation, and canonical mitigation techniques, such as prompt engineering and fine-tuning, either fail to substantially improve separation or reduce model utility.

****

Chi-Heng Lin,Shangqian Gao,James Seale Smith,Abhishek Patel,Shikhar Tuli,Yilin Shen,Hongxia Jin,Yen-Chang Hsu
MoDeGPT: Modular Decomposition for Large Language Model Compression
Large Language Models (LLMs) have significantly advanced AI with their exceptional performance across a wide range of tasks. However, their extensive computational requirements restrict their use on devices with limited resources.
While recent compression methods based on low-rank matrices show potential solutions, they often suffer from significant loss of accuracy or introduce substantial
overhead in parameters and inference time. In this paper, we introduce Modular Decomposition (MoDeGPT), a new, efficient, and structured compression framework that overcomes these limitations. MoDeGPT jointly decomposes pairs of consecutive subcomponents within Transformer blocks, reduces hidden dimensions through output reconstruction on a larger structural scale than conventional low-rank methods, and repurposes three classical matrix decomposition algorithms—Nyström approximation, CR decomposition, and SVD—to ensure bounded errors in our novel decomposition approach. Our experiments show that MoDeGPT, without relying on backward propagation, consistently matches or surpasses the performance of prior techniques that depend on gradient information, while achieving a 98% reduction in compute costs when compressing a 13B-parameter model. On LLaMA-2/3 and OPT models, MoDeGPT retains 90-95% of zero-shot performance with compression rates of 25-30%. The compression process can be completed on a single GPU in a few hours, boosting inference throughput by up to 46%.
****
Weiwen Liu,Xu Huang,Xingshan Zeng,xinlong hao,Shuai Yu,Dexun Li,Shuai Wang,Weinan Gan,Zhengying Liu,Yuanqing Yu,Zezhong WANG,Yuxian Wang,Wu Ning,Yutai Hou,Bin Wang,Chuhan Wu,Wang Xinzhi,Yong Liu,Yasheng Wang,Duyu Tang,Dandan Tu,Lifeng Shang,Xin Jiang,Ruiming Tang,Defu Lian,Qun Liu,Enhong Chen
ToolACE: Winning the Points of LLM Function Calling
Function calling significantly extends the application boundary of large language models (LLMs), where high-quality and diverse training data is critical for unlocking this capability. However, collecting and annotating real function-calling data is challenging, while synthetic data from existing pipelines often lack coverage and accuracy. In this paper, we present ToolACE, an automatic agentic pipeline designed to generate accurate, complex, and diverse tool-learning data, specifically tailored to the capabilities of LLMs. ToolACE leverages a novel self-evolution synthesis process to curate a comprehensive API pool of 26,507 diverse APIs. Dialogs are further generated through the interplay among multiple agents, under the guidance of a complexity evaluator. To ensure data accuracy, we implement a dual-layer verification system combining rule-based and model-based checks. We demonstrate that models trained on our synthesized data---even with only 8B parameters---achieve state-of-the-art performance, comparable to the latest GPT-4 models. Our model and a subset of the data are publicly available at https://huggingface.co/Team-ACE.
****
Max Liu,Chan-Hung Yu,Wei-Hsu Lee,Cheng-Wei Hung,Yen-Chun Chen,Shao-Hua Sun
Synthesizing Programmatic Reinforcement Learning Policies with Large Language Model Guided Search
Programmatic reinforcement learning (PRL) has been explored for representing policies through programs as a means to achieve interpretability and generalization. Despite promising outcomes, current state-of-the-art PRL methods are hindered by sample inefficiency, necessitating tens of millions of program-environment interactions. To tackle this challenge, we introduce a novel LLM-guided search framework (LLM-GS). Our key insight is to leverage the programming expertise and common sense reasoning of LLMs to enhance the efficiency of assumption-free, random-guessing search methods. We address the challenge of LLMs' inability to generate precise and grammatically correct programs in domain-specific languages (DSLs) by proposing a Pythonic-DSL strategy — an LLM is instructed to initially gener

ate Python codes and then convert them into DSL programs. To further optimize th
e LLM-generated programs, we develop a search algorithm named Scheduled Hill Cli
mbing, designed to efficiently explore the programmatic search space to improve
the programs consistently. Experimental results in the Karel domain demonstrate
our LLM-GS framework's superior effectiveness and efficiency. Extensive ablation
 studies further verify the critical role of our Pythonic-DSL strategy and Sched
uled Hill Climbing algorithm. Moreover, we conduct experiments with two novel ta
sks, showing that LLM-GS enables users without programming skills and knowledge
of the domain or DSL to describe the tasks in natural language to obtain perform
ant programs.
****

Mingyuan Sun,Zheng Fang,Jiaxu Wang,Junjie Jiang,Delei Kong,Chenming Hu,Yuetong F
ANG,Renjing Xu

Optimal Brain Apoptosis

The increasing complexity and parameter count of Convolutional Neural Networks (
CNNs) and Transformers pose challenges in terms of computational efficiency and
resource demands. Pruning has been identified as an effective strategy to addres
s these challenges by removing redundant elements such as neurons, channels, or
connections, thereby enhancing computational efficiency without heavily compromi
sing performance. This paper builds on the foundational work of Optimal Brain Da
mage (OBD) by advancing the methodology of parameter importance estimation using
 the Hessian matrix. Unlike previous approaches that rely on approximations, we
introduce Optimal Brain Apoptosis (OBA), a novel pruning method that calculates
the Hessian-vector product value directly for each parameter. By decomposing the
 Hessian matrix across network layers and identifying conditions under which int
er-layer Hessian submatrices are non-zero, we propose a highly efficient techniq
ue for computing the second-order Taylor expansion of parameters. This approach
allows for a more precise pruning process, particularly in the context of CNNs a
nd Transformers, as validated in our experiments including VGG19, ResNet32, ResN
et50, and ViT-B/16 on CIFAR10, CIFAR100 and Imagenet datasets. Our code is avail
able at https://github.com/NEU-REAL/OBA.
****

Tian Yu Liu,Alessandro Achille,Matthew Trager,Aditya Golatkar,Luca Zancato,Stefa
no Soatto

PICASO: Permutation-Invariant Context Composition with State Space Models

Providing Large Language Models with relevant contextual knowledge at inference
time has been shown to greatly improve the quality of their generations. This is
 often achieved by prepending informative passages of text, or 'contexts', retri
eved from external knowledge bases to their input. However, processing additiona
l contexts online incurs significant computation costs that scale with their len
gth. State Space Models (SSMs) offer a promising solution by allowing a database
 of contexts to be mapped onto fixed-dimensional states from which to start the
generation. A key challenge arises when attempting to leverage information prese
nt across multiple contexts, since there is no straightforward way to condition
generation on multiple independent states in existing SSMs. To address this, we
leverage a simple mathematical relation derived from SSM dynamics to compose mul
tiple states into one that efficiently approximates the effect of concatenating
raw context tokens. Since the temporal ordering of contexts can often be uninfor
mative, we enforce permutation-invariance by efficiently averaging states obtain
ed via our composition algorithm across all possible context orderings. We evalu
ate our resulting method on WikiText and MSMARCO in both zero-shot and fine-tune
d settings, and show that we can match the strongest performing baseline while e
njoying on average $5.4\times$ speedup.
****

Lin Zhu,Xinbing Wang,Chenghu Zhou,Qinying Gu,Nanyang Ye

Less is More: Masking Elements in Image Condition Features Avoids Content Leakag
es in Style Transfer Diffusion Models

Given a style-reference image as the additional image condition, text-to-image d
iffusion models have demonstrated impressive capabilities in generating images t
hat possess the content of text prompts while adopting the visual style of the r

eference image. However, current state-of-the-art methods often struggle to dise
ntangle content and style from style-reference images, leading to issues such as
 content leakages. To address this issue, we propose a masking-based method that
 efficiently decouples content from style without the need of tuning any model p
arameters. By simply masking specific elements in the style reference's image fe
atures, we uncover a critical yet under-explored principle: guiding with appropr
iately-selected fewer conditions (e.g., dropping several image feature elements)
 can efficiently avoid unwanted content flowing into the diffusion models, enhan
cing the style transfer performances of text-to-image diffusion models. In this
paper, we validate this finding both theoretically and experimentally. Extensive
 experiments across various styles demonstrate the effectiveness of our masking-
based method and support our theoretical results.
****
Tianqi Liu,Wei Xiong,Jie Ren,Lichang Chen,Junru Wu,Rishabh Joshi,Yang Gao,Jiamin
g Shen,Zhen Qin,Tianhe Yu,Daniel Sohn,Anastasia Makarova,Jeremiah Zhe Liu,Yuan L
iu,Bilal Piot,Abe Ittycheriah,Aviral Kumar,Mohammad Saleh
RRM: Robust Reward Model Training Mitigates Reward Hacking
Reward models (RMs) play a pivotal role in aligning large language models (LLMs)
 with human preferences. However, traditional RM training, which relies on respo
nse pairs tied to specific prompts, struggles to disentangle prompt-driven prefe
rences from prompt-independent artifacts, such as response length and format. In
 this work, we expose a fundamental limitation of current RM training methods, w
here RMs fail to effectively distinguish between contextual signals and irreleva
nt artifacts when determining preferences. To address this, we introduce a causa
l framework that learns preferences independent of these artifacts and propose a
 novel data augmentation technique designed to eliminate them. Extensive experi
ments show that our approach successfully filters out undesirable artifacts, yie
lding a more robust reward model (RRM). Our RRM improves the performance of a pa
irwise reward model trained on Gemma-2-9b-it, on Reward-Bench, increasing accura
cy from 80.61% to 84.15%. Additionally, we train two DPO policies using both the
 RM and RRM, demonstrating that the RRM significantly enhances DPO-aligned polic
ies, improving MT-Bench scores from 7.27 to 8.31 and length-controlled win-rates
 in AlpacaEval-2 from 33.46% to 52.49%.
****
Yuto Nishimura,Takumi Hirose,Masanari Ohi,Hideki Nakayama,Nakamasa Inoue
HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-
to-Speech Synthesis
Recently, Text-to-speech (TTS) models based on large language models (LLMs)
that translate natural language text into sequences of discrete audio tokens hav
e
gained great research attention, with advances in neural audio codec (NAC) mod-
els using residual vector quantization (RVQ). However, long-form speech synthe-
sis remains a significant challenge due to the high frame rate, which increases
the
length of audio tokens and makes it difficult for autoregressive language models
to generate audio tokens for even a minute of speech. To address this challenge,
this paper introduces two novel post-training approaches: 1) Multi-Resolution Re
-
quantization (MReQ) and 2) HALL-E. MReQ is a framework to reduce the frame
rate of pre-trained NAC models. Specifically, it incorporates multi-resolution
residual vector quantization (MRVQ) module that hierarchically reorganizes dis-
crete audio tokens through teacher-student distillation. HALL-E is an LLM-based
TTS model designed to predict hierarchical tokens of MReQ. Specifically, it inco
r-
porates the technique of using MRVQ sub-modules and continues training from a
pre-trained LLM-based TTS model. Furthermore, to promote TTS research, we
create MinutesSpeech, a new benchmark dataset consisting of 40k hours of filtere
d
speech data for training and evaluating speech synthesis ranging from 3s up to
180s. In experiments, we demonstrated the effectiveness of our approaches by ap-

plying our post-training framework to VALL-E. We achieved the frame rate down
to as low as 8 Hz, enabling the stable minitue-long speech synthesis in a single
inference step. Audio samples, dataset, codes and pre-trained models are availab
le
at https://yutonishimura-v2.github.io/HALL-E_DEMO.
****
Rui Sun,Huayu Mai,Wangkai Li,Tianzhu Zhang
Towards Unbiased Learning in Semi-Supervised Semantic Segmentation
Semi-supervised semantic segmentation aims to learn from a limited amount of lab
eled data and a large volume of unlabeled data, which has witnessed impressive p
rogress with the recent advancement of deep neural networks. However, existing m
ethods tend to neglect the fact of class imbalance issues, leading to the Matthe
w effect, that is, the poorly calibrated model’s predictions can be biased towar
ds the majority classes and away from minority classes with fewer samples. In th
is work, we analyze the Matthew effect present in previous methods that hinder m
odel learning from a discriminative perspective. In light of this background, we
 integrate generative models into semi-supervised learning, taking advantage of
their better class-imbalance tolerance. To this end, we propose DiffMatch to for
mulate the semi-supervised semantic segmentation task as a conditional discrete
data generation problem to alleviate the Matthew effect of discriminative soluti
ons from a generative perspective. Plus, to further reduce the risk of overfitti
ng to the head classes and to increase coverage of the tail class distribution,
we mathematically derive a debiased adjustment to adjust the conditional reverse
 probability towards unbiased predictions during each sampling step. Extensive e
xperimental results across multiple benchmarks, especially in the most limited l
abel scenarios with the most serious class imbalance issues, demonstrate that Di
ffMatch performs favorably against state-of-the-art methods.
****
Zheyuan Zhang,Fengyuan Hu,Jayjun Lee,Freda Shi,Parisa Kordjamshidi,Joyce Chai,Zi
qiao Ma
Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of R
eference under Ambiguities
Spatial expressions in situated communication can be ambiguous, as their meaning
s vary depending on the frames of reference (FoR) adopted by speakers and listen
ers. While spatial language understanding and reasoning by vision-language model
s (VLMs) have gained increasing attention, potential ambiguities in these models
 are still under-explored. To address this issue, we present the COnsistent Mult
ilingual Frame Of Reference Test (COMFORT), an evaluation protocol to systematic
ally assess the spatial reasoning capabilities of VLMs. We evaluate nine state-o
f-the-art VLMs using COMFORT. Despite showing some alignment with English conven
tions in resolving ambiguities, our experiments reveal significant shortcomings
of VLMs: notably, the models (1) exhibit poor robustness and consistency, (2) la
ck the flexibility to accommodate multiple FoRs, and (3) fail to adhere to langu
age-specific or culture-specific conventions in cross-lingual tests, as English
tends to dominate other languages. With a growing effort to align vision-languag
e models with human cognitive intuitions, we call for more attention to the ambi
guous nature and cross-cultural diversity of spatial reasoning.
****
Austin Henry Cheng,Alston Lo,Kin Long Kelvin Lee,Santiago Miret,Alan Aspuru-Guzi
k
Stiefel Flow Matching for Moment-Constrained Structure Elucidation
Molecular structure elucidation is a fundamental step in understanding chemical
phenomena, with applications in identifying molecules in natural products, lab s
yntheses, forensic samples, and the interstellar medium.
We consider the task of predicting a molecule's all-atom 3D structure given only
 its molecular formula and moments of inertia, motivated by the ability of rotat
ional spectroscopy to measure these moments.
While existing generative models can conditionally sample 3D structures with app
roximately correct moments, this soft conditioning fails to leverage the many di
gits of precision afforded by experimental rotational spectroscopy.

To address this, we first show that the space of n-atom point clouds with a fi
xed set of moments of inertia is embedded in the Stiefel manifold $\mathrm{St}(n
, 4)$.
We then propose Stiefel Flow Matching as a generative model for elucidating 3D s
tructure under exact moment constraints.
Additionally, we learn simpler and shorter flows by finding approximate solution
s for equivariant optimal transport on the Stiefel manifold.
Empirically, enforcing exact moment constraints allows Stiefel Flow Matching to
achieve higher success rates and faster sampling than Euclidean diffusion models
, even on high-dimensional manifolds corresponding to large molecules in the GEO
M dataset.
****
Robin Staab,Mark Vero,Mislav Balunovic,Martin Vechev
Language Models are Advanced Anonymizers
Recent privacy research on large language models (LLMs) has shown that they achi
eve near-human-level performance at inferring personal data from online texts. W
ith ever-increasing model capabilities, existing text anonymization methods are
currently lacking behind regulatory requirements and adversarial threats. In thi
s work, we take two steps to bridge this gap: First, we present a new setting fo
r evaluating anonymization in the face of adversarial LLM inferences, allowing f
or a natural measurement of anonymization performance while remedying some of th
e shortcomings of previous metrics. Then, within this setting, we develop a nove
l LLM-based adversarial anonymization framework leveraging the strong inferentia
l capabilities of LLMs to inform our anonymization procedure. We conduct a compr
ehensive experimental evaluation of adversarial anonymization across 13 LLMs on
real-world and synthetic online texts, comparing it against multiple baselines a
nd industry-grade anonymizers. Our evaluation shows that adversarial anonymizati
on outperforms current commercial anonymizers both in terms of the resulting uti
lity and privacy. We support our findings with a human study (n=50) highlighting
 a strong and consistent human preference for LLM-anonymized texts.
****
Jiaxiang Tang,Zhaoshuo Li,Zekun Hao,Xian Liu,Gang Zeng,Ming-Yu Liu,Qinsheng Zhan
g
EdgeRunner: Auto-regressive Auto-encoder for Artistic Mesh Generation
Current auto-regressive mesh generation methods suffer from issues such as incom
pleteness, insufficient detail, and poor generalization.
In this paper, we propose an Auto-regressive Auto-encoder (ArAE) model capable o
f generating high-quality 3D meshes with up to 4,000 faces at a spatial resoluti
on of 512^3.
We introduce a novel mesh tokenization algorithm that efficiently compresses tri
angular meshes into 1D token sequences, significantly enhancing training efficie
ncy.
Furthermore, our model compresses variable-length triangular meshes into a fixed
-length latent space, enabling training latent diffusion models for better gener
alization.
Extensive experiments demonstrate the superior quality, diversity, and generaliz
ation capabilities of our model in both point cloud and image-conditioned mesh g
eneration tasks.
****
Jaeho Kim,Kwangryeol Park,Sukmin Yun,Seulki Lee
PPT: Patch Order Do Matters In Time Series Pretext Task
Recently, patch-based models have been widely discussed in time series analysis.
 However, existing pretext tasks for patch-based learning, such as masking, may
not capture essential time and channel-wise patch interdependencies in time seri
es data, presumed to result in subpar model performance. In this work, we introd
uce *Patch order-aware Pretext Task (PPT)*, a new self-supervised patch order le
arning pretext task for time series classification. PPT exploits the intrinsic s
equential order information among patches across time and channel dimensions of
time series data, where model training is aided by channel-wise patch permutatio
ns. The permutation disrupts patch order consistency across time and channel dim

ensions with controlled intensity to provide supervisory signals for learning time series order characteristics. To this end, we propose two patch order-aware learning methods: patch order consistency learning, which quantifies patch order correctness, and contrastive learning, which distinguishes weakly permuted patch sequences from strongly permuted ones. With patch order learning, we observe enhanced model performance, e.g., improving up to 7% accuracy for the supervised cardiogram task and outperforming mask-based learning by 5% in the self-supervised human activity recognition task. We also propose ACF-CoS, an evaluation metric that measures the *importance of orderness* for time series datasets, which enables pre-examination of the efficacy of PPT in model training.
****

Dongyang Ma,Yan Wang,Tian Lan

Block-Attention for Efficient Prefilling

We introduce Block-attention, an attention mechanism designed to address the increased inference latency and cost in Retrieval-Augmented Generation (RAG) scenarios.

Traditional approaches often encode the entire context in an auto-regressive manner.

Instead, Block-attention divides retrieved documents into discrete blocks, with each block independently calculating key-value (KV) states except for the final block.

In RAG scenarios, by defining each passage as a block, Block-attention enables us to reuse the KV states of passages that have been seen before, thereby significantly reducing the latency and the computation overhead during inference.

The implementation of Block-attention involves block segmentation, position re-encoding, and fine-tuning the LLM to adapt to the Block-attention mechanism.

Experiments on 11 diverse benchmarks, including RAG, ICL, and general domains, demonstrate that after block fine-tuning, the Block-attention model not only achieves performance comparable to that of full-attention models, but can also seamlessly switch between the block and full attention modes without any performance loss.

Notably, Block-attention significantly reduces the time to first token (TTFT) and floating point operations (FLOPs) to a very low level. It only takes 45 ms to output the first token for an input sequence with a total length of 32K. Compared to the full-attention models, the TTFT and corresponding FLOPs are reduced by 98.7\% and 99.8\%, respectively. Additionally, in Appendix A, we elaborate on how Block-attention is applied in Game AI scenario and the substantial potential benefits it entails. We strongly suggest researchers in the gaming field not to overlook this section.
****

Ian Wu,Patrick Fernandes,Amanda Bertsch,Seungone Kim,Sina Khoshfetrat Pakazad,Graham Neubig

Better Instruction-Following Through Minimum Bayes Risk

General-purpose LLM judges capable of human-level evaluation provide not only a scalable and accurate way of evaluating instruction-following LLMs but also new avenues for supervising and improving their performance. One promising way of leveraging LLM judges for supervision is through Minimum Bayes Risk (MBR) decoding, which uses a reference-based evaluator to select a high-quality output from amongst a set of candidate outputs. In the first part of this work, we explore using MBR decoding as a method for improving the test-time performance of instruction-following LLMs. We find that MBR decoding with reference-based LLM judges substantially improves over greedy decoding, best-of-N decoding with reference-free judges and MBR decoding with lexical and embedding-based metrics on AlpacaEval and MT-Bench. These gains are consistent across LLMs with up to 70B parameters, demonstrating that smaller LLM judges can be used to supervise much larger LLMs. Then, seeking to retain the improvements from MBR decoding while mitigating additional test-time costs, we explore iterative self-training on MBR-decoded outputs. We find that self-training using Direct Preference Optimisation leads to significant performance gains, such that the self-trained models with greedy decoding generally match and sometimes exceed the performance of their base models wit

h MBR decoding.
****
Nicklas Hansen,Jyothir S V,Vlad Sobal,Yann LeCun,Xiaolong Wang,Hao Su
Hierarchical World Models as Visual Whole-Body Humanoid Controllers
Whole-body control for humanoids is challenging due to the high-dimensional natu re of the problem, coupled with the inherent instability of a bipedal morphology . Learning from visual observations further exacerbates this difficulty. In this work, we explore highly data-driven approaches to visual whole-body humanoid co ntrol based on reinforcement learning, without any simplifying assumptions, rewa rd design, or skill primitives. Specifically, we propose a hierarchical world mo del in which a high-level agent generates commands based on visual observations for a low-level agent to execute, both of which are trained with rewards. Our ap proach produces highly performant control policies in 8 tasks with a simulated 5 6-DoF humanoid, while synthesizing motions that are broadly preferred by humans. Code and videos: https://www.nicklashansen.com/rlpuppeteer
****
Qingxiu Dong,Li Dong,Xingxing Zhang,Zhifang Sui,Furu Wei
Self-Boosting Large Language Models with Synthetic Preference Data
Through alignment with human preferences, Large Language Models (LLMs) have adva nced significantly in generating honest, harmless, and helpful responses. Howeve r, collecting high-quality preference data is a resource-intensive and creativit y-demanding process, especially for the continual improvement of LLMs. We introd uce SynPO, a self-boosting paradigm that leverages synthetic preference data for model alignment. SynPO employs an iterative mechanism wherein a self-prompt gen erator creates diverse prompts, and a response improver refines model responses progressively. This approach trains LLMs to autonomously learn the generative re wards for their own outputs and eliminates the need for large-scale annotation of prompts and human preferences. After four SynPO iterations, Llama3-8B and Mis tral-7B show significant enhancements in instruction-following abilities, achiev ing over 22.1% win rate improvements on AlpacaEval 2.0 and ArenaHard. Simultaneo usly, SynPO improves the general performance of LLMs on various tasks, validated by a 3.2 to 5.0 average score increase on the well-recognized Open LLM leaderbo ard.
****
Seil Kang,Jinyeong Kim,Junhyeok Kim,Seong Jae Hwang
See What You Are Told: Visual Attention Sink in Large Multimodal Models
Large multimodal models (LMMs) "see" images by leveraging the attention mechanis m between text and visual tokens in the transformer decoder. Ideally, these mode ls should focus on key visual information relevant to the text token. However, r ecent findings indicate that LMMs have an extraordinary tendency to consistently allocate high attention weights to specific visual tokens, even when these toke ns are irrelevant to the corresponding text. In this study, we investigate the p roperty behind the appearance of these irrelevant visual tokens and examine thei r characteristics. Our findings show that this behavior arises due to the massiv e activation of certain hidden state dimensions, which resembles the attention s ink found in language models. Hence, we refer to this phenomenon as the visual a ttention sink. In particular, our analysis reveals that removing the irrelevant visual sink tokens does not impact model performance, despite receiving high att ention weights. Consequently, we recycle the attention to these tokens as surplu s resources, redistributing the attention budget to enhance focus on the image. To achieve this, we introduce Visual Attention Redistribution (VAR), a method th at redistributes attention in image-centric heads, which we identify as innately focusing on visual information. VAR can be seamlessly applied across different LMMs to improve performance on a wide range of tasks, including general vision-l anguage tasks, visual hallucination tasks, and vision-centric tasks, all without the need for additional training, models, or inference steps. Experimental resu lts demonstrate that VAR enables LMMs to process visual information more effecti vely by adjusting their internal attention mechanisms, offering a new direction to enhancing the multimodal capabilities of LMMs.
****

Xianghan Meng,Zhiyuan Huang,Wei He,Xianbiao Qi,Rong Xiao,Chun-Guang Li
Exploring a Principled Framework for Deep Subspace Clustering
Subspace clustering is a classical unsupervised learning task, built on a basic assumption that high-dimensional data can be approximated by a union of subspaces (UoS). Nevertheless, the real-world data are often deviating from the UoS assumption. To address this challenge, state-of-the-art deep subspace clustering algorithms attempt to jointly learn UoS representations and self-expressive coefficients. However, the general framework of the existing algorithms suffers from feature collapse and lacks a theoretical guarantee to learn desired UoS representation. In this paper, we present a Principled fRamewOrk for Deep Subspace Clustering (PRO-DSC), which is designed to learn structured representations and self-expressive coefficients in a unified manner. Specifically, in PRO-DSC, we incorporate an effective regularization on the learned representations into the self-expressive model, prove that the regularized self-expressive model is able to prevent feature space collapse, and demonstrate that the learned optimal representations under certain condition lie on a union of orthogonal subspaces. Moreover, we provide a scalable and efficient approach to implement our PRO-DSC and conduct extensive experiments to verify our theoretical findings and demonstrate the superior performance of our proposed deep subspace clustering approach.
****
Yunlu Yan,Chun-Mei Feng,Wangmeng Zuo,Salman Khan,Lei Zhu,Yong Liu
On the Importance of Language-driven Representation Learning for Heterogeneous Federated Learning
Non-Independent and Identically Distributed (Non-IID) training data significantly challenge federated learning (FL), impairing the performance of the global model in distributed frameworks. Inspired by the superior performance and generalizability of language-driven representation learning in centralized settings, we explore its potential to enhance FL for handling non-IID data. In specific, this paper introduces FedGLCL, a novel language-driven FL framework for image-text learning that uniquely integrates global language and local image features through contrastive learning, offering a new approach to tackle non-IID data in FL. FedGLCL redefines FL by avoiding separate local training models for each client. Instead, it uses contrastive learning to harmonize local image features with global textual data, enabling uniform feature learning across different local models. The utilization of a pre-trained text encoder in FedGLCL serves a dual purpose: it not only reduces the variance in local feature representations within FL by providing a stable and rich language context but also aids in mitigating overfitting, particularly to majority classes, by leveraging broad linguistic knowledge. Extensive experiments show that FedGLCL significantly outperforms state-of-the-art FL algorithms across different non-IID scenarios.
****
Wei Lin,Jia Wan,Antoni B. Chan
Proximal Mapping Loss: Understanding Loss Functions in Crowd Counting & Localization
Crowd counting and localization involve extracting the number and distribution of crowds from images or videos using computer vision techniques. Most counting methods are based on density regression and are based on an ``intersection'' hypothesis, *i.e.*, one pixel is influenced by multiple points in the ground truth, which is inconsistent with reality since one pixel would not contain two objects. This paper proposes Proximal Mapping Loss (PML), a density regression method that eliminates this hypothesis. {PML} divides the predicted density map into multiple point-neighbor cases through the nearest neighbor, and then dynamically constructs a learning target for each sub-case via proximal mapping, leading to more robust and accurate training. {Furthermore}, PML is theoretically linked to various existing loss functions, such as Gaussian-blurred L2 loss, Bayesian loss, and the training schemes in P2PNet and DMC, demonstrating its versatility and adaptability. Experimentally, PML significantly improves the performance of crowd counting and localization, and illustrates the robustness against annotation noise. The code is available at https://github.com/Elin24/pml.

****
Souradip Chakraborty,Sujay Bhatt,Udari Madhushani Sehwag,Soumya Suvra Ghosal,Jia hao Qiu,Mengdi Wang,Dinesh Manocha,Furong Huang,Alec Koppel,Sumitra Ganesh
Collab: Controlled Decoding using Mixture of Agents for LLM Alignment
Alignment of Large Language models (LLMs) is crucial for safe and trustworthy de ployment in applications. Reinforcement learning from human feedback (RLHF) has emerged as an effective technique to align LLMs to human preferences, and broade r utilities, but it requires updating billions of model parameters which is comp utationally expensive. Controlled Decoding, by contrast, provides a mechanism fo r aligning a model at inference time without retraining. However, single-agent d ecoding approaches often struggle to adapt to diverse tasks due to the complexit y and variability inherent in these tasks. To strengthen the test-time performan ce w.r.t the target task, we propose a mixture of agents-based decoding strategi es leveraging the existing off-the-shelf aligned LLM policies. Treating each pri or policy as an agent in the spirit of mixture of agent collaboration, we develo p a decoding method that allows for inference-time alignment through a token-lev el selection strategy among multiple agents. For each token, the most suitable L LM is dynamically chosen from a pool of models based on a long-term utility metr ic. This policy-switching mechanism ensures optimal model selection at each step , enabling efficient collaboration and alignment among LLMs during decoding. The oretical analysis of our proposed algorithm establishes optimal performance with respect to the target task represented via a target reward, for the given off-t he-shelf models. We conduct comprehensive empirical evaluations with open-source aligned models on diverse tasks and preferences, which demonstrates the merits of this approach over single-agent decoding baselines. Notably, COLLAB surpasses the current SoTA decoding strategy, achieving an improvement of {up to 1.56x} i n average reward and 71.89% in GPT-4 based win-tie rate.
****
Kush Jain,Gabriel Synnaeve,Baptiste Roziere
TestGenEval: A Real World Unit Test Generation and Test Completion Benchmark
Code generation models can help improve many common software tasks ranging from code completion to defect prediction. Most of the existing benchmarks for code g eneration LLMs focus on code authoring or code completion. Surprisingly, there h as been far less effort dedicated to benchmarking software testing, despite the strong correlation between well-tested software and effective bug detection. To address this gap, we create and release TestGenEval, a large-scale benchmark to measure test generation performance. Based on SWEBench, TestGenEval comprises 68 ,647 tests from 1,210 code and test file pairs across 11 well-maintained Python repositories. It covers initial tests authoring, test suite completion, and code coverage improvements. Test authoring simulates the process of a developer writ ing a test suite from scratch, while test completion mimics the scenario where a developer aims to improve the coverage of an existing test suite. We evaluate s everal popular models, with sizes ranging from 7B to 405B parameters. Our detail ed analysis highlights TestGenEval's contribution to a comprehensive evaluation of test generation performance. In particular, models struggle to generate high-coverage test suites, with the best model, GPT-4o, achieving an average coverage of only 35.2\%. This is primarily due to models struggling to reason about exec ution, and their frequent assertion errors when addressing complex code paths.
****
Minyoung Kim,Timothy Hospedales
LiFT: Learning to Fine-Tune via Bayesian Parameter Efficient Meta Fine-Tuning
We tackle the problem of parameter-efficient fine-tuning (PEFT) of a pre-trained large deep model on many different but related tasks. Instead of the simple but strong baseline strategy of task-wise independent fine-tuning, we aim to meta-l earn the core shared information that can be used for unseen test tasks to impro ve the prediction performance further. That is, we propose a method for {\em lea rning-to-fine-tune} (LiFT). LiFT introduces a novel hierarchical Bayesian model that can be superior to both existing general meta learning algorithms like MAML and recent LoRA zoo mixing approaches such as LoRA-Retriever and model-based cl ustering. In our Bayesian model, the parameters of the task-specific LoRA module

s are regarded as random variables where these task-wise LoRA modules are govern
ed/regularized by higher-level latent random variables, which represents the pri
or of the LoRA modules that capture the shared information across all training t
asks. To make the posterior inference feasible, we propose a novel SGLD-Gibbs sa
mpling algorithm that is computationally efficient. To represent the posterior s
amples from the SGLD-Gibbs, we propose an online EM algorithm that maintains a G
aussian mixture representation for the posterior in an online manner in the cour
se of iterative posterior sampling. We demonstrate the effectiveness of LiFT on
NLP and vision multi-task meta learning benchmarks.
****

Jinpeng Li,Haiping Wang,Jiabin chen,Yuan Liu,Zhiyang Dou,Yuexin Ma,Sibei Yang,Yu
an Li,Wenping Wang,Zhen Dong,Bisheng Yang
CityAnchor: City-scale 3D Visual Grounding with Multi-modality LLMs
In this paper, we present a 3D visual grounding method called CityAnchor for loc
alizing an urban object in a city-scale point cloud. Recent developments in mult
iview reconstruction enable us to reconstruct city-scale point clouds but how to
 conduct visual grounding on such a large-scale urban point cloud remains an ope
n problem. Previous 3D visual grounding system mainly concentrates on localizing
 an object in an image or a small-scale point cloud, which is not accurate and e
fficient enough to scale up to a city-scale point cloud. We address this problem
 with a multi-modality LLM which consists of two stages, a coarse localization a
nd a fine-grained matching. Given the text descriptions, the coarse localization
 stage locates possible regions on a projected 2D map of the point cloud while t
he fine-grained matching stage accurately determines the most matched object in
these possible regions. We conduct experiments on the CityRefer dataset and a ne
w synthetic dataset annotated by us, both of which demonstrate our method can pr
oduce accurate 3D visual grounding on a city-scale 3D point cloud.
****

Junru Song,Yang Yang,Huan Xiao,Wei Peng,Wen Yao,Feifei Wang
LASeR: Towards Diversified and Generalizable Robot Design with Large Language Mo
dels
Recent advances in Large Language Models (LLMs) have stimulated a significant pa
radigm shift in evolutionary optimization, where hand-crafted search heuristics
are gradually replaced with LLMs serving as intelligent search operators. Howeve
r, these studies still bear some notable limitations, including a challenge to b
alance exploitation with exploration, often leading to inferior solution diversi
ty, as well as poor generalizability of problem solving across different task se
ttings. These unsolved issues render the prowess of LLMs in robot design automat
ion largely untapped. In this work, we present LASeR -- Large Language Model-Aid
ed Evolutionary Search for Robot Design Automation. Leveraging a novel reflectio
n mechanism termed DiRect, we elicit more knowledgeable exploratory behaviors fr
om LLMs based on past search trajectories, reshaping the exploration-exploitatio
n tradeoff with dual improvements in optimization efficiency and solution divers
ity. Additionally, with evolution fully grounded in task-related background info
rmation, we unprecedentedly uncover the inter-task reasoning capabilities of LLM
s, facilitating generalizable design processes that effectively inspire zero-sho
t robot proposals for new applications. Our simulated experiments on voxel-based
 soft robots showcase distinct advantages of LASeR over competitive baselines. C
ode at https://github.com/WoodySJR/LASeR.
****

Jinlan Fu,huangfushenzhen,Hao Fei,Xiaoyu Shen,Bryan Hooi,Xipeng Qiu,See-Kiong Ng
CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLM
s
Multimodal Large Language Models (MLLMs) still struggle with hallucinations desp
ite their impressive capabilities. Recent studies have attempted to mitigate thi
s by applying Direct Preference Optimization (DPO) to multimodal scenarios using
 preference pairs from text-based responses. However, our analysis of representa
tion distributions reveals that multimodal DPO struggles to align image and text
 representations and to distinguish between hallucinated and non-hallucinated de
scriptions. To address these challenges,

In this work, we propose a Cross-modal Hierarchical Direct Preference Optimizati
on (CHiP) to address these limitations.
We introduce a visual preference optimization module within the DPO framework, e
nabling MLLMs to learn from both textual and visual preferences simultaneously.
Furthermore, we propose a hierarchical textual preference optimization module th
at allows the model to capture preferences at multiple granular levels, includin
g response, segment, and token levels. We evaluate CHiP through both quantitativ
e and qualitative analyses, with results across multiple benchmarks demonstratin
g its effectiveness in reducing hallucinations. On the Object HalBench dataset,
CHiP outperforms DPO in hallucination reduction, achieving improvements of 52.7%
 and 55.5% relative points based on the base model Muffin and LLaVA models, resp
ectively. We make all our datasets and code publicly available.
****
Yingzhen Yang,Ping Li
Sketching for Convex and Nonconvex Regularized Least Squares with Sharp Guarante
es
Randomized algorithms play a crucial role in efficiently solving large-scale opt
imization problems. In this paper, we introduce Sketching for Regularized Optimi
zation (SRO), a fast sketching algorithm designed for least squares problems wit
h convex or nonconvex regularization. SRO operates by first creating a sketch of
 the original data matrix and then solving the sketched problem. We establish mi
nimax optimal rates for sparse signal estimation by addressing the sketched spar
se convex and nonconvex learning problems. Furthermore, we propose a novel Itera
tive SRO algorithm, which reduces the approximation error geometrically for sket
ched convex regularized problems. To the best of our knowledge, this work is amo
ng the first to provide a unified theoretical framework demonstrating minimax ra
tes for convex and nonconvex sparse learning problems via sketching. Experimenta
l results validate the efficiency and effectiveness of both the SRO and Iterativ
e SRO algorithms.
****
John Vastola
Generalization through variance: how noise shapes inductive biases in diffusion
models
How diffusion models generalize beyond their training set is not known, and is s
omewhat mysterious given two facts: the optimum of the denoising score matching
(DSM) objective usually used to train diffusion models is the score function of
the training distribution; and the networks usually used to learn the score func
tion are expressive enough to learn this score to high accuracy. We claim that a
 certain feature of the DSM objective—the fact that its target is not the traini
ng distribution's score, but a noisy quantity only equal to it in expectation—st
rongly impacts whether and to what extent diffusion models generalize. In this p
aper, we develop a mathematical theory that partly explains this 'generalization
 through variance' phenomenon. Our theoretical analysis exploits a physics-inspi
red path integral approach to compute the distributions typically learned by a f
ew paradigmatic under- and overparameterized diffusion models. We find that the
distributions diffusion models effectively learn to sample from resemble their t
raining distributions, but with `gaps' filled in, and that this inductive bias i
s due to the covariance structure of the noisy target used during training. We a
lso characterize how this inductive bias interacts with feature-related inductiv
e biases.
****
Hyunin Lee,Chanwoo Park,David Abel,Ming Jin
A Black Swan Hypothesis: The Role of Human Irrationality in AI Safety
Black swan events are statistically rare occurrences that carry extremely high r
isks. A typical view of defining black swan events is heavily assumed to origina
te from an unpredictable time-varying environments; however, the community lacks
 a comprehensive definition of black swan events. To this end, this paper challe
nges that the standard view is incomplete and claims that high-risk, statistical
ly rare events can also occur in unchanging environments due to human mispercept
ion of their value and likelihood, which we call as spatial black swan event. We

first carefully categorize black swan events, focusing on spatial black swan events, and mathematically formalize the definition of black swan events. We hope these definitions can pave the way for the development of algorithms to prevent such events by rationally correcting human perception.
****
Amirkeivan Mohtashami,Matteo Pagliardini,Martin Jaggi
CoTFormer: A Chain of Thought Driven Architecture with Budget-Adaptive Computation Cost at Inference
Scaling language models to larger and deeper sizes has led to significant boosts in performance. Even though the size of these models limits their application in compute-constrained environments, the race to continually develop ever larger and deeper foundational models is underway. At the same time---regardless of the model size---task-specific techniques continue to play a pivotal role in achieving optimal downstream performance. One of these techniques, called Chain-of-Thought (CoT), is particularly interesting since, as we point out in this work, it resembles employing a deeper transformer through re-applying the model multiple times. However, a key subtlety in computing the attention of past tokens differentiates CoT from simply applying the model several times. Based on this insight, we propose CoTFormer, a novel architecture which closely mimics CoT at the token level, allowing us to obtain significantly improved accuracies close to much larger models. While applying CoT introduces additional computation costs, we compensate for it by leveraging CoTFormer's special compatibility with token-wise variable depth. Through a compute adaptive model---which automatically allocates the compute to tokens that need it most---we show that it is possible to reduce the computation cost significantly without any reduction in accuracy, and with further compute cost reductions possible while maintaining a competitive accuracy.
****
Huajian Huang,Yingshu Chen,Longwei Li,Hui Cheng,Tristan Braud,Yajie Zhao,Sai-Kit Yeung
SC-OmniGS: Self-Calibrating Omnidirectional Gaussian Splatting
360-degree cameras streamline data collection for radiance field 3D reconstruction by capturing comprehensive scene data. However, traditional radiance field methods do not address the specific challenges inherent to 360-degree images. We present SC-OmniGS, a novel self-calibrating omnidirectional Gaussian splatting system for fast and accurate omnidirectional radiance field reconstruction using 360-degree images. Rather than converting 360-degree images to cube maps and performing perspective image calibration, we treat 360-degree images as a whole sphere and derive a mathematical framework that enables direct omnidirectional camera pose calibration accompanied by 3D Gaussians optimization. Furthermore, we introduce a differentiable omnidirectional camera model in order to rectify the distortion of real-world data for performance enhancement. Overall, the omnidirectional camera intrinsic model, extrinsic poses, and 3D Gaussians are jointly optimized by minimizing weighted spherical photometric loss. Extensive experiments have demonstrated that our proposed SC-OmniGS is able to recover a high-quality radiance field from noisy camera poses or even no pose prior in challenging scenarios characterized by wide baselines and non-object-centric configurations. The noticeable performance gain in the real-world dataset captured by consumer-grade omnidirectional cameras verifies the effectiveness of our general omnidirectional camera model in reducing the distortion of 360-degree images.
****
Jonathan Brokman,Amit Giloni,Omer Hofman,Roman Vainshtein,Hisashi Kojima,Guy Gilboa
Manifold Induced Biases for Zero-shot and Few-shot Detection of Generated Images
Distinguishing between real and AI-generated images, commonly referred to as 'image detection', presents a timely and significant challenge. Despite extensive research in the (semi-)supervised regime, zero-shot and few-shot solutions have only recently emerged as promising alternatives. Their main advantage is in alleviating the ongoing data maintenance, which quickly becomes outdated due to advances in generative technologies. We identify two main gaps: (1) a lack of theore

tical grounding for the methods, and (2) significant room for performance improv
ements in zero-shot and few-shot regimes. Our approach is founded on understandi
ng and quantifying the biases inherent in generated content, where we use these
quantities as criteria for characterizing generated images. Specifically, we exp
lore the biases of the implicit probability manifold, captured by a pre-trained
diffusion model. Through score-function analysis, we approximate the curvature,
gradient, and bias towards points on the probability manifold, establishing crit
eria for detection in the zero-shot regime. We further extend our contribution t
o the few-shot setting by employing a mixture-of-experts methodology. Empirical
results across 20 generative models demonstrate that our method outperforms curr
ent approaches in both zero-shot and few-shot settings. This work advances the t
heoretical understanding and practical usage of generated content biases through
 the lens of manifold analysis.
****

Nikola ■uki■,Tim Lebailly,Tinne Tuytelaars

Object-Centric Pretraining via Target Encoder Bootstrapping

Object-centric representation learning has recently been successfully applied to
 real-world datasets. This success can be attributed to pretrained non-object-ce
ntric foundation models, whose features serve as reconstruction targets for slot
 attention. However, targets must remain frozen throughout the training, which s
ets an upper bound on the performance object-centric models can attain. Attempts
 to update the target encoder by bootstrapping result in large performance drops
, which can be attributed to its lack of object-centric inductive biases, causin
g the object-centric model’s encoder to drift away from representations useful a
s reconstruction targets. To address these limitations, we propose **O**bject-**
CE**ntric Pretraining by Target Encoder **BO**otstrapping, a self-distillation s
etup for training object-centric models from scratch, on real-world data, for th
e first time ever. In OCEBO, the target encoder is updated as an exponential mov
ing average of the object-centric model, thus explicitly being enriched with obj
ect-centric inductive biases introduced by slot attention while removing the upp
er bound on performance present in other models. We mitigate the slot collapse c
aused by random initialization of the target encoder by introducing a novel cros
s-view patch filtering approach that limits the supervision to sufficiently info
rmative patches. When pretrained on 241k images from COCO, OCEBO achieves unsupe
rvised object discovery performance comparable to that of object-centric models
with frozen non-object-centric target encoders pretrained on hundreds of million
s of images. The code and pretrained models are publicly available at https://gi
thub.com/djukicn/ocebo.
****

Maria Drencheva,Ivo Petrov,Maximilian Baader,Dimitar Iliev Dimitrov,Martin Veche
v

GRAIN: Exact Graph Reconstruction from Gradients

Federated learning claims to enable collaborative model training among multiple
clients with data privacy by transmitting gradient updates instead of the actual
 client data. However, recent studies have shown the client privacy is still at
risk due to the, so called, gradient inversion attacks which can precisely recon
struct clients' text and image data from the shared gradient updates. While thes
e attacks demonstrate severe privacy risks for certain domains and architectures
, the vulnerability of other commonly-used data types, such as graph-structured
data, remain under-explored. To bridge this gap, we present GRAIN, the first exa
ct gradient inversion attack on graph data in the honest-but-curious setting tha
t recovers both the structure of the graph and the associated node features. Con
cretely, we focus on Graph Convolutional Networks (GCN) and Graph Attention Netw
orks (GAT) -- two of the most widely used frameworks for learning on graphs. Our
 method first utilizes the low-rank structure of GNN gradients to efficiently re
construct and filter the client subgraphs which are then joined to complete the
input graph. We evaluate our approach on molecular, citation, and social network
 datasets using our novel metric.
We show that GRAIN reconstructs up to 80\% of all graphs exactly, significantly
outperforming the baseline, which achieves up to 20\% correctly positioned nodes

.
****
Jialong Chen,Bowen Deng,Zhen WANG,Chuan Chen,Zibin Zheng
Graph Neural Ricci Flow: Evolving Feature from a Curvature Perspective
Differential equations provide a dynamical perspective for understanding and des
igning graph neural networks (GNNs). By generalizing the discrete Ricci flow (DR
F) to attributed graphs, we can leverage a new paradigm for the evolution of nod
e features with the help of curvature. We show that in the attributed graphs, DR
F guarantees a vital property: The curvature of each edge concentrates toward ze
ro over time. This property leads to two interesting consequences: 1) graph Diri
chlet energy with bilateral bounds and 2) data-independent curvature decay rate.
 Based on these theoretical results, we propose the Graph Neural Ricci Flow (GNR
F), a novel curvature-aware continuous-depth GNN. Compared to traditional curvat
ure-based graph learning methods, GNRF is not limited to a specific curvature de
finition. It computes and adjusts time-varying curvature efficiently in linear t
ime. We also empirically illustrate the operating mechanism of GNRF and verify t
hat it performs excellently on diverse datasets.
****
Ariel Elnekave,Yair Weiss
Do WGANs succeed because they minimize the Wasserstein Distance? Lessons from Di
screte Generators
Since WGANs were first introduced, there has been considerable debate whether th
eir success in generating realistic images can be attributed to minimizing the W
asserstein distance between the distribution of generated images and the traini
ng distribution. In this paper we present theoretical and experimental results t
hat show that successful WGANs {\em do} minimize the Wasserstein distance but th
e form of the distance that is minimized depends highly on the discriminator arc
hitecture and its inductive biases. Specifically, we show that when the discrimi
nator is convolutional, WGANs minimize the Wasserstein distance between {\em pa
tches} in the generated images and the training images, not the Wasserstein dis
tance between images.
Our results are obtained by considering {\em discrete} generators for which the
Wasserstein distance between the generator distribution and the training distrib
ution can be computed exactly and the minimum can be characterized analytically.
   We present experimental results with discrete GANs that generate realistic f
ake images (comparable in quality to their continuous counterparts) and present
evidence that they are minimizing the Wasserstein distance between real and fake
 patches and not the distance between real and fake images.
****
Yuheng Li,Wang Panpan,Haipeng Chen
Can Reinforcement Learning Solve Asymmetric Combinatorial-Continuous Zero-Sum Ga
mes?
There have been extensive studies on learning in zero-sum games, focusing on the
 analysis of the existence and algorithmic convergence of Nash equilibrium (NE).
 Existing studies mainly focus on symmetric games where the strategy spaces of t
he players are of the same type and size. For the few studies that do consider a
symmetric games, they are mostly restricted to matrix games. In this paper, we d
efine and study a new practical class of asymmetric games called two-player Asym
metric Combinatorial-Continuous zEro-Sum (ACCES) games, featuring a combinatoria
l action space for one player and an infinite compact space for the other. Such
ACCES games have broad implications in the real world, particularly in combinato
rial optimization problems (COPs) where one player optimizes a solution in a com
binatorial space, and the opponent plays against it in an infinite (continuous)
compact space (e.g., a nature player deciding epistemic parameters of the enviro
nmental model). Our first key contribution is to prove the existence of NE for t
wo-player ACCES games, using the idea of essentially finite game approximation.
Building on the theoretical insights and double oracle (DO)-based solutions to c
omplex zero-sum games, our second contribution is to design the novel algorithm,
 Combinatorial Continuous DO (CCDO), to solve ACCES games, and prove the converg
ence of the proposed algorithm. Considering the NP-hardness of most COPs and rec

ent advancements in reinforcement learning (RL)-based solutions to COPs, our thi
rd contribution is to propose a practical algorithm to solve NE in the real worl
d, CCDORL (based on CCDO) and provide the novel convergence analysis in the ACCE
S game. Experimental results across diverse instances of COPs demonstrate the em
pirical effectiveness of our algorithms.
****
Hongbo Li,Sen Lin,Lingjie Duan,Yingbin Liang,Ness Shroff
Theory on Mixture-of-Experts in Continual Learning
Continual learning (CL) has garnered significant attention because of its abilit
y to adapt to new tasks that arrive over time. Catastrophic forgetting (of old t
asks) has been identified as a major issue in CL, as the model adapts to new tas
ks. The Mixture-of-Experts (MoE) model has recently been shown to effectively mi
tigate catastrophic forgetting in CL, by employing a gating network to sparsify
and distribute diverse tasks among multiple experts. However, there is a lack of
 theoretical analysis of MoE and its impact on the learning performance in CL. T
his paper provides the first theoretical results to characterize the impact of M
oE in CL via the lens of overparameterized linear regression tasks. We establish
 the benefit of MoE over a single expert by proving that the MoE model can diver
sify its experts to specialize in different tasks, while its router learns to se
lect the right expert for each task and balance the loads across all experts. Ou
r study further suggests an intriguing fact that the MoE in CL needs to terminat
e the update of the gating network after sufficient training rounds to attain sy
stem convergence, which is not needed in the existing MoE studies that do not co
nsider the continual task arrival. Furthermore, we provide explicit expressions
for the expected forgetting and overall generalization error to characterize the
 benefit of MoE in the learning performance in CL. Interestingly, adding more ex
perts requires additional rounds before convergence, which may not enhance the l
earning performance. Finally, we conduct experiments on both synthetic and real
datasets to extend these insights from linear models to deep neural networks (DN
Ns), which also shed light on the practical algorithm design for MoE in CL.
****
Debargha Ganguly,Warren Richard Morningstar,Andrew Seohwan Yu,Vipin Chaudhary
Forte : Finding Outliers with Representation Typicality Estimation
Generative models can now produce photorealistic synthetic data which is virtual
ly indistinguishable from the real data used to train it. This is a significant
evolution over previous models which could produce reasonable facsimiles of the
training data, but ones which could be visually distinguished from the training
data by human evaluation. Recent work on OOD detection has raised doubts that ge
nerative model likelihoods are optimal OOD detectors due to issues involving lik
elihood misestimation, entropy in the generative process, and typicality. We spe
culate that generative OOD detectors also failed because their models focused on
 the pixels rather than the semantic content of the data, leading to failures in
 near-OOD cases where the pixels may be similar but the information content is s
ignificantly different. We hypothesize that estimating typical sets using self-s
upervised learners leads to better OOD detectors. We introduce a novel approach
that leverages representation learning, and informative summary statistics based
 on manifold estimation, to address all of the aforementioned issues. Our method
 outperforms other unsupervised approaches and achieves state-of-the art perform
ance on well-established challenging benchmarks, and new synthetic data detectio
n tasks.
****
Wenlong Wang,Ivana Dusparic,Yucheng Shi,Ke Zhang,Vinny Cahill
Drama: Mamba-Enabled Model-Based Reinforcement Learning Is Sample and Parameter
Efficient
Model-based reinforcement learning (RL) offers a solution to the data inefficien
cy that plagues most model-free RL algorithms. However, learning a robust world
model often requires complex and deep architectures, which are computationally e
xpensive and challenging to train. Within the world model, sequence models play
a critical role in accurate predictions, and various architectures have been exp
lored, each with its own challenges. Currently, recurrent neural network (RNN)-b

ased world models struggle with vanishing gradients and capturing long-term depe
ndencies. Transformers, on the other hand, suffer from the quadratic memory and
computational complexity of self-attention mechanisms, scaling as $O(n^2)$, wher
e n is the sequence length.

To address these challenges, we propose a state space model (SSM)-based world mo
del, Drama, specifically leveraging Mamba, that achieves $O(n)$ memory and compu
tational complexity while effectively capturing long-term dependencies and enabl
ing efficient training with longer sequences. We also introduce a novel sampling
 method to mitigate the suboptimality caused by an incorrect world model in the
early training stages. Combining these techniques, Drama achieves a normalised s
core on the Atari100k benchmark that is competitive with other state-of-the-art
(SOTA) model-based RL algorithms, using only a 7 million-parameter world model.
Drama is accessible and trainable on off-the-shelf hardware, such as a standard
laptop. Our code is available at https://github.com/realwenlongwang/Drama.git.
****
Shiyao Li,Antoine Guedon,Clémentin Boittiaux,Shizhe Chen,Vincent Lepetit
NextBestPath: Efficient 3D Mapping of Unseen Environments
This work addresses the problem of active 3D mapping, where an agent must find a
n efficient trajectory to exhaustively reconstruct a new scene.
Previous approaches mainly predict the next best view near the agent's location,
 which is prone to getting stuck in local areas. Additionally, existing indoor d
atasets are insufficient due to limited geometric complexity and inaccurate grou
nd truth meshes.
To overcome these limitations, we introduce a novel dataset AiMDoom with a map g
enerator for the Doom video game, enabling to better benchmark active 3D mapping
 in diverse indoor environments.
Moreover, we propose a new method we call next-best-path (NBP), which predicts l
ong-term goals rather than focusing solely on short-sighted views.
The model jointly predicts accumulated surface coverage gains for long-term goal
s and obstacle maps, allowing it to efficiently plan optimal paths with a unifie
d model.
By leveraging online data collection, data augmentation and curriculum learning,
 NBP significantly outperforms state-of-the-art methods on both the existing MP3
D dataset and our AiMDoom dataset, achieving more efficient mapping in indoor en
vironments of varying complexity.
****
Che-Ping Tsai,Ganyu Teng,Phillip Wallis,Wei Ding
AnoLLM: Large Language Models for Tabular Anomaly Detection
We introduce AnoLLM, a novel framework that leverages large language models (LLM
s) for unsupervised tabular anomaly detection. By converting tabular data into
a standardized text format, we further adapt a pre-trained LLM with this seriali
zed data, and assign anomaly scores based on the negative log likelihood generat
ed by the LLM. Unlike traditional methods that can require extensive feature eng
ineering, and often lose textual information during data processing, AnoLLM pres
erves data integrity and streamlines the preprocessing required for tabular anom
aly detection. This approach can effectively handle mixed-type data, especially
those containing textual features. Our empirical results indicate that AnoLLM de
livers the best performance on six benchmark datasets with mixed feature types.
Additionally, across 30 datasets from the ODDS library, which are predominantly
numerical, AnoLLM performs on par with top performing baselines.
****
Qingchen Yu,Zifan Zheng,Shichao Song,Zhiyu li,Feiyu Xiong,Bo Tang,Ding Chen
xFinder: Large Language Models as Automated Evaluators for Reliable Evaluation
The continuous advancement of large language models (LLMs) has brought increasin
g attention to the critical issue of developing fair and reliable methods for ev
aluating their performance. Particularly, the emergence of cheating phenomena, s
uch as test set leakage and prompt format overfitting, poses significant challen
ges to the reliable evaluation of LLMs. As evaluation frameworks commonly use Re
gular Expression (RegEx) for answer extraction, models may adjust their response

s to fit formats easily handled by RegEx. Nevertheless, the key answer extractio
n module based on RegEx frequently suffers from extraction errors. Furthermore,
recent studies proposing fine-tuned LLMs as judge models for automated evaluatio
n face challenges in terms of generalization ability and fairness. This paper co
mprehensively analyzes the entire LLM evaluation chain and demonstrates that opt
imizing the key answer extraction module improves extraction accuracy and enhanc
es evaluation reliability. Our findings suggest that improving the key answer ex
traction module can lead to higher judgment accuracy and improved evaluation eff
iciency compared to the judge models. To address these issues, we propose xFinde
r, a novel evaluator for answer extraction and matching in LLM evaluation. As pa
rt of this process, we create a specialized dataset, the Key Answer Finder (KAF)
 dataset, to ensure effective model training and evaluation. Generalization test
s and real-world evaluations show that the smallest xFinder model, with only 500
 million parameters, achieves an average extraction accuracy of 93.42\%. In cont
rast, RegEx accuracy in the best evaluation framework is 74.38\%. The final judg
ment accuracy of xFinder reaches 97.61\%, outperforming existing evaluation fram
eworks and judge models.
****
Zhiyang Xu,Minqian Liu,Ying Shen,Joy Rimchala,Jiaxin Zhang,Qifan Wang,Yu Cheng,L
ifu Huang
Modality-Specialized Synergizers for Interleaved Vision-Language Generalists
Recent advancements in Vision-Language Models (VLMs) have led to the emergence o
f Vision-Language Generalists (VLGs) capable of understanding and generating bot
h text and images. However, seamlessly generating an arbitrary sequence of text
and images remains a challenging task for the current VLGs. One primary limitati
on lies in applying a unified architecture and the same set of parameters to sim
ultaneously model discrete text tokens and continuous image features. Recent wor
ks attempt to tackle this fundamental problem by introducing modality-aware expe
rt models. However, they employ identical architectures to process both text and
 images, disregarding the intrinsic inductive biases in these two modalities. In
 this work, we introduce Modality-Specialized Synergizers (MoSS), a novel design
 that efficiently optimizes existing unified architectures of VLGs with modality
-specialized adaptation layers, i.e., a Convolutional LoRA for modeling the loca
l priors of image patches and a Linear LoRA for processing sequential text. This
 design enables more effective modeling of modality-specific features while main
taining the strong cross-modal integration gained from pretraining. In addition,
 to improve the instruction-following capability on interleaved text-and-image g
eneration, we introduce LeafInstruct, the first open-sourced interleaved instruc
tion tuning dataset comprising 184,982 high-quality instances on more than 10 di
verse domains. Extensive experiments show that VLGs integrated with MoSS achieve
 state-of-the-art performance, significantly surpassing baseline VLGs in complex
 interleaved generation tasks. Furthermore, our method exhibits strong generaliz
ability on different VLGs.
****
Haoyuan Sun,Zihao Wu,Bo Xia,Pu Chang,Zibin Dong,Yifu Yuan,Yongzhe Chang,Xueqian
Wang
Entropy-based Activation Function Optimization: A Method on Searching Better Act
ivation Functions
The success of artificial neural networks (ANNs) hinges greatly on the judicious
 selection of an activation function, introducing non-linearity into network and
 enabling them to model sophisticated relationships in data. However, the search
 of activation functions has largely relied on empirical knowledge in the past,
lacking theoretical guidance, which has hindered the identification of more effe
ctive activation functions. In this work, we offer a proper solution to such iss
ue. Firstly, we theoretically demonstrate the existence of the worst activation
function with boundary conditions (WAFBC) from the perspective of information en
tropy. Furthermore, inspired by the Taylor expansion form of information entropy
 functional, we propose the Entropy-based Activation Function Optimization (EAFO
) methodology. EAFO methodology presents a novel perspective for designing stati
c activation functions in deep neural networks and the potential of dynamically

optimizing activation during iterative training. Utilizing EAFO methodology, we derive a novel activation function from ReLU, known as Correction Regularized ReLU (CRReLU). Experiments conducted with vision transformer and its variants on CIFAR-10, CIFAR-100 and ImageNet-1K datasets demonstrate the superiority of CRReLU over existing corrections of ReLU. Extensive empirical studies on task of large language model (LLM) fine-tuning, CRReLU exhibits superior performance compared to GELU, suggesting its broader potential for practical applications.
****
Teun van der Weij,Felix Hofstätter,Oliver Jaffe,Samuel F. Brown,Francis Rhys Ward
AI Sandbagging: Language Models can Strategically Underperform on Evaluations
Trustworthy capability evaluations are crucial for ensuring the safety of AI systems, and are becoming a key component of AI regulation. However, the developers of an AI system, or the AI system itself, may have incentives for evaluations to understate the AI's actual capability. These conflicting interests lead to the problem of *sandbagging* – which we define as *strategic underperformance on an evaluation*. In this paper we assess sandbagging capabilities in contemporary language models (LMs). We prompt frontier LMs, like GPT-4 and Claude 3 Opus, to selectively underperform on dangerous capability evaluations, while maintaining performance on general (harmless) capability evaluations. Moreover, we find that models can be fine-tuned, on a synthetic dataset, to hide specific capabilities unless given a password. This behaviour generalizes to high-quality, held-out benchmarks such as WMDP. In addition, we show that both frontier and smaller models can be prompted or password-locked to target specific scores on a capability evaluation. We have mediocre success in password-locking a model to mimic the answers a weaker model would give. Overall, our results suggest that capability evaluations are vulnerable to sandbagging. This vulnerability decreases the trustworthiness of evaluations, and thereby undermines important safety decisions regarding the development and deployment of advanced AI systems.

See our code at https://github.com/TeunvdWeij/sandbagging
****
Zakhar Shumaylov,Peter Zaika,James Rowbottom,Ferdia Sherry,Melanie Weber,Carola-Bibiane Schönlieb
Lie Algebra Canonicalization: Equivariant Neural Operators under arbitrary Lie Groups
The quest for robust and generalizable machine learning models has driven recent interest in exploiting symmetries through equivariant neural networks. In the context of PDE solvers, recent works have shown that Lie point symmetries can be a useful inductive bias for Physics-Informed Neural Networks (PINNs) through data and loss augmentation. Despite this, directly enforcing equivariance within the model architecture for these problems remains elusive. This is because many PDEs admit non-compact symmetry groups, oftentimes not studied beyond their infinitesimal generators, making them incompatible with most existing equivariant architectures. In this work, we propose Lie aLgebrA Canonicalization (LieLAC), a novel approach that exploits only the action of infinitesimal generators of the symmetry group, circumventing the need for knowledge of the full group structure. To achieve this, we address existing theoretical issues in the canonicalization literature, establishing connections with frame averaging in the case of continuous non-compact groups. Operating within the framework of canonicalization, LieLAC can easily be integrated with unconstrained pre-trained models, transforming inputs to a canonical form before feeding them into the existing model, effectively aligning the input for model inference according to allowed symmetries. LieLAC utilizes standard Lie group descent schemes, achieving equivariance in pre-trained models. Finally, we showcase LieLAC's efficacy on tasks of invariant image classification and Lie point symmetry equivariant neural PDE solvers using pre-trained models.
****
Diego Calanzone,Stefano Teso,Antonio Vergari
Logically Consistent Language Models via Neuro-Symbolic Integration

Large language models (LLMs) are a promising venue for natural language understanding and generation tasks. However, current LLMs are far from reliable: they are prone to generate non-factual information and, more crucially, to contradict themselves when prompted to reason about relations between real entities of the world. These problems are currently addressed with large scale fine-tuning or by delegating consistent reasoning to external tools. In this work, we strive for a middle ground and leverage a training objective based on a principled neuro-symbolic loss that teaches a LLM to be consistent with external knowledge in the form of a set of facts and rules. Fine-tuning with such a loss on a limited set of facts enables our LLMs to be more logically consistent than previous baselines for a given constraint. Our approach also allows to easily combine multiple logical constraints at once in a principled way, delivering LLMs that are more consistent w.r.t. all the selected rules. Moreover, our method allows LLMs to extrapolate to unseen but semantically similar factual knowledge, represented in unseen datasets, more systematically.
****
Shaonan Wu,Shuai Lu,Yeyun Gong,Nan Duan,Ping Wei
Alchemy: Amplifying Theorem-Proving Capability Through Symbolic Mutation
Formal proofs are challenging to write even for experienced experts. Recent progress in Neural Theorem Proving (NTP) shows promise in expediting this process. However, the formal corpora available on the Internet are limited compared to the general text, posing a significant data scarcity challenge for NTP. To address this issue, this work proposes Alchemy, a general framework for data synthesis that constructs formal theorems through symbolic mutation. Specifically, for each candidate theorem in Mathlib, we identify all invocable theorems that can be used to rewrite or apply to it. Subsequently, we mutate the candidate theorem by replacing the corresponding term in the statement with its equivalent form or antecedent. As a result, our method increases the number of theorems in Mathlib by an order of magnitude, from 110k to 6M. Furthermore, we perform continual pretraining and supervised finetuning on this augmented corpus for large language models. Experimental results demonstrate the effectiveness of our approach, achieving a 4.70% absolute performance improvement on Leandojo benchmark. Additionally, our approach achieves a 2.47% absolute performance gain on the out-of-distribution miniF2F benchmark based on the synthetic data. To provide further insights, we conduct a comprehensive analysis of synthetic data composition and the training paradigm, offering valuable guidance for developing a strong theorem prover.
****
Sullam Jeoung,Yubin Ge,Haohan Wang,Jana Diesner
Examining Alignment of Large Language Models through Representative Heuristics: the case of political stereotypes
Examining the alignment of large language models (LLMs) has become increasingly important, e.g., when LLMs fail to operate as intended. This study examines the alignment of LLMs with human values for the domain of politics. Prior research has shown that LLM-generated outputs can include political leanings and mimic the stances of political parties on various issues. However, the extent and conditions under which LLMs deviate from empirical positions are insufficiently examined. To address this gap, we analyze the factors that contribute to LLMs' deviations from empirical positions on political issues, aiming to quantify these deviations and identify the conditions that cause them.

Drawing on findings from cognitive science about representativeness heuristics, i.e., situations where humans lean on representative attributes of a target group in a way that leads to exaggerated beliefs, we scrutinize LLM responses through this heuristics' lens. We conduct experiments to determine how LLMs inflate predictions about political parties, which results in stereotyping. We find that while LLMs can mimic certain political parties' positions, they often exaggerate these positions more than human survey respondents do. Also, LLMs tend to overemphasize representativeness more than humans. This study highlights the susceptibility of LLMs to representativeness heuristics, suggesting a potential vulnerabi

lity of LLMs that facilitates political stereotyping. We also test prompt-based mitigation strategies, finding that strategies that can mitigate representative heuristics in humans are also effective in reducing the influence of representat iveness on LLM-generated responses.
****
Rana Shahout,eran malach,Chunwei Liu,Weifan Jiang,Minlan Yu,Michael Mitzenmacher
DON’T STOP ME NOW: EMBEDDING BASED SCHEDULING FOR LLMS
Efficient scheduling is crucial for interactive Large Language Model (LLM) appli cations, where low request completion time directly impacts user engagement. Siz e-based scheduling algorithms like Shortest Remaining Process Time (SRPT) aim to reduce average request completion time by leveraging known or estimated request sizes and allowing preemption by incoming jobs with shorter service times. Howe ver, two main challenges arise when applying size-based scheduling to LLM system s. First, accurately predicting output lengths from prompts is challenging and o ften resource-intensive, making it impractical for many systems. As a result, th e state-of-the-art LLM systems default to first-come, first-served scheduling, w hich can lead to head-of-line blocking and reduced system efficiency. Second, pr eemption introduces extra memory overhead to LLM systems as they must maintain i ntermediate states for unfinished (preempted) requests.
In this paper, we propose TRAIL, a method to obtain output predictions from the target LLM itself. After generating each output token, we recycle the embedding of its internal structure as input for a lightweight classifier that predicts th e remaining length for each running request. Using these predictions, we propose a prediction-based SRPT variant with limited preemption designed to account for memory overhead in LLM systems. This variant allows preemption early in request execution when memory consumption is low but restricts preemption as requests a pproach completion to optimize resource utilization. On the theoretical side, we derive a closed-form formula for this SRPT variant in an M/G/1 queue model, whi ch demonstrates its potential value. In our system, we implement this preemption policy alongside our embedding-based prediction method. Our refined predictions from layer embeddings achieve 2.66x lower mean absolute error compared to BERT predictions from sequence prompts. TRAIL achieves 1.66x to 2.01x lower mean late ncy on the Alpaca dataset and 1.76x to 24.07x lower mean time to the first token compared to the state-of-the-art serving system.
****
Jiahua Dong,Yu-Xiong Wang
3DGS-Drag: Dragging Gaussians for Intuitive Point-Based 3D Editing
The transformative potential of 3D content creation has been progressively unloc ked through advancements in generative models. Recently, intuitive drag editing with geometric changes has attracted significant attention in 2D editing yet rem ains challenging for 3D scenes. In this paper, we introduce 3DGS-Drag, a point-b ased 3D editing framework that provides efficient, intuitive drag manipulation o f real 3D scenes. Our approach bridges the gap between deformation-based and 2D-editing-based 3D editing methods, addressing their limitations to geometry-relat ed content editing. We leverage two key innovations: deformation guidance utiliz ing 3D Gaussian Splatting for consistent geometric modifications and diffusion g uidance for content correction and visual quality enhancement. A progressive edi ting strategy further supports aggressive 3D drag edits. Our method enables a wi de range of edits, including motion change, shape adjustment, inpainting, and co ntent extension. Experimental results demonstrate the effectiveness of 3DGS-Drag in various scenes, achieving state-of-the-art performance in geometry-related 3 D content editing. Notably, the editing is efficient, taking 10 to 20 minutes on a single RTX 4090 GPU.
****
Matteo Gallici,Mattie Fellows,Benjamin Ellis,Bartomeu Pou,Ivan Masmitja,Jakob Ni colaus Foerster,Mario Martin
Simplifying Deep Temporal Difference Learning
Q-learning played a foundational role in the field reinforcement learning (RL) .
However, TD algorithms with off-policy data, such as Q-learning, or nonlinear

function approximation like deep neural networks require several additional tric ks to stabilise training, primarily a large replay buffer and target networks. U nfortunately, the delayed updating of frozen network parameters in the target ne twork harms the sample efficiency and, similarly, the large replay buffer introd uces memory and implementation overheads. In this paper, we investigate whether it is possible to accelerate and simplify off-policy TD training while maintaini ng its stability. Our key theoretical result demonstrates for the first time tha t regularisation techniques such as LayerNorm can yield provably convergent TD a lgorithms without the need for a target network or replay buffer, even with off-policy data. Empirically, we find that online, parallelised sampling enabled by vectorised environments stabilises training without the need for a large replay buffer. Motivated by these findings, we propose PQN, our simplified deep online Q-Learning algorithm. Surprisingly, this simple algorithm is competitive with more complex methods like: Rainbow in Atari, PPO-RNN in Craftax, QMix in Smax, a nd can be up to 50x faster than traditional DQN without sacrificing sample effic iency. In an era where PPO has become the go-to RL algorithm, PQN reestablishes off-policy Q-learning as a viable alternative.
****

Seulki Park,Youren Zhang,Stella X. Yu,Sara Beery,Jonathan Huang
Visually Consistent Hierarchical Image Classification
Hierarchical classification requires predicting an entire taxonomy, demanding bo th per-level accuracy and cross-level consistency. However, per-level accuracy o ften suffers because different levels rely on distinct visual cues, making it a multi-task problem. For example, identifying specific bird species (e.g., Anna' s hummingbird) depends on localized features like head and neck color, while dis tinguishing a bird from a plant requires global features like feathers versus le aves. Existing methods address this by using separate model components for each level, but this often leads to inconsistencies, as classifiers focus on unrelate d regions.

Our key insight is that classification across levels should be grounded in consi stent visual cues that link fine details to broader structures. Instead of treat ing fine and coarse classification separately, our model first captures key deta ils (e.g., beak, wings) for species recognition, then integrates them into globa l features (e.g., bird shape) for higher-level classification. For this, we intr oduce a hierarchical model that aligns fine-to-coarse semantic parsing with cons istent visual segmentation. Additionally, we propose a tree-path KL divergence l oss to enforce semantic consistency across levels. Our approach significantly o utperforms zero-shot CLIP and other state-of-the-art methods on hierarchical cla ssification benchmarks.
****

Jingling Li,Zeyu Tang,Xiaoyu Liu,Peter Spirtes,Kun Zhang,Liu Leqi,Yang Liu
Prompting Fairness: Integrating Causality to Debias Large Language Models
Large language models (LLMs), despite their remarkable capabilities, are suscept ible to generating biased and discriminatory responses. As LLMs increasingly inf luence high-stakes decision-making (e.g., hiring and healthcare), mitigating the se biases becomes critical. In this work, we propose a causality-guided debiasin g framework to tackle social biases, aiming to reduce the objectionable dependen ce between LLMs' decisions and the social information in the input. Our framewor k introduces a novel perspective to identify how social information can affect a n LLM's decision through different causal pathways. Leveraging these causal insi ghts, we outline principled prompting strategies that regulate these pathways th rough selection mechanisms. This framework not only unifies existing prompting-b ased debiasing techniques, but also opens up new directions for reducing bias by encouraging the model to prioritize fact-based reasoning over reliance on biase d social cues. We validate our framework through extensive experiments on real-w orld datasets across multiple domains, demonstrating its effectiveness in debias ing LLM decisions, even with only black-box access to the model.
****

Kai Jungel,Dario Paccagnan,Axel Parmentier,Maximilian Schiffer

WardropNet: Traffic Flow Predictions via Equilibrium-Augmented Learning
When optimizing transportation systems, anticipating traffic flows is a central element. Yet, computing such traffic equilibria remains computationally expensive. Against this background, we introduce a novel combinatorial optimization augmented neural network pipeline that allows for fast and accurate traffic flow predictions. We propose WardropNet, a neural network that combines classical layers with a subsequent equilibrium layer: the first ones inform the latter by predicting the parameterization of the equilibrium problem's latency functions. Using supervised learning we minimize the difference between the actual traffic flow and the predicted output. We show how to leverage a Bregman divergence fitting the geometry of the equilibria, which allows for end-to-end learning. WardropNet outperforms pure learning-based approaches in predicting traffic equilibria for realistic and stylized traffic scenarios. On realistic scenarios, WardropNet improves on average for time-invariant predictions by up to 72\% and for time-varian t predictions by up to 23\% over pure learning-based approaches.
****
Ricardo Dominguez-Olmedo,Vedant Nanda,Rediet Abebe,Stefan Bechtold,Christoph Engel,Jens Frankenreiter,Krishna P. Gummadi,Moritz Hardt,Michael Livermore
Lawma: The Power of Specialization for Legal Annotation
Annotation and classification of legal text are central components of empirical legal research. Traditionally, these tasks are often delegated to trained research assistants. Motivated by the advances in language modeling, empirical legal scholars are increasingly turning to commercial models, hoping that it will alleviate the significant cost of human annotation. In this work, we present a comprehensive analysis of large language models' current abilities to perform legal annotation tasks. To do so, we construct CaselawQA, a benchmark comprising 260 legal text classification tasks, nearly all new to the machine learning community. Starting from GPT-4 as a baseline, we show that it has non-trivial but highly varied accuracy, often exhibiting performance that may be insufficient for legal work. We then demonstrate that a lightly fine-tuned Llama 3 8B model vastly outperforms GPT-4 on almost all tasks, typically by double-digit percentage points. A few tens to hundreds of examples suffice to achieve high classification accuracy. Our work points to a viable alternative to the predominant practice of prompting commercial models. For concrete legal tasks with some available labeled data, researchers are better off using a specialized open-source model.
****
Yusu Qian,Hanrong Ye,Jean-Philippe Fauconnier,Peter Grasch,Yinfei Yang,Zhe Gan
MIA-Bench: Towards Better Instruction Following Evaluation of Multimodal LLMs
Effective evaluation of Multimodal Large Language Models (MLLMs) is essential for understanding their capabilities and limitations. In this paper, we introduce MIA-Bench, a benchmark designed to assess MLLMs’ ability to strictly adhere to complex instructions. Our benchmark comprises a diverse set of 400 image-prompt pairs, each crafted to challenge the models’ compliance with layered instructions in generating accurate and contextually appropriate responses. Evaluation results from a wide array of state-of-the-art MLLMs reveal significant variations in performance, highlighting areas for improvement in instruction fidelity. Additionally, we create extra training data and explore supervised fine-tuning and direct preference optimization to enhance the models’ ability to strictly follow instructions without compromising performance on other tasks. We hope this benchmark not only serves as a tool for measuring MLLM adherence to instructions, but also guides future developments in MLLM training methods.
****
Fang Kong,Jingqi Tang,Mingzhu Li,Pinyan Lu,John C.S. Lui,Shuai Li
Bandit Learning in Matching Markets with Indifference
A rich line of recent works studies how participants in matching markets learn their unknown preferences through iterative interactions with each other. The two sides of participants in the market can be respectively formulated as players and arms in the bandit problem. To ensure market stability, the objective is to minimize the stable regret of each player. Though existing works provide significant theoretical upper bounds for players' stable regret, the results heavily rel

y on the assumption that each participant has a strict preference ranking. Howev
er, in real applications, multiple candidates (e.g., workers in the labor market
 and students in school admission) usually demonstrate comparable performance le
vels, making it challenging for participants (e.g., employers and schools) to di
fferentiate and rank their preferences. To deal with the potential indifferent p
references, we propose an adaptive exploration algorithm based on arm-guided Gal
e-Shapley (AE-AGS). We show that its stable regret is of order $O(NK \log T / \D
elta^2)$, where N is the number of players, K the number of arms, T the to
tal time horizon, and Δ the minimum non-zero preference gap. Extensive ex
periments demonstrate the algorithm's effectiveness in handling such complex sit
uations and its consistent superiority over baselines.
****
Zhengqing Wang,Jiacheng Chen,Yasutaka Furukawa
PuzzleFusion++: Auto-agglomerative 3D Fracture Assembly by Denoise and Verify
This paper proposes a novel “auto-agglomerative” 3D fracture assembly method, Pu
zzleFusion++, resembling how humans solve challenging spatial puzzles. Starting
from individual fragments, the approach 1) aligns and merges fragments into larg
er groups akin to agglomerative clustering and 2) repeats the process iterativel
y in completing the assembly akin to auto-regressive methods. Concretely, a diff
usion model denoises the 6-DoF alignment parameters of the fragments simultaneou
sly,
and a transformer model verifies and merges pairwise alignments into larger ones
, whose process repeats iteratively. Extensive experiments on the Breaking Bad d
ataset show that PuzzleFusion++ outperforms all other state-of-the-art technique
s by significant margins across all metrics In particular by over 10% in part ac
curacy and 50% in Chamfer distance. We will release code and model.
****
Haofei Lu,Dongqi Han,Yifei Shen,Dongsheng Li
What Makes a Good Diffusion Planner for Decision Making?
Diffusion models have recently shown significant potential in solving decision-m
aking problems, particularly in generating behavior plans -- also known as diffu
sion planning. While numerous studies have demonstrated the impressive performan
ce of diffusion planning, the mechanisms behind the key components of a good dif
fusion planner remain unclear and the design choices are highly inconsistent in
existing studies. In this work, we address this issue through systematic empiric
al experiments on diffusion planning in an offline reinforcement learning (RL) s
etting, providing practical insights into the essential components of diffusion
planning. We trained and evaluated over 6,000 diffusion models, identifying the
critical components such as guided sampling, network architecture, action genera
tion and planning strategy. We revealed that some design choices opposite to the
 common practice in previous work in diffusion planning actually lead to better
performance, e.g., unconditional sampling with selection can be better than guid
ed sampling and Transformer outperforms U-Net as denoising network. Based on the
se insights, we suggest a simple yet strong diffusion planning baseline that ach
ieves state-of-the-art results on standard offline RL benchmarks. Code: https://
github.com/Josh00-Lu/DiffusionVeteran.
****
Tai Hoang,Huy Le,Philipp Becker,Vien Anh Ngo,Gerhard Neumann
Geometry-aware RL for Manipulation of Varying Shapes and Deformable Objects
Manipulating objects with varying geometries and deformable objects is a major c
hallenge in robotics. Tasks such as insertion with different objects or cloth ha
nging require precise control and effective modelling of complex dynamics. In th
is work, we frame this problem through the lens of a heterogeneous graph that co
mprises smaller sub-graphs, such as actuators and objects, accompanied by differ
ent edge types describing their interactions. This graph representation serves a
s a unified structure for both rigid and deformable objects tasks, and can be ex
tended further to tasks comprising multiple actuators. To evaluate this setup, w
e present a novel and challenging reinforcement learning benchmark, including ri
gid insertion of diverse objects, as well as rope and cloth manipulation with mu
ltiple end-effectors. These tasks present a large search space, as both the init

ial and target configurations are uniformly sampled in 3D space. To address this issue, we propose a novel graph-based policy model, dubbed Heterogeneous Equivariant Policy (HEPi), utilizing $SE(3)$ equivariant message passing networks as the main backbone to exploit the geometric symmetry. In addition, by modeling explicit heterogeneity, HEPi can outperform Transformer-based and non-heterogeneous equivariant policies in terms of average returns, sample efficiency, and generalization to unseen objects. Our project page is available at https://thobotics.github.io/hepi.

****

Robert Ganian,Mathis Rocton,Simon Wietheger

Training One-Dimensional Graph Neural Networks is NP-Hard

We initiate the study of the computational complexity of training graph neural networks (GNNs). We consider the classical node classification setting; there, the intractability of training multidimensonal GNNs immediately follows from known lower bounds for training classical neural networks (and holds even for trivial GNNs). However, one-dimensional GNNs form a crucial case of interest: the computational complexity of training such networks depends on both the graphical structure of the network and the properties of the involved activation and aggregation functions. As our main result, we establish the NP-hardness of training ReLU-activated one-dimensional GNNs via a highly non-trivial reduction. We complement this result with algorithmic upper bounds for the training problem in the ReLU-activated and linearly-activated settings.

****

Sujay Nagaraj,Yang Liu,Flavio Calmon,Berk Ustun

Regretful Decisions under Label Noise

Machine learning models are routinely used to support decisions that affect individuals – be it to screen a patient for a serious illness or to gauge their response to treatment. In these tasks, we are limited to learning models from datasets with noisy labels. In this paper, we study the instance-level impact of learning under label noise. We introduce a notion of regret for this regime which measures the number of unforeseen mistakes due to noisy labels. We show that standard approaches to learning under label noise can return models that perform well at a population level while subjecting individuals to a lottery of mistakes. We present a versatile approach to estimate the likelihood of mistakes at the individual level from a noisy dataset by training models over plausible realizations of datasets without label noise. This is supported by a comprehensive empirical study of label noise in clinical prediction tasks. Our results reveal how failure to anticipate mistakes can compromise model reliability and adoption, and demonstrate how we can address these challenges by anticipating and avoiding regretful decisions.

****

Guibin Zhang,Xiangguo Sun,Yanwei Yue,Chonghe Jiang,Kun Wang,Tianlong Chen,Shirui Pan

Graph Sparsification via Mixture of Graphs

Graph Neural Networks (GNNs) have demonstrated superior performance across various graph learning tasks but face significant computational challenges when applied to large-scale graphs. One effective approach to mitigate these challenges is graph sparsification, which involves removing non-essential edges to reduce computational overhead. However, previous graph sparsification methods often rely on a single global sparsity setting and uniform pruning criteria, failing to provide customized sparsification schemes for each node's complex local context.

In this paper, we introduce Mixture-of-Graphs (MoG), leveraging the concept of Mixture-of-Experts (MoE), to dynamically select tailored pruning solutions for each node. Specifically, MoG incorporates multiple sparsifier experts, each characterized by unique sparsity levels and pruning criteria, and selects the appropriate experts for each node. Subsequently, MoG performs a mixture of the sparse graphs produced by different experts on the Grassmann manifold to derive an optimal sparse graph. One notable property of MoG is its entirely local nature, as it depends on the specific circumstances of each individual node. Extensive experiments on four large-scale OGB datasets and two superpixel datasets, equipped with

five GNN backbones, demonstrate that MoG (I) identifies subgraphs at higher sparsity levels ($8.67\\%\sim 50.85\\%$), with performance equal to or better than the dense graph, (II) achieves $1.47-2.62\times$ speedup in GNN inference with negligible performance drop, and (III) boosts ``top-student'' GNN performance ($1.02\\%\uparrow$ on RevGNN+\textsc{ogbn-proteins} and $1.74\\%\\uparrow$ on DeeperGCN+\textsc{ogbg-ppa}). The source code is available at \url{https://github.com/yanweiyue/MoG}.
****
Mingyuan Fan,Zhanyi Hu,Fuyi Wang,Cen Chen
Bad-PFL: Exploiting Backdoor Attacks against Personalized Federated Learning
Data heterogeneity and backdoor attacks rank among the most significant challenges facing federated learning (FL). For data heterogeneity, personalized federated learning (PFL) enables each client to maintain a private personalized model to cater to client-specific knowledge. Meanwhile, vanilla FL has proven vulnerable to backdoor attacks. However, recent advancements in PFL community have demonstrated a potential immunity against such attacks. This paper explores this intersection further, revealing that existing federated backdoor attacks fail in PFL because backdoors about manually designed triggers struggle to survive in personalized models. To tackle this, we degisn Bad-PFL, which employs features from natural data as our trigger. As long as the model is trained on natural data, it inevitably embeds the backdoor associated with our trigger, ensuring its longevity in personalized models. Moreover, our trigger undergoes mutual reinforcement training with the model, further solidifying the backdoor's durability and enhancing attack effectiveness. The large-scale experiments across three benchmark datasets demonstrate the superior performance of Bad-PFL against various PFL methods, even when equipped with state-of-the-art defense mechanisms.
****
Alaa Khaddaj,Logan Engstrom,Aleksander Madry
Small-to-Large Generalization: Training Data Influences Models Consistently Across Scale
Choice of training data distribution greatly influences model behavior. Yet, in large-scale settings, precisely characterizing *how* changes in training data affects predictions is often difficult due to model training costs. Current practice is to instead extrapolate from scaled down, inexpensive-to-train proxy models. However, changes in data do not influence smaller and larger models identically. Therefore, understanding how choice of data affects large-scale models raises the question: how does training data distribution influence model behavior across compute scale? We find that small- and large-scale language model predictions (generally) *do* highly correlate across choice of training data. Equipped with these findings, we characterize how proxy scale affects effectiveness in two downstream proxy model applications: data attribution and dataset selection.
****
Yinkai Wang,Xiaohui Chen,Liping Liu,Soha Hassoun
MADGEN: Mass-Spec attends to De Novo Molecular generation
The annotation (assigning structural chemical identities) of MS/MS spectra remains a significant challenge due to the enormous molecular diversity in biological samples and the limited scope of reference databases. Currently, the vast majority of spectral measurements remain in the "dark chemical space" without structural annotations. To improve annotation, we propose MADGEN (Mass-spec Attends to De Novo Molecular GENeration), a scaffold-based method for de novo molecular structure generation guided by mass spectrometry data. MADGEN operates in two stages: scaffold retrieval and spectra-conditioned molecular generation starting with the scaffold. In the first stage, given an MS/MS spectrum, we formulate scaffold retrieval as a ranking problem and employ contrastive learning to align mass spectra with candidate molecular scaffolds. In the second stage, starting from the retrieved scaffold, we employ the MS/MS spectrum to guide an attention-based generative model to generate the final molecule. Our approach constrains the molecular generation search space, reducing its complexity and improving generation accuracy. We evaluate MADGEN on three datasets (NIST23, CANOPUS, and MassSpecG

ym) and evaluate MADGEN's performance with a predictive scaffold retriever and with an oracle retriever. We demonstrate the effectiveness of using attention to integrate spectral information throughout the generation process to achieve strong results with the oracle retriever.
****
Xiangming Gu,Tianyu Pang,Chao Du,Qian Liu,Fengzhuo Zhang,Cunxiao Du,Ye Wang,Min Lin
When Attention Sink Emerges in Language Models: An Empirical View
Auto-regressive language Models (LMs) assign significant attention to the first token, even if it is not semantically important, which is known as **attention sink**. This phenomenon has been widely adopted in applications such as streaming/long context generation, KV cache optimization, inference acceleration, model quantization, and others. Despite its widespread use, a deep understanding of attention sink in LMs is still lacking. In this work, we first demonstrate that attention sinks exist universally in auto-regressive LMs with various inputs, even in small models. Furthermore, attention sink is observed to emerge during the LM pre-training, motivating us to investigate how *optimization*, *data distribution*, *loss function*, and *model architecture* in LM pre-training influence its emergence. We highlight that attention sink emerges after effective optimization on sufficient training data. The sink position is highly correlated with the loss function and data distribution. Most importantly, we find that attention sink acts more like key biases, *storing extra attention scores*, which could be non-informative and not contribute to the value computation. We also observe that this phenomenon (at least partially) stems from tokens' inner dependence on attention scores as a result of softmax normalization. After relaxing such dependence by replacing softmax attention with other attention operations, such as sigmoid attention without normalization, attention sinks do not emerge in LMs up to 1B parameters. The code is available at https://github.com/sail-sg/Attention-Sink.
****
Yongyi Su,Yushu Li,Nanqing Liu,Kui Jia,Xulei Yang,Chuan-Sheng Foo,Xun Xu
On the Adversarial Risk of Test Time Adaptation: An Investigation into Realistic Test-Time Data Poisoning
Test-time adaptation (TTA) updates the model weights during the inference stage using testing data to enhance generalization. However, this practice exposes TTA to adversarial risks. Existing studies have shown that when TTA is updated with crafted adversarial test samples, also known as test-time poisoned data, the performance on benign samples can deteriorate. Nonetheless, the perceived adversarial risk may be overstated if the poisoned data is generated under overly strong assumptions. In this work, we first review realistic assumptions for test-time data poisoning, including white-box versus grey-box attacks, access to benign data, attack order, and more. We then propose an effective and realistic attack method that better produces poisoned samples without access to benign samples, and derive an effective in-distribution attack objective. We also design two TTA-aware attack objectives. Our benchmarks of existing attack methods reveal that the TTA methods are more robust than previously believed. In addition, we analyze effective defense strategies to help develop adversarially robust TTA methods. The source code is available at https://github.com/Gorilla-Lab-SCUT/RTTDP.
****
Yinlam Chow,Guy Tennenholtz,Izzeddin Gur,Vincent Zhuang,Bo Dai,Aviral Kumar,Rishabh Agarwal,Sridhar Thiagarajan,Craig Boutilier,Aleksandra Faust
Inference-Aware Fine-Tuning for Best-of-N Sampling in Large Language Models
Recent studies indicate that effectively utilizing inference-time compute is crucial for attaining good performance from large language models (LLMs). Specifically, the Best-of-N (BoN) inference strategy, where an LLM generates multiple responses and a verifier selects the best, has shown strong empirical performance. Motivated by this, we develop a novel inference-aware fine-tuning paradigm, which encompasses the BoN-aware inference framework as a special case. We devise the first imitation learning and reinforcement learning (RL) methods for fine-tuning LLMs using BoN, overcoming the challenging, non-differentiable argmax operator in BoN. We empirically demonstrate that our BoN-aware models implicitly learn a

per-example "meta-strategy", which interleaves best responses with more diverse responses that might be better suited to a test-time input—a process reminiscent of the exploration-exploitation trade-off in RL. Our experiments demonstrate the effectiveness of BoN-aware fine-tuning in terms of improved performance and inference-time compute. In particular, we show that our methods improve the BoN performance of Gemma 2B on Hendrycks MATH from 26.8% to 30.8%, and Pass@K from 60% to 67%.

****

Xiangyu Wu,Feng Yu,Yang Yang,Qing-Guo Chen,Jianfeng Lu

Multi-Label Test-Time Adaptation with Bound Entropy Minimization

Mainstream test-time adaptation (TTA) techniques endeavor to mitigate distribution shifts via entropy minimization for multi-class classification, inherently increasing the probability of the most confident class. However, when encountering multi-label instances, the primary challenge stems from the varying number of labels per image, and prioritizing only the highest probability class inevitably undermines the adaptation of other positive labels. To address this issue, we investigate TTA within multi-label scenario (ML--TTA), developing Bound Entropy Minimization (BEM) objective to simultaneously increase the confidence of multiple top predicted labels. Specifically, to determine the number of labels for each augmented view, we retrieve a paired caption with yielded textual labels for that view. These labels are allocated to both the view and caption, called weak label set and strong label set with the same size k. Following this, the proposed BEM considers the highest top-k predicted labels from view and caption as a single entity, respectively, learning both view and caption prompts concurrently. By binding top-k predicted labels, BEM overcomes the limitation of vanilla entropy minimization, which exclusively optimizes the most confident class. Across the MSCOCO, VOC, and NUSWIDE multi-label datasets, our ML--TTA framework equipped with BEM exhibits superior performance compared to the latest SOTA methods, across various model architectures, prompt initialization, and varying label scenarios. The code is available at https://github.com/Jinx630/ML-TTA.

****

Brandon Trabucco,Max A Gurinas,Kyle Doherty,Russ Salakhutdinov

Wayward Concepts In Multimodal Models

Large multimodal models such as Stable Diffusion can generate, detect, and classify new visual concepts after optimizing just the prompt. How are prompt embeddings for visual concepts found by prompt tuning methods different from typical discrete prompts? We conduct a large-scale analysis on three state-of-the-art models in text-to-image generation, open-set object detection, and zero-shot classification, and find that prompts optimized to represent new visual concepts are akin to an adversarial attack on the text encoder. Across 4,800 new embeddings trained for 40 diverse visual concepts on four standard datasets, we find perturbations within an ϵ-ball to any prompt that reprogram models to generate, detect, and classify arbitrary subjects. These perturbations target the final-layers in text encoders, and steer pooling tokens towards the subject. We explore the transferability of these prompts, and find that perturbations reprogramming multimodal models are initialization-specific, and model-specific. Code for reproducing our work is available at the following site: https://wayward-concepts.github.io.

****

Biao Zhang,Peter Wonka

LaGeM: A Large Geometry Model for 3D Representation Learning and Diffusion

This paper introduces a novel hierarchical autoencoder that maps 3D models into a highly compressed latent space. The hierarchical autoencoder is specifically designed to tackle the challenges arising from large-scale datasets and generative modeling using diffusion. Different from previous approaches that only work on a regular image or volume grid, our hierarchical autoencoder operates on unordered sets of vectors. Each level of the autoencoder controls different geometric levels of detail. We show that the model can be used to represent a wide range of 3D models while faithfully representing high-resolution geometry details. The training of the new architecture takes 0.70x time and 0.58x memory compared to t

he baseline.
We also explore how the new representation can be used for generative modeling. Specifically, we propose a cascaded diffusion framework where each stage is conditioned on the previous stage. Our design extends existing cascaded designs for image and volume grids to vector sets.
****
Chaoyi Zhu,Jiayi Tang,Jeroen M. Galjaard,Pin-Yu Chen,Robert Birke,Cornelis Bos,Lydia Y. Chen
TabWak: A Watermark for Tabular Diffusion Models
Synthetic data offers alternatives for data augmentation and sharing. Till date, it remains unknown how to use watermarking techniques to trace and audit synthetic tables generated by tabular diffusion models to mitigate potential misuses. In this paper, we design TabWak, the first watermarking method to embed invisible signatures that control the sampling of Gaussian latent codes used to synthesize table rows via the diffusion backbone. TabWak has two key features. Different from existing image watermarking techniques, TabWak uses self-cloning and shuffling to embed the secret key in positional information of random seeds that control the Gaussian latents, allowing to use different seeds at each row for high inter-row diversity and enabling row-wise detectability. To further boost the robustness of watermark detection against post-editing attacks, TabWak uses a valid-bit mechanism that focuses on the tail of the latent code distribution for superior noise resilience. We provide theoretical guarantees on the row diversity and effectiveness of detectability. We evaluate TabWak on five datasets against baselines to show that the quality of watermarked tables remains nearly indistinguishable from non-watermarked tables while achieving high detectability in the presence of strong post-editing attacks, with a 100% true positive rate at a 0.1% false positive rate on synthetic tables with fewer than 300 rows. Our code is available at the following anonymized repository https://github.com/chaoyitud/TabWak.
****
Mingxin Huang,Yuliang Liu,Dingkang Liang,Lianwen Jin,Xiang Bai
Mini-Monkey: Alleviating the Semantic Sawtooth Effect for Lightweight MLLMs via Complementary Image Pyramid
Recently, scaling images to high resolution has received much attention in multimodal large language models (MLLMs). Most existing practices adopt a sliding-window-style cropping strategy to adapt to resolution increase. Such a cropping strategy, however, can easily cut off objects and connected regions, which introduces semantic discontinuity and therefore impedes MLLMs from recognizing small or irregularly shaped objects or text, leading to a phenomenon we call the semantic sawtooth effect. This effect is particularly evident in lightweight MLLMs. To address this issue, we introduce a Complementary Image Pyramid (CIP), a simple, effective, and plug-and-play solution designed to mitigate semantic discontinuity during high-resolution image processing. In particular, CIP dynamically constructs an image pyramid to provide complementary semantic information for the cropping-based MLLMs, enabling it rich acquire semantics at all levels. Furthermore, we introduce a Scale Compression Mechanism (SCM) to reduce the additional computational overhead by compressing the redundant visual tokens. Our experiments demonstrate that CIP can consistently enhance the performance across diverse architectures (e.g., MiniCPM-V-2, InternVL2, and LLaVA-OneVision), various model capacity (1B$\rightarrow$8B), and different usage configurations (training-free and fine-tuning). Leveraging the proposed CIP and SCM, we introduce a lightweight MLLM, Mini-Monkey, which achieves remarkable performance in both general multimodal understanding and document understanding. On the OCRBench, the 2B-version Mini-Monkey even surpasses the 8B model InternVL2-8B by 12 score. Additionally, training Mini-Monkey is cheap, requiring only eight RTX 3090 GPUs. Code and models are available at
https://github.com/Yuliang-Liu/Monkey.
****
Ziru Chen,Shijie Chen,Yuting Ning,Qianheng Zhang,Boshi Wang,Botao Yu,Yifei Li,Zeyi Liao,Chen Wei,Zitong Lu,Vishal Dey,Mingyi Xue,Frazier N. Baker,Benjamin Burns

,Daniel Adu-Ampratwum,Xuhui Huang,Xia Ning,Song Gao,Yu Su,Huan Sun
ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery
The advancements of language language models (LLMs) have piqued growing interest in developing LLM-based language agents to automate scientific discovery end-to-end, which has sparked both excitement and skepticism about the true capabilities of such agents. In this work, we argue that for an agent to fully automate scientific discovery, it must be able to complete all essential tasks in the workflow. Thus, we call for rigorous assessment of agents on individual tasks in a scientific workflow before making bold claims on end-to-end automation. To this end, we present ScienceAgentBench, a new benchmark for evaluating language agents for data-driven scientific discovery. To ensure the scientific authenticity and real-world relevance of our benchmark, we extract 102 tasks from 44 peer-reviewed publications in four disciplines and engage nine subject matter experts to validate them. We unify the target output for every task to a self-contained Python program file and employ an array of evaluation metrics to examine the generated programs, execution results, and costs. Each task goes through multiple rounds of manual validation by annotators and subject matter experts to ensure its annotation quality and scientific plausibility. We also propose two effective strategies to mitigate data contamination concerns. Using our benchmark, we evaluate five open-weight and proprietary LLMs, each with three frameworks: direct prompting, OpenHands, and self-debug. Given three attempts for each task, the best-performing agent can only solve 32.4% of the tasks independently and 34.3% with expert-provided knowledge. These results underscore the limited capacities of current language agents in generating code for data-driven discovery, let alone end-to-end automation for scientific research.
****
Alexis Huet,Zied Ben Houidi,Dario Rossi
Episodic Memories Generation and Evaluation Benchmark for Large Language Models
Episodic memory -- the ability to recall specific events grounded in time and space -- is a cornerstone of human cognition, enabling not only coherent storytelling, but also planning and decision-making. Despite their remarkable capabilities, Large Language Models (LLMs) lack a robust mechanism for episodic memory: we argue that integrating episodic memory capabilities into LLM is essential for advancing AI towards human-like cognition, increasing their potential to reason consistently and ground their output in real-world episodic events, hence avoiding confabulations. To address this challenge, we introduce a comprehensive framework to model and evaluate LLM episodic memory capabilities. Drawing inspiration from cognitive science, we develop a structured approach to represent episodic events, encapsulating temporal and spatial contexts, involved entities, and detailed descriptions. We synthesize a unique episodic memory benchmark, free from contamination, and release open source code and datasets to assess LLM performance across various recall and episodic reasoning tasks. Our evaluation of state-of-the-art models, including GPT-4 and Claude variants, Llama 3.1, and o1-mini, reveals that even the most advanced LLMs struggle with episodic memory tasks, particularly when dealing with multiple related events or complex spatio-temporal relationships -- even in contexts as short as 10k-100k tokens.
****
Sirui Li,Janardhan Kulkarni,Ishai Menache,Cathy Wu,Beibin Li
Towards Foundation Models for Mixed Integer Linear Programming
Mixed Integer Linear Programming (MILP) is essential for modeling complex decision-making problems but faces challenges in computational tractability and interpretability. Current deep learning approaches for MILP focus on specific problem classes and do not generalize to unseen classes. To address this shortcoming, we take a foundation model training approach, where we train a single deep learning model on a diverse set of MILP problems to generalize across problem classes. As existing datasets for MILP lack diversity and volume, we introduce MILP-Evolve, a novel LLM-based evolutionary framework that is capable of generating a large set of diverse MILP classes with an unlimited amount of instances. We study our methodology on three key learning tasks that capture diverse aspects of MILP:

(1) integrality gap prediction, (2) learning to branch, and (3) a new task of al
igning MILP instances with natural language descriptions. Our empirical results
show that models trained on the data generated by MILP-Evolve achieve significan
t improvements on unseen problems, including MIPLIB benchmarks. Our work highlig
hts the potential of moving towards a foundation model approach for MILP that ca
n generalize to a broad range of MILP problem classes. Our code and data are pub
licly available at https://github.com/microsoft/OptiGuide.
****

Xiaoyu Cui,Weixing Chen,Jiandong Su

A Multiscale Frequency Domain Causal Framework for Enhanced Pathological Analysi
s

Multiple Instance Learning (MIL) in digital pathology Whole Slide Image (WSI) an
alysis has shown significant progress. However, due to data bias and unobservabl
e confounders, this paradigm still faces challenges in terms of performance and
interpretability. Existing MIL methods might identify patches that do not have t
rue diagnostic significance, leading to false correlations, and experience diffi
culties in integrating multi-scale features and handling unobservable confounder
s. To address these issues, we propose a new Multi-Scale Frequency Domain Causal
 framework (MFC). This framework employs an adaptive memory module to estimate t
he overall data distribution through multi-scale frequency-domain information du
ring training and simulates causal interventions based on this distribution to m
itigate confounders in pathological diagnosis tasks. The framework integrates th
e Multi-scale Spatial Representation Module (MSRM), Frequency Domain Structure R
epresentation Module (FSRM), and Causal Memory Intervention Module (CMIM) to enh
ance the model's performance and interpretability. Furthermore, the plug-and-pla
y nature of this framework allows it to be broadly applied across various models
. Experimental results on Camelyon16 and TCGA-NSCLC dataset show that, compared
to previous work, our method has significantly improved accuracy and generalizat
ion ability, providing a new theoretical perspective for medical image analysis
and potentially advancing the field further. The code will be released at https:
//github.com/WissingChen/MFC-MIL.
****

Yilun Li,Miaomiao Cheng,Xu Han,Wei Song

Divergence-enhanced Knowledge-guided Context Optimization for Visual-Language Pr
ompt Tuning

Prompt tuning vision-language models like CLIP has shown great potential in lear
ning transferable representations for various downstream tasks. The main issue i
s how to mitigate the over-fitting problem on downstream tasks with limited trai
ning samples. While knowledge-guided context optimization has been proposed by c
onstructing consistency constraints to handle catastrophic forgetting in the pre
-trained backbone, it also introduces a bias toward pre-training.
This paper proposes a novel and simple Divergence-enhanced Knowledge-guided Prom
pt Tuning (DeKg) method to address this issue.
The key insight is that the bias toward pre-training can be alleviated by encour
aging the independence between the learnable and the crafted prompt. Specificall
y, DeKg employs the Hilbert-Schmidt Independence Criterion (HSIC) to regularize
the learnable prompts, thereby reducing their dependence on prior general knowle
dge, and enabling divergence induced by target knowledge.
Comprehensive evaluations demonstrate that DeKg serves as a plug-and-play module
 can seamlessly integrate with existing knowledge-guided context optimization me
thods and achieves superior performance in three challenging benchmarks. We make
 our code available at https://github.com/cnunlp/DeKg.
****

Zhong Zheng,Haochen Zhang,Lingzhou Xue

Gap-Dependent Bounds for Q-Learning using Reference-Advantage Decomposition

We study the gap-dependent bounds of two important algorithms for on-policy Q-
learning for finite-horizon episodic tabular Markov Decision Processes (MDPs): U
CB-Advantage (Zhang et al. 2020) and Q-EarlySettled-Advantage (Li et al. 2021).
UCB-Advantage and Q-EarlySettled-Advantage improve upon the results based on Hoe
ffding-type bonuses and achieve the {almost optimal} $\sqrt{T}$-type regret boun

d in the worst-case scenario, where T is the total number of steps. However, t
he benign structures of the MDPs such as a strictly positive suboptimality gap c
an significantly improve the regret. While gap-dependent regret bounds have been
 obtained for Q-learning with Hoeffding-type bonuses, it remains an open quest
ion to establish gap-dependent regret bounds for Q-learning using variance est
imators in their bonuses and reference-advantage decomposition for variance redu
ction. We develop a novel error decomposition
framework to prove gap-dependent regret bounds of UCB-Advantage and Q-EarlySettl
ed-Advantage that are logarithmic in T and improve upon existing ones for Q-
learning algorithms. Moreover, we establish the gap-dependent bound for the poli
cy switching cost of UCB-Advantage and improve that under the worst-case MDPs. T
o our knowledge, this paper presents the first gap-dependent regret analysis for
 Q-learning using variance estimators and reference-advantage decomposition an
d also provides the first gap-dependent analysis on policy switching cost for $Q
$-learning.
****
Jun Shern Chan,Neil Chowdhury,Oliver Jaffe,James Aung,Dane Sherburn,Evan Mays,Gi
ulio Starace,Kevin Liu,Leon Maksin,Tejal Patwardhan,Aleksander Madry,Lilian Weng
MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering
We introduce MLE-bench, a benchmark for measuring how well AI agents perform at
machine learning engineering. To this end, we curate 75 ML engineering-related c
ompetitions from Kaggle, creating a diverse set of challenging tasks that test r
eal-world ML engineering skills such as training models, preparing datasets, and
 running experiments. We establish human baselines for each competition using Ka
ggle's publicly available leaderboards. We use open-source agent scaffolds to ev
aluate several frontier language models on our benchmark, finding that the best-
performing setup — OpenAI's o1-preview with AIDE scaffolding — achieves at least
 the level of a Kaggle bronze medal in 16.9% of competitions. In addition to our
 main results, we investigate various forms of resource-scaling for AI agents an
d the impact of contamination from pre-training. We open-source our benchmark co
de https://github.com/openai/mle-bench to facilitate future research in understa
nding the ML engineering capabilities of AI agents.
****
Mehul Damani,Idan Shenfeld,Andi Peng,Andreea Bobu,Jacob Andreas
Learning How Hard to Think: Input-Adaptive Allocation of LM Computation
Computationally intensive decoding procedures---including search, reranking, and
 self-critique---can improve the quality of language model (LM) outputs in probl
ems spanning code generation, numerical reasoning, and dialog.
Existing work typically applies the same decoding procedure for every input to a
n LM. But not all inputs require the same amount of computation to process. Can
we allocate decoding computation adaptively, using more resources to answer ques
tions whose answers will be harder to compute? We present an approach that predi
cts the distribution of rewards given an input and computation budget, then allo
cates additional computation to inputs for which it is predicted to be most usef
ul. We apply this approach in two decoding procedures: first, an adaptive best-o
f-k procedure that dynamically selects the number of samples to generate as in
put to a reranker; second, a routing procedure that dynamically responds to a qu
ery using a decoding procedure that is expensive but accurate, or one that is ch
eaper but less capable. Across a suite of programming, mathematics, and dialog t
asks, we show that accurate computation-allocation procedures can be learned, an
d reduce computation by up to 50% at no cost to quality.
****
Felix Koulischer,Johannes Deleu,Gabriel Raya,Thomas Demeester,Luca Ambrogioni
Dynamic Negative Guidance of Diffusion Models
Negative Prompting (NP) is widely utilized in diffusion models, particularly in
text-to-image applications, to prevent the generation of undesired features. In
this paper, we show that conventional NP is limited by the assumption of a const
ant guidance scale, which may lead to highly suboptimal results, or even complet
e failure, due to the non-stationarity and state-dependence of the reverse proce
ss. Based on this analysis, we derive a principled technique called ***D**ynamic

Negative **G**uidance*, which relies on a near-optimal time and state depen
dent modulation of the guidance without requiring additional training. Unlike NP
, negative guidance requires estimating the posterior class probability during t
he denoising process, which is achieved with limited additional computational ov
erhead by tracking the discrete Markov Chain during the generative process. We e
valuate the performance of DNG class-removal on MNIST and CIFAR10, where we show
 that DNG leads to higher safety, preservation of class balance and image qualit
y when compared with baseline methods. Furthermore, we show that it is possible
to use DNG with Stable Diffusion to obtain more accurate and less invasive guida
nce than NP.
****
Christian Klötergens,Vijaya Krishna Yalavarthi,Randolf Scholz,Maximilian Stubbem
ann,Stefan Born,Lars Schmidt-Thieme
Physiome-ODE: A Benchmark for Irregularly Sampled Multivariate Time-Series Forec
asting Based on Biological ODEs
State-of-the-art methods for forecasting irregularly sampled time series with mi
ssing values predominantly rely on just four datasets and a few small toy exampl
es for evaluation. While ordinary differential equations (ODE) are the prevalent
 models in science and engineering, a baseline model that forecasts a constant v
alue outperforms ODE-based models from the last five years on three of these exi
sting datasets. This unintuitive finding hampers further research on ODE-based m
odels, a more plausible model family.
In this paper, we develop a methodology to generate irregularly sampled multivar
iate time series (IMTS) datasets from ordinary differential
equations and to select challenging instances via rejection sampling. Using this
 methodology, we create Physiome-ODE, a large and sophisticated benchmark of IMT
S datasets consisting of 50 individual datasets, derived from real-world ordinar
y differential equations from research in biology. Physiome-ODE is the first ben
chmark for IMTS forecasting that we are aware of and an order of magnitude large
r than the current evaluation setting of four datasets. Using our benchmark Phys
iome-ODE, we show qualitatively completely different results than those derived
from the current four datasets: on Physiome-ODE ODE-based models can play to the
ir strength and our benchmark can differentiate in a meaningful way between diff
erent IMTS forecasting models. This way, we expect to give a new impulse to rese
arch on ODE-based time series modeling.
****
Guoxuan Xia,Olivier Laurent,Gianni Franchi,Christos-Savvas Bouganis
Towards Understanding Why Label Smoothing Degrades Selective Classification and
How to Fix It
Label smoothing (LS) is a popular regularisation method for training neural netw
orks as it is effective in improving test accuracy and is simple to implement. '
'Hard'' one-hot labels are ''smoothed'' by uniformly distributing probability ma
ss to other classes, reducing overfitting. Prior work has shown that in some cas
es *LS can degrade selective classification (SC)* -- where the aim is to reject
misclassifications using a model's uncertainty. In this work, we first demonstra
te empirically across an extended range of large-scale tasks and architectures t
hat LS *consistently* degrades SC.
We then address a gap in existing knowledge, providing an *explanation* for this
 behaviour by analysing logit-level gradients: LS degrades the uncertainty rank
ordering of correct vs incorrect predictions by regularising the max logit *more
* when a prediction is likely to be correct, and *less* when it is likely to be
wrong.
    This elucidates previously reported experimental results where strong classi
fiers underperform in SC.
    We then demonstrate the empirical effectiveness of post-hoc *logit normalisa
tion* for recovering lost SC performance caused by LS. Furthermore, linking back
 to our gradient analysis, we again provide an explanation for why such normalis
ation is effective.
****
Yan Sun,Stanley Kok

Unsupervised Multiple Kernel Learning for Graphs via Ordinality Preservation
Learning effective graph similarities is crucial for tasks like clustering, yet selecting the optimal kernel to evaluate such similarities in unsupervised settings remains a major challenge. Despite the development of various graph kernels, determining the most appropriate one for a specific task is particularly difficult in the absence of labeled data. Existing methods often struggle to handle the complex structure of graph data and rely on heuristic approaches that fail to adequately capture the global relationships between graphs. To overcome these limitations, we propose Unsupervised Multiple Kernel Learning for Graphs (UMKL-G), a model that combines multiple graph kernels without requiring labels or predefined local neighbors. Our approach preserves the topology of the data by maintaining ordinal relationships among graphs through a probability simplex, allowing for a unified and adaptive kernel learning process. We provide theoretical guarantees on the stability, robustness, and generalization of our method. Empirical results demonstrate that UMKL-G outperforms individual kernels and other state-of-the-art methods, offering a robust solution for unsupervised graph analysis.
****
Wassim Bouaziz,Nicolas Usunier,El-Mahdi El-Mhamdi
Data Taggants: Dataset Ownership Verification Via Harmless Targeted Data Poisoning
Dataset ownership verification, the process of determining if a dataset is used in a model's training data, is necessary for detecting unauthorized data usage and data contamination.
Existing approaches, such as backdoor watermarking, rely on inducing a detectable behavior into the trained model on a part of the data distribution.
However, these approaches have limitations, as they can be harmful to the model's performances or require unpractical access to the model's internals.
Most importantly, previous approaches lack guarantee against false positives.\
This paper introduces *data taggants*, a novel non-backdoor dataset ownership verification technique.
Our method uses pairs of out-of-distribution samples and random labels as secret *keys*, and leverages clean-label targeted data poisoning to subtly alter a dataset, so that models trained on it respond to the key samples with the corresponding key labels.
The keys are built as to allow for statistical certificates with black-box access only to the model.\
We validate our approach through comprehensive and realistic experiments on ImageNet1k using ViT and ResNet models with state-of-the-art training recipes.
Our findings demonstrate that data taggants can reliably detect models trained on the protected dataset with high confidence, without compromising validation accuracy, and show their superiority over backdoor watermarking.
We demonstrate the stealthiness and robustness of our method
% shows to be stealthy and robust
against various defense mechanisms.
****
Hayk Manukyan,Andranik Sargsyan,Barsegh Atanyan,Zhangyang Wang,Shant Navasardyan,Humphrey Shi
HD-Painter: High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models
Recent progress in text-guided image inpainting, based on the unprecedented success of text-to-image diffusion models, has led to exceptionally realistic and visually plausible results. However, there is still significant potential for improvement in current text-to-image inpainting models, particularly in better aligning the inpainted area with user prompts. Therefore, we introduce $\textit{HD-Painter}$, a $\textbf{training-free}$ approach that $\textbf{accurately follows prompts}$. To this end, we design the $\textit{Prompt-Aware Introverted Attention (PAIntA)}$ layer enhancing self-attention scores by prompt information resulting in better text aligned generations. To further improve the prompt coherence we introduce the $\textit{Reweighting Attention Score Guidance (RASG)}$ mechanism seamlessly integrating a post-hoc sampling strategy into the general form of DDIM

to prevent out-of-distribution latent shifts. Our experiments demonstrate that HD-Painter surpasses existing state-of-the-art approaches quantitatively and qualitatively across multiple metrics and a user study. Code is publicly available at: https://github.com/Picsart-AI-Research/HD-Painter
****
Minh-Quan Le,Gaurav Mittal,Tianjian Meng,A S M Iftekhar,Vishwas Suryanarayanan,Barun Patra,Dimitris Samaras,Mei Chen
Hummingbird: High Fidelity Image Generation via Multimodal Context Alignment
While diffusion models are powerful in generating high-quality, diverse synthetic data for object-centric tasks, existing methods struggle with scene-aware tasks such as Visual Question Answering (VQA) and Human-Object Interaction (HOI) Reasoning, where it is critical to preserve scene attributes in generated images consistent with a multimodal context, i.e. a reference image with accompanying text guidance query. To address this, we introduce **Hummingbird**, the first diffusion-based image generator which, given a multimodal context, generates highly diverse images w.r.t. the reference image while ensuring high fidelity by accurately preserving scene attributes, such as object interactions and spatial relationships from the text guidance. Hummingbird employs a novel Multimodal Context Evaluator that simultaneously optimizes our formulated Global Semantic and Fine-grained Consistency Rewards to ensure generated images preserve the scene attributes of reference images in relation to the text guidance while maintaining diversity. As the first model to address the task of maintaining both diversity and fidelity given a multimodal context, we introduce a new benchmark formulation incorporating MME Perception and Bongard HOI datasets. Benchmark experiments show Hummingbird outperforms all existing methods by achieving superior fidelity while maintaining diversity, validating Hummingbird's potential as a robust multimodal context-aligned image generator in complex visual tasks. Project page: https://roar-ai.github.io/hummingbird
****
Shange Tang,Jiayun Wu,Jianqing Fan,Chi Jin
Benign Overfitting in Out-of-Distribution Generalization of Linear Models
Benign overfitting refers to the phenomenon where an over-parameterized model fits the training data perfectly, including noise in the data, but still generalizes well to the unseen test data. While prior work provides some theoretical understanding of this phenomenon under the in-distribution setup, modern machine learning often operates in a more challenging Out-of-Distribution (OOD) regime, where the target (test) distribution can be rather different from the source (training) distribution. In this work, we take an initial step towards understanding benign overfitting in the OOD regime by focusing on the basic setup of over-parameterized linear models under covariate shift. We provide non-asymptotic guarantees proving that benign overfitting occurs in standard ridge regression, even under the OOD regime when the target covariance satisfies certain structural conditions. We identify several vital quantities relating to source and target covariance, which govern the performance of OOD generalization. Our result is sharp, which provably recovers prior in-distribution benign overfitting guarantee (Tsigler & Bartlett, 2023), as well as under-parameterized OOD guarantee (Ge et al., 2024) when specializing to each setup. Moreover, we also present theoretical results for a more general family of target covariance matrix, where standard ridge regression only achieves a slow statistical rate of $\mathcal{O}(1/\sqrt{n})$ for the excess risk, while Principal Component Regression (PCR) is guaranteed to achieve the fast rate $\mathcal{O}(1/n)$, where n is the number of samples.
****
Yangxuan Zhou,Sha Zhao,Jiquan Wang,Haiteng Jiang,Shijian Li,Tao Li,Gang Pan
BrainUICL: An Unsupervised Individual Continual Learning Framework for EEG Applications
Electroencephalography (EEG) is a non-invasive brain-computer interface technology used for recording brain electrical activity. It plays an important role in human life and has been widely uesd in real life, including sleep staging, emotion recognition, and motor imagery. However, existing EEG-related models cannot be

well applied in practice, especially in clinical settings, where new patients w
ith individual discrepancies appear every day. Such EEG-based model trained on f
ixed datasets cannot generalize well to the continual flow of numerous unseen su
bjects in real-world scenarios. This limitation can be addressed through continu
al learning (CL), wherein the CL model can continuously learn and advance over t
ime. Inspired by CL, we introduce a novel Unsupervised Individual Continual Lear
ning paradigm for handling this issue in practice. We propose the BrainUICL fram
ework, which enables the EEG-based model to continuously adapt to the incoming n
ew subjects. Simultaneously, BrainUICL helps the model absorb new knowledge duri
ng each adaptation, thereby advancing its generalization ability for all unseen
subjects. The effectiveness of the proposed BrainUICL has been evaluated on thre
e different mainstream EEG tasks. The BrainUICL can effectively balance both the
plasticity and stability during CL, achieving better plasticity on new individu
als and better stability across all the unseen individuals, which holds signific
ance in a practical setting.
****
Pei Wang,Yanan Wu,Noah Wang,Jiaheng Liu,Xiaoshuai Song,Z.Y. Peng,Ken Deng,Chench
en Zhang,JiakaiWang,Junran Peng,Ge Zhang,Hangyu Guo,Zhaoxiang Zhang,Wenbo Su,Bo
Zheng
MTU-Bench: A Multi-granularity Tool-Use Benchmark for Large Language Models
Large Language Models (LLMs) have displayed massive improvements in reason- ing
and decision-making skills and can hold natural conversations with users. Recent
ly, many tool-use benchmark datasets have been proposed. However, existing datas
ets have the following limitations: (1). Insufficient evaluation scenarios (e.g.
, only cover limited tool-use scenes). (2). Extensive evaluation costs (e.g., GP
T API costs). To address these limitations, in this work, we propose a multi-gra
nularity tool-use benchmark for large language models called MTU-Bench. For the
"multi-granularity" property, our MTU-Bench covers five tool usage scenes (i.e.,
single-turn and single-tool, single-turn and multiple-tool, multiple-turn and s
ingle-tool, multiple-turn and multiple-tool, and out-of-distribution tasks). Bes
ides, all evaluation metrics of our MTU-Bench are based on the prediction result
s and the ground truth without using any GPT or human evaluation metrics. Moreov
er, our MTU-Bench is collected by transforming existing high-quality datasets to
simulate real-world tool usage scenarios, and we also propose an instruction da
taset called MTU-Instruct data to enhance the tool-use abilities of existing LLM
s. Comprehensive experimental results demonstrate the effectiveness of our MTU-B
ench.
****
Jiuding Sun,Jing Huang,Sidharth Baskaran,Karel D'Oosterlinck,Christopher Potts,M
ichael Sklar,Atticus Geiger
HyperDAS: Towards Automating Mechanistic Interpretability with Hypernetworks
Mechanistic interpretability has made great strides in identifying neural networ
k features (e.g., directions in hidden activation space) that mediate concepts (
e.g., *the birth year of a Nobel laureate*) and enable predictable manipulation.
Distributed alignment search (DAS) leverages supervision from counterfactual da
ta to learn concept features within hidden states, but DAS assumes we can afford
to conduct a brute force search over potential feature locations. To address th
is, we present HyperDAS, a transformer-based hypernetwork architecture that (1)
automatically locates the token-positions of the residual stream that a concept
is realized in and (2) learns features of those residual stream vectors for the
concept. In experiments with Llama3-8B, HyperDAS achieves state-of-the-art perfo
rmance on the RAVEL benchmark for disentangling concepts in hidden states. In ad
dition, we review the design decisions we made to mitigate the concern that Hype
rDAS (like all powerful interpretabilty methods) might inject new information in
to the target model rather than faithfully interpreting it.
****
Farzad Farhadzadeh,Debasmit Das,Shubhankar Borse,Fatih Porikli
LoRA-X: Bridging Foundation Models with Training-Free Cross-Model Adaptation
The rising popularity of large foundation models has led to a heightened demand
for parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA),

which offer performance comparable to full model fine-tuning while requiring only a few additional parameters tailored to the specific base model. When such base models are deprecated and replaced, all associated LoRA modules must be retrained, requiring access to either the original training data or a substantial amount of synthetic data that mirrors the original distribution. However, the original data is often inaccessible due to privacy or licensing issues, and generating synthetic data may be impractical and insufficiently representative. These factors complicate the fine-tuning process considerably. To address this challenge, we introduce a new adapter, Cross-Model Low-Rank Adaptation (LoRA-X), which enables the training-free transfer of LoRA parameters across source and target models, eliminating the need for original or synthetic training data. Our approach imposes the adapter to operate within the subspace of the source base model. This constraint is necessary because our prior knowledge of the target model is limited to its weights, and the criteria for ensuring the adapter's transferability are restricted to the target base model's weights and subspace. To facilitate the transfer of LoRA parameters of the source model to a target model, we employ the adapter only in the layers of the target model that exhibit an acceptable level of subspace similarity. Our extensive experiments demonstrate the effectiveness of LoRA-X for text-to-image generation, including Stable Diffusion v1.5 and Stable Diffusion XL.
****

Arth Shukla,Stone Tao,Hao Su
ManiSkill-HAB: A Benchmark for Low-Level Manipulation in Home Rearrangement Tasks
High-quality benchmarks are the foundation for embodied AI research, enabling significant advancements in long-horizon navigation, manipulation and rearrangement tasks. However, as frontier tasks in robotics get more advanced, they require faster simulation speed, more intricate test environments, and larger demonstration datasets. To this end, we present MS-HAB, a holistic benchmark for low-level manipulation and in-home object rearrangement. First, we provide a GPU-accelerated implementation of the Home Assistant Benchmark (HAB). We support realistic low-level control and achieve over 3x the speed of prior magical grasp implementations at a fraction of the GPU memory usage. Second, we train extensive reinforcement learning (RL) and imitation learning (IL) baselines for future work to compare against. Finally, we develop a rule-based trajectory filtering system to sample specific demonstrations from our RL policies which match predefined criteria for robot behavior and safety. Combining demonstration filtering with our fast environments enables efficient, controlled data generation at scale.
****

Luofeng Liao,Christian Kroer,Sergei Leonenkov,Okke Schrijvers,Liang Shi,Nicolas Stier Moses,Congshan Zhang
Interference Among First-Price Pacing Equilibria: A Bias and Variance Analysis
A/B testing is widely used in the internet industry. For online marketplaces (such as advertising markets), standard approaches to A/B testing may lead to biased results when buyers have budget constraints, as budget consumption in one arm of the experiment impacts performance of the other arm.
This is often addressed using a budget-split design. Yet such splitting may degrade statistical performance as budgets become too small in each arm.
We propose a parallel budget-controlled A/B testing design where we use market segmentation to identify submarkets in the larger market, and we run parallel budget-split experiments in each submarket.
We demonstrate the effectiveness of this approach on real experiments on advertising markets at Meta.
Then, we formally study interference that derives from such experimental designs, using the first-price pacing equilibrium framework as our model of market equilibration.
We propose a debiased surrogate that eliminates the first-order bias of FPPE, and derive a plug-in estimator for the surrogate and establish its asymptotic normality. We then provide an estimation procedure for submarket parallel budget-controlled A/B tests. Finally, we present numerical examples on semi-synthetic data

, confirming that the debiasing technique achieves the desired coverage properti
es.
****

Yinuo Ren,Haoxuan Chen,Grant M. Rotskoff,Lexing Ying
How Discrete and Continuous Diffusion Meet: Comprehensive Analysis of Discrete D
iffusion Models via a Stochastic Integral Framework
Discrete diffusion models have gained increasing attention for their ability to
model complex distributions with tractable sampling and inference. However, the
error analysis for discrete diffusion models remains less well-understood. In th
is work, we propose a comprehensive framework for the error analysis of discrete
 diffusion models based on Lévy-type stochastic integrals. By generalizing the P
oisson random measure to that with a time-independent and state-dependent intens
ity, we rigorously establish a stochastic integral formulation of discrete diffu
sion models and provide the corresponding change of measure theorems that are in
triguingly analogous to Itô integrals and Girsanov's theorem for their continuou
s counterparts. Our framework unifies and strengthens the current theoretical re
sults on discrete diffusion models and obtains the first error bound for the $\t
au$-leaping scheme in KL divergence. With error sources clearly identified, our
analysis gives new insight into the mathematical properties of discrete diffusio
n models and offers guidance for the design of efficient and accurate algorithms
 for real-world discrete diffusion model applications.
****

Zhenting Qi,Mingyuan MA,Jiahang Xu,Li Lyna Zhang,Fan Yang,Mao Yang
Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solver
This paper introduces rStar, a self-play mutual reasoning approach that signific
antly improves reasoning capabilities of small language models (SLMs) without fi
ne-tuning or superior models. rStar decouples reasoning into a self-play mutual
generation-discrimination process. First, a target SLM augments the Monte Carlo
Tree Search (MCTS) with a rich set of human-like reasoning actions to construct
higher quality reasoning trajectories. Next, another SLM, with capabilities simi
lar to the target SLM, acts as a discriminator to verify each trajectory generat
ed by the target SLM. The mutually agreed reasoning trajectories are considered
mutual consistent, thus are more likely to be correct. Extensive experiments acr
oss five SLMs demonstrate rStar can effectively solve diverse reasoning problems
, including GSM8K, GSM-Hard, MATH, SVAMP, and StrategyQA. Remarkably, rStar boos
ts GSM8K accuracy from 12.51\% to 63.91\% for LLaMA2-7B, from 36.46\% to 81.88\%
 for Mistral-7B, from 74.53\% to 91.13\% for LLaMA3-8B-Instruct. Code is availab
le at https://github.com/zhentingqi/rStar.
****

Peijie Dong,Lujun Li,Yuedong Zhong,DaYou Du,Ruibo FAN,Yuhan Chen,Zhenheng Tang,Q
iang Wang,Wei Xue,Yike Guo,Xiaowen Chu
STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs
In this paper, we present the first structural binarization method for LLM compr
ession to less than 1-bit precision. Although LLMs have achieved remarkable perf
ormance, their memory-bound nature during the inference stage hinders the adopti
on of resource-constrained devices. Reducing weights to 1-bit precision through
binarization substantially enhances computational efficiency. We observe that ra
ndomly flipping some weights in binarized LLMs does not significantly degrade th
e model's performance, suggesting the potential for further compression. To expl
oit this, our STBLLM employs an N:M sparsity technique to achieve structural bin
arization of the weights. Specifically, we introduce a novel Standardized Import
ance (SI) metric, which considers weight magnitude and input feature norm to mor
e accurately assess weight significance. Then, we propose a layer-wise approach,
 allowing different layers of the LLM to be sparsified with varying N:M ratios,
thereby balancing compression and accuracy. Furthermore, we implement a fine-gra
ined grouping strategy for less important weights, applying distinct quantizatio
n schemes to sparse, intermediate, and dense regions. Finally, we design a speci
alized CUDA kernel to support structural binarization. We conduct extensive expe
riments on LLaMA, OPT, and Mistral family. STBLLM achieves a perplexity of 11.07
 at 0.55 bits per weight, outperforming the BiLLM by 3×. The results demonstrate

that our approach performs better than other compressed binarization LLM methods while significantly reducing memory requirements. Code is released at https://github.com/pprp/STBLLM.
****
Xingrui Wang,Wufei Ma,Angtian Wang,Shuo Chen,Adam Kortylewski,Alan Yuille
Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering
For vision-language models (VLMs), understanding the dynamic properties of objects and their interactions in 3D scenes from videos is crucial for effective reasoning about high-level temporal and action semantics. Although humans are adept at understanding these properties by constructing 3D and temporal (4D) representations of the world, current video understanding models struggle to extract these dynamic semantics, arguably because these models use cross-frame reasoning without underlying knowledge of the 3D/4D scenes.
In this work, we introduce **DynSuperCLEVR**, the first video question answering dataset that focuses on language understanding of the dynamic properties of 3D objects. We concentrate on three physical concepts—*velocity*, *acceleration*, and *collisions*—within 4D scenes. We further generate three types of questions, including factual queries, future predictions, and counterfactual reasoning that involve different aspects of reasoning on these 4D dynamic properties.
To further demonstrate the importance of explicit scene representations in answering these 4D dynamics questions, we propose **NS-4DPhysics**, a **N**eural-**S**ymbolic VideoQA model integrating **Physics** prior for **4D** dynamic properties with explicit scene representation of videos.
Instead of answering the questions directly from the video text input, our method first estimates the 4D world states with a 3D generative model powered by a physical prior, and then uses neural symbolic reasoning to answer the questions based on the 4D world states.
Our evaluation on all three types of questions in DynSuperCLEVR shows that previous video question answering models and large multimodal models struggle with questions about 4D dynamics, while our NS-4DPhysics significantly outperforms previous state-of-the-art models.
****
Kevin Li,Sachin Goyal,João D. Semedo,J Zico Kolter
Inference Optimal VLMs Need Fewer Visual Tokens and More Parameters
Vision Language Models (VLMs) have demonstrated strong capabilities across various visual understanding and reasoning tasks, driven by incorporating image representations into the token inputs of Large Language Models (LLMs). However, their real-world deployment is often constrained by high latency during inference due to the substantial compute required by the LLM to process the large number of input tokens, predominantly arising from the image. To reduce inference costs, one can either downsize the LLM or reduce the number of input tokens needed to represent the image, the latter of which has been the focus of many recent efforts around token compression. However, it is unclear what the optimal trade-off is given a fixed inference budget.
We first characterize this optimal trade-off between the number of visual tokens and LLM parameters by establishing scaling laws that capture variations in performance with these two factors. Our results reveal a surprising trend: for visual reasoning tasks, the inference-optimal behavior in VLMs is achieved by using the largest LLM that fits within the inference budget while minimizing visual token count - often to a single token. While the token reduction literature has mainly focused on maintaining base model performance by modestly reducing the token count (e.g., $5-10\times$), our results indicate that the compute-optimal inference regime requires operating under even higher token compression ratios. Based on these insights, we take the first steps toward designing token compression algorithms tailored for high-compression settings, utilizing prompt-based compression of tokens. Our work underscores the performance and efficiency benefits of operating in low visual token regimes and the importance of developing tailored token reduction algorithms for such conditions.
****

Alexey Skrynnik,Anton Andreychuk,Anatolii Borzilov,Alexander Chernyavskiy,Konstantin Yakovlev,Aleksandr Panov
POGEMA: A Benchmark Platform for Cooperative Multi-Agent Pathfinding
Multi-agent reinforcement learning (MARL) has recently excelled in solving challenging cooperative and competitive multi-agent problems in various environments, typically involving a small number of agents and full observability. Moreover, a range of crucial robotics-related tasks, such as multi-robot pathfinding, which have traditionally been approached with classical non-learnable methods (e.g., heuristic search), are now being suggested for solution using learning-based or hybrid methods. However, in this domain, it remains difficult, if not impossible, to conduct a fair comparison between classical, learning-based, and hybrid approaches due to the lack of a unified framework that supports both learning and evaluation. To address this, we introduce POGEMA, a comprehensive set of tools that includes a fast environment for learning, a problem instance generator, a collection of predefined problem instances, a visualization toolkit, and a benchmarking tool for automated evaluation. We also introduce and define an evaluation protocol that specifies a range of domain-related metrics, computed based on primary evaluation indicators (such as success rate and path length), enabling a fair multi-fold comparison. The results of this comparison, which involves a variety of state-of-the-art MARL, search-based, and hybrid methods, are presented.
****
Haoyuan Li,Zhou Yanpeng,Tao Tang,Jifei Song,Yihan Zeng,Michael Kampffmeyer,Hang Xu,Xiaodan Liang
UniGS: Unified Language-Image-3D Pretraining with Gaussian Splatting
Recent advancements in multi-modal 3D pre-training methods have shown promising efficacy in learning joint representations of text, images, and point clouds. However, adopting point clouds as 3D representation fails to fully capture the intricacies of the 3D world and exhibits a noticeable gap between the discrete points and the dense 2D pixels of images. To tackle this issue, we propose UniGS, integrating 3D Gaussian Splatting (3DGS) into multi-modal pre-training to enhance the 3D representation. We first rely on the 3DGS representation to model the 3D world as a collection of 3D Gaussians with color and opacity, incorporating all the information of the 3D scene while establishing a strong connection with 2D images. Then, to achieve Language-Image-3D pertaining, UniGS starts with a pretrained vision-language model to establish a shared visual and textual space through extensive real-world image-text pairs. Subsequently, UniGS employs a 3D encoder to align the optimized 3DGS with the Language-Image representations to learn unified multi-modal representations. To facilitate the extraction of global explicit 3D features by the 3D encoder and achieve better cross-modal alignment, we additionally introduce a novel Gaussian-Aware Guidance module that guides the learning of fine-grained representations of the 3D domain. Through extensive experiments across the Objaverse, ABO, MVImgNet and SUN RGBD datasets with zero-shot classification, text-driven retrieval and open-world understanding tasks, we demonstrate the effectiveness of UniGS in learning a more general and stronger aligned multi-modal representation. Specifically, UniGS achieves leading results across different 3D tasks with remarkable improvements over previous SOTA, Uni3D, including on zero-shot classification (+9.36%), text-driven retrieval (+4.3%) and open-world understanding (+7.92%).
****
Haoyi Zhu,Honghui Yang,Yating Wang,Jiange Yang,Limin Wang,Tong He
SPA: 3D Spatial-Awareness Enables Effective Embodied Representation
In this paper, we introduce SPA, a novel representation learning framework that emphasizes the importance of 3D spatial awareness in embodied AI. Our approach leverages differentiable neural rendering on multi-view images to endow a vanilla Vision Transformer (ViT) with intrinsic spatial understanding. We present the most comprehensive evaluation of embodied representation learning to date, covering 268 tasks across 8 simulators with diverse policies in both single-task and language-conditioned multi-task scenarios. The results are compelling: SPA consistently outperforms more than 10 state-of-the-art representation methods, including those specifically designed for embodied AI, vision-centric tasks, and multi-

modal applications, while using less training data. Furthermore, we conduct a se
ries of real-world experiments to confirm its effectiveness in practical scenari
os. These results highlight the critical role of 3D spatial awareness for embodi
ed representation learning. Our strongest model takes more than 6000 GPU hours t
o train and we are committed to open-sourcing all code and model weights to fost
er future research in embodied representation learning.
****

Yanjie Wang,Liqun Chen,Tianming Zhao,Tao Zhang,Guodong Wang,Luxin Yan,Sheng Zhon
g,Jiahuan Zhou,Xu Zou
High-dimension Prototype is a Better Incremental Object Detection Learner
Incremental object detection (IOD), surpassing simple classification, requires t
he simultaneous overcoming of catastrophic forgetting in both recognition and lo
calization tasks, primarily due to the significantly higher feature space comple
xity. Integrating Knowledge Distillation (KD) would mitigate the occurrence of c
atastrophic forgetting. However, the challenge of knowledge shift caused by invi
sible previous task data hampers existing KD-based methods, leading to limited i
mprovements in IOD performance. This paper aims to alleviate knowledge shift by
enhancing the accuracy and granularity in describing complex high-dimensional fe
ature spaces. To this end, we put forth a novel higher-dimension-prototype learn
ing approach for KD-based IOD, enabling a more flexible, accurate, and fine-grai
ned representation of feature distributions without the need to retain any previ
ous task data. Existing prototype learning methods calculate feature centroids o
r statistical Gaussian distributions as prototypes, disregarding actual irregula
r distribution information or leading to inter-class feature overlap, which is n
ot directly applicable to the more difficult task of IOD with complex feature sp
ace. To address the above issue, we propose a Gaussian Mixture Distribution-base
d Prototype (GMDP), which explicitly models the distribution relationships of di
fferent classes by directly measuring the likelihood of embedding from new and o
ld models into class distribution prototypes in a higher dimension manner. Speci
fically, GMDP dynamically adapts the component weights and corresponding means/
variances of class distribution prototypes to represent both intra-class and int
er-class variability more accurately. Progressing into a new task, GMDP constrai
ns the distance between the distribution of new and previous task classes, minim
izing overlap with existing classes and thus striking a balance between stabilit
y and adaptability. GMDP can be readily integrated into existing IOD methods to
enhance performance further. Extensive experiments on the PASCAL VOC and MS-COCO
 show that our method consistently exceeds four baselines by a large margin and
significantly outperforms other SOTA results under various settings.
****

Takashi Furuya,Maarten V. de Hoop,Gabriel Peyré
Transformers are Universal In-context Learners
Transformers are deep architectures that define ``in-context mappings'' which en
able predicting new tokens based on a given set of tokens (such as a prompt in N
LP applications or a set of patches for a vision transformer). In this work, we
study in particular the ability of these architectures to handle an arbitrarily
large number of context tokens. To mathematically, uniformly address their expre
ssivity, we consider the case that the mappings are conditioned on a context rep
resented by a probability distribution of tokens which becomes discrete for a fi
nite number of these. The relevant notion of smoothness then corresponds to cont
inuity in terms of the Wasserstein distance between these contexts. We demonstra
te that deep transformers are universal and can approximate continuous in-contex
t mappings to arbitrary precision, uniformly over compact token domains. This re
sult implies, as a special case, that transformers are universal approximators f
or continuous permutation-invariant mappings over a fixed number of tokens. It a
lso establishes the universal approximation capability of transformers for certa
in in-context learning tasks, demonstrating in particular their ability to perfo
rm regression within context. A key aspect of our results, compared to existing
findings, is that for a fixed precision, a single transformer can operate on an
arbitrary (even infinite) number of tokens. Additionally, it operates with a fix
ed embedding dimension of tokens (this dimension does not increase with precisio

n) and a fixed number of heads (proportional to the dimension). The use of MLPs between multi-head attention layers is also explicitly controlled. We consider b oth unmasked attentions (as used for the vision transformer) and masked causal a ttentions (as used for NLP and time series applications). We tackle the causal s etting leveraging a space-time lifting to analyze causal attention as a mapping over probability distributions of tokens.
****
Claas A Voelcker,Marcel Hussing,Eric Eaton,Amir-massoud Farahmand,Igor Gilitsche nski
MAD-TD: Model-Augmented Data stabilizes High Update Ratio RL
Building deep reinforcement learning (RL) agents that find a good policy with fe w samples has proven notoriously challenging. To achieve sample efficiency, rece nt work has explored updating neural networks with large numbers of gradient ste ps for every new sample. While such high update-to-data (UTD) ratios have shown strong empirical performance, they also introduce instability to the training pr ocess. Previous approaches need to rely on periodic neural network parameter re sets to address this instability, but restarting the training process is infeasi ble in many real-world applications and requires tuning the resetting interval. In this paper, we focus on one of the core difficulties of stable training with limited samples: the inability of learned value functions to generalize to unobs erved on-policy actions. We mitigate this issue directly by augmenting the off-p olicy RL training process with a small amount of data generated from a learned w orld model. Our method, Model-Augmented Data for TD Learning (MAD-TD) uses small amounts of generated data to stabilize high UTD training and achieve competitiv e performance on the most challenging tasks in the DeepMind control suite. Our e xperiments further highlight the importance of employing a good model to generat e data, MAD-TD's ability to combat value overestimation, and its practical stabi lity gains for continued learning.
****
Changlong Shi,Jinmeng Li,He Zhao,Dan dan Guo,Yi Chang
FedLWS: Federated Learning with Adaptive Layer-wise Weight Shrinking
In Federated Learning (FL), weighted aggregation of local models is conducted to generate a new global model, and the aggregation weights are typically normaliz ed to 1. A recent study identifies the global weight shrinking effect in FL, ind icating an enhancement in the global model’s generalization when the sum of weig hts (i.e., the shrinking factor) is smaller than 1, where how to learn the shrin king factor becomes crucial. However, principled approaches to this solution hav e not been carefully studied from the adequate consideration of privacy concerns and layer-wise distinctions. To this end, we propose a novel model aggregation strategy, Federated Learning with Adaptive Layer-wise Weight Shrinking (FedLWS), which adaptively designs the shrinking factor in a layer-wise manner and avoids optimizing the shrinking factors on a proxy dataset. We initially explored the factors affecting the shrinking factor during the training process. Then we calc ulate the layer-wise shrinking factors by considering the distinctions among eac h layer of the global model. FedLWS can be easily incorporated with various exis ting methods due to its flexibility. Extensive experiments under diverse scenari os demonstrate the superiority of our method over several state-of-the-art appro aches, providing a promising tool for enhancing the global model in FL.
****
Renqiu Xia,Mingsheng Li,Hancheng Ye,Wenjie Wu,Hongbin Zhou,Jiakang Yuan,Tianshuo Peng,Xinyu Cai,Xiangchao Yan,Bin Wang,Conghui He,Botian Shi,Tao Chen,Junchi Yan ,Bo Zhang
GeoX: Geometric Problem Solving Through Unified Formalized Vision-Language Pre-t raining
Despite their proficiency in general tasks, Multi-modal Large Language Models (M LLMs) struggle with automatic Geometry Problem Solving (GPS), which demands unde rstanding diagrams, interpreting symbols, and performing complex reasoning. This limitation arises from their pre-training on natural images and texts, along wi th the lack of automated verification in the problem-solving process. Besides, c urrent geometric specialists are limited by their task-specific designs, making

them less effective for broader geometric problems. To this end, we present GeoX , a multi-modal large model focusing on geometric understanding and reasoning ta sks. Given the significant differences between geometric diagram-symbol and natu ral image-text, we introduce unimodal pre-training to develop a diagram encoder and symbol decoder, enhancing the understanding of geometric images and corpora. Furthermore, we introduce geometry-language alignment, an effective pre-trainin g paradigm that bridges the modality gap between unimodal geometric experts. We propose a Generator-And-Sampler Transformer (GS-Former) to generate discriminati ve queries and eliminate uninformative representations from unevenly distributed geometric signals. Finally, GeoX benefits from visual instruction tuning, empow ering it to take geometric images and questions as input and generate verifiable solutions. Experiments show that GeoX outperforms both generalists and geometri c specialists on publicly recognized benchmarks, such as GeoQA, UniGeo, Geometry 3K, and PGPS9k. Our data and code will be released soon to accelerate future res earch on automatic GPS.
****
Yuke Li,Yujia Zheng,Guangyi Chen,Kun Zhang,Heng Huang
Identification of Intermittent Temporal Latent Process
Identifying time-delayed latent causal process is crucial for understanding temp oral dynamics and enabling downstream reasoning. While recent methods have made progress in identifying latent time-delayed causal processes, they cannot addres s the dynamics in which the influence of some latent factors on both the subsequ ent latent states and the observed data can become inactive or irrelevant at dif ferent time steps. Therefore, we introduce intermittent temporal latent processe s, where: (1) any subset of latent factors may be missing during nonlinear data generation at any time step, and (2) the active latent factors at each step are unknown. This framework encompasses both nonstationary and stationary transition s, accommodating changing or consistent active factors over time.
Our work shows that under certain assumptions, the latent causal variables are b lock-wise identifiable. With further conditional independence assumption, each l atent variable can even be recovered up to component-wise transformations.
Using this identification theory, we propose an unsupervised approach, InterLate nt, to reliably uncover the representations of the intermittent temporal latent process. The experimental findings on both synthetic and real-world datasets ver ify our theoretical claims.
****
Sarah Wiegreffe,Oyvind Tafjord,Yonatan Belinkov,Hannaneh Hajishirzi,Ashish Sabha rwal
Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions
Multiple-choice question answering (MCQA) is a key competence of performant tran sformer language models that is tested by mainstream benchmarks. However, recent evidence shows that models can have quite a range of performance, particularly when the task format is diversified slightly (such as by shuffling answer choice order). In this work we ask: how do successful models perform formatted MCQA? W e employ vocabulary projection and activation patching methods to localize key h idden states that encode relevant information for predicting the correct answer. We find that prediction of a specific answer symbol is causally attributed to a few middle layers, and specifically their multi-head self-attention mechanisms. We show that subsequent layers increase the probability of the predicted answer symbol in vocabulary space, and that this probability increase is associated wi th a sparse set of attention heads with unique roles. We additionally uncover di fferences in how different models adjust to alternative symbols. Finally, we dem onstrate that a synthetic task can disentangle sources of model error to pinpoin t when a model has learned formatted MCQA, and show that logit differences betwe en answer choice tokens continue to grow over the course of training.
****
Masih Eskandar,Tooba Imtiaz,Davin Hill,Zifeng Wang,Jennifer Dy
STAR: Stability-Inducing Weight Perturbation for Continual Learning
Humans can naturally learn new and varying tasks in a sequential manner.
Continual learning is a class of learning algorithms that updates its learned

model as it sees new data (on potentially new tasks) in a sequence.
A key challenge in continual learning is that as the model is updated to learn new tasks, it becomes susceptible to \textit{catastrophic forgetting}, where knowledge of previously learned tasks is lost. A popular approach to mitigate forgetting during continual learning is to maintain a small buffer of previously-seen samples, and to replay them during training. However, this approach is limited by the small buffer size and, while forgetting is reduced, it is still present.
In this paper, we propose
a novel loss function STAR that exploits the worst-case parameter perturbation that reduces the KL-divergence of model predictions with that of its local parameter neighborhood to promote stability and alleviate forgetting. STAR can be combined with almost any existing rehearsal-based methods as a plug-and-play component. We empirically show that STAR consistently improves performance of existing methods by up to $\sim15\\%$ across varying baselines, and achieves superior or competitive accuracy to that of state-of-the-art methods aimed at improving rehearsal-based continual learning. Our implementation is available at https://github.com/Gnomy17/STAR_CL.
****

Xuehao Wang,Zhan Zhuang,Feiyang Ye,Yu Zhang
MTSAM: Multi-Task Fine-Tuning for Segment Anything Model
The Segment Anything Model (SAM), with its remarkable zero-shot capability, has the potential to be a foundation model for multi-task learning. However, adopting SAM to multi-task learning faces two challenges: (a) SAM has difficulty generating task-specific outputs with different channel numbers, and (b) how to fine-tune SAM to adapt multiple downstream tasks simultaneously remains unexplored. To address these two challenges, in this paper, we propose the Multi-Task SAM (MTSAM) framework, which enables SAM to work as a foundation model for multi-task learning. MTSAM modifies SAM's architecture by removing the prompt encoder and implementing task-specific no-mask embeddings and mask decoders, enabling the generation of task-specific outputs. Furthermore, we introduce Tensorized low-Rank Adaptation (ToRA) to perform multi-task fine-tuning on SAM. Specifically, ToRA injects an update parameter tensor into each layer of the encoder in SAM and leverages a low-rank tensor decomposition method to incorporate both task-shared and task-specific information.
Extensive experiments conducted on benchmark datasets substantiate the efficacy of MTSAM in enhancing the performance of multi-task learning. Our code is available at https://github.com/XuehaoWangFi/MTSAM.
****

Xiangyu Qi,Ashwinee Panda,Kaifeng Lyu,Xiao Ma,Subhrajit Roy,Ahmad Beirami,Prateek Mittal,Peter Henderson
Safety Alignment Should be Made More Than Just a Few Tokens Deep
The safety alignment of current Large Language Models (LLMs) is vulnerable. Simple attacks, or even benign fine-tuning, can jailbreak aligned models. We note that many of these vulnerabilities are related to a shared underlying issue: safety alignment can take shortcuts, wherein the alignment adapts a model's generative distribution primarily over only its very first few output tokens. We unifiedly refer to this issue as shallow safety alignment. In this paper, we present case studies to explain why shallow safety alignment can exist and show how this issue universally contributes to multiple recently discovered vulnerabilities in LLMs, including the susceptibility to adversarial suffix attacks, prefilling attacks, decoding parameter attacks, and fine-tuning attacks. The key contribution of this work is that we demonstrate how this consolidated notion of shallow safety alignment sheds light on promising research directions for mitigating these vulnerabilities. We show that deepening the safety alignment beyond the first few tokens can meaningfully improve robustness against some common exploits. We also design a regularized fine-tuning objective that makes the safety alignment more persistent against fine-tuning attacks by constraining updates on initial tokens. Overall, we advocate that future safety alignment should be made more than just a few tokens deep.
****

Po-han Li,Sandeep P. Chinchali,ufuk topcu
CSA: Data-efficient Mapping of Unimodal Features to Multimodal Features
Multimodal encoders like CLIP excel in tasks such as zero-shot image classificat
ion and cross-modal retrieval. However, they require excessive training data.
We propose canonical similarity analysis (CSA), which uses two unimodal encoders
 to replicate multimodal encoders using limited data.
CSA maps unimodal features into a multimodal space, using a new similarity score
 to retain only the multimodal information.
CSA only involves the inference of unimodal encoders and a cubic-complexity matr
ix decomposition, eliminating the need for extensive GPU-based model training.
Experiments show that CSA outperforms CLIP while requiring $50,000\times$ fewer
multimodal data pairs to bridge the modalities given pre-trained unimodal encode
rs on ImageNet classification and misinformative news caption detection.
CSA surpasses the state-of-the-art method to map unimodal features to multimodal
 features.
We also demonstrate the ability of CSA with modalities beyond image and text, p
aving the way for future modality pairs with limited paired multimodal data but
abundant unpaired unimodal data, such as lidar and text.
****
Isaac Reid,Kumar Avinava Dubey,Deepali Jain,William F Whitney,Amr Ahmed,Joshua A
inslie,Alex Bewley,Mithun George Jacob,Aranyak Mehta,David Rendleman,Connor Sche
nck,Richard E. Turner,René Wagner,Adrian Weller,Krzysztof Marcin Choromanski
Linear Transformer Topological Masking with Graph Random Features
When training transformers on graph-structured data, incorporating information a
bout the underlying topology is crucial for good performance. Topological maskin
g, a type of relative position encoding, achieves this by upweighting or downwei
ghting attention depending on the relationship between the query and keys in the
 graph. In this paper, we propose to parameterise topological masks as a learnab
le function of a weighted adjacency matrix -- a novel, flexible approach which i
ncorporates a strong structural inductive bias. By approximating this mask with
graph random features (for which we prove the first known concentration bounds),
 we show how this can be made fully compatible with linear attention, preserving
 $\mathcal{O}(N)$ time and space complexity with respect to the number of input
tokens. The fastest previous alternative was $\mathcal{O}(N \log N)$ and only su
itable for specific graphs. Our efficient masking algorithms provide strong perf
ormance gains for image and point cloud data, including with >30k nodes.
****
Mehak Preet Dhaliwal,Andong Hua,Laya Pullela,Ryan Burke,Yao Qin
NutriBench: A Dataset for Evaluating Large Language Models in Nutrition Estimati
on from Meal Descriptions
Accurate nutrition estimation helps people make informed dietary choices and is
essential in the prevention of serious health complications. We present NutriBen
ch, the first publicly available natural language meal description nutrition ben
chmark. NutriBench consists of 11,857 meal descriptions generated from real-worl
d global dietary intake data. The data is human-verified and annotated with macr
o-nutrient labels, including carbohydrates, proteins, fats, and calories. We con
duct an extensive evaluation of Nutribench on the task of carbohydrate estimatio
n, testing twelve leading Large Language Models (LLMs), including GPT-4o, Llama3
.1, Qwen2, Gemma2, and OpenBioLLM models, using standard, Chain-of-Thought and R
etrieval-Augmented Generation strategies. Additionally, we present a study invol
ving professional nutritionists, finding that LLMs can provide comparable but si
gnificantly faster estimates. Finally, we perform a real-world risk assessment b
y simulating the effect of carbohydrate predictions on the blood glucose levels
of individuals with type 1 diabetes. Our work highlights the opportunities and c
hallenges of using LLMs for nutrition estimation, demonstrating their potential
to aid professionals and laypersons and improve health outcomes. Our benchmark i
s publicly available at: https://mehak126.github.io/nutribench.html
****
Jiarui Jin,Haoyu Wang,Hongyan Li,Jun Li,Jiahui Pan,Shenda Hong
Reading Your Heart: Learning ECG Words and Sentences via Pre-training ECG Langua

ge Model
Electrocardiogram (ECG) is essential for the clinical diagnosis of arrhythmias a
nd other heart diseases, but deep learning methods based on ECG often face limit
ations due to the need for high-quality annotations. Although previous ECG self-
supervised learning (eSSL) methods have made significant progress in representat
ion learning from unannotated ECG data, they typically treat ECG signals as ordi
nary time-series data, segmenting the signals using fixed-size and fixed-step ti
me windows, which often ignore the form and rhythm characteristics and latent se
mantic relationships in ECG signals. In this work, we introduce a novel perspect
ive on ECG signals, treating heartbeats as words and rhythms as sentences. Based
 on this perspective, we first designed the QRS-Tokenizer, which generates seman
tically meaningful ECG sentences from the raw ECG signals. Building on these, we
 then propose HeartLang, a novel self-supervised learning framework for ECG lang
uage processing, learning general representations at form and rhythm levels. Add
itionally, we construct the largest heartbeat-based ECG vocabulary to date, whic
h will further advance the development of ECG language processing. We evaluated
HeartLang across six public ECG datasets, where it demonstrated robust competiti
veness against other eSSL methods. Our data and code are publicly available at h
ttps://github.com/PKUDigitalHealth/HeartLang.
****
Jung Hyun Lee,June Yong Yang,Byeongho Heo,Dongyoon Han,Kyungsu Kim,Eunho Yang,Ka
ng Min Yoo
Token-Supervised Value Models for Enhancing Mathematical Problem-Solving Capabil
ities of Large Language Models
With the rapid advancement of test-time compute search strategies to improve the
 mathematical problem-solving capabilities of large language models (LLMs), the
need for building robust verifiers has become increasingly important. However, a
ll these inference strategies rely on existing verifiers originally designed for
 Best-of-N search, which makes them sub-optimal for tree search techniques at te
st time. During tree search, existing verifiers can only offer indirect and impl
icit assessments of partial solutions or under-value prospective intermediate st
eps, thus resulting in the premature pruning of promising intermediate steps. To
 overcome these limitations, we propose token-supervised value models (TVMs) --
a new class of verifiers that assign each token a probability that reflects the
likelihood of reaching the correct final answer. This new token-level supervisio
n enables TVMs to directly and explicitly evaluate partial solutions, effectivel
y distinguishing between promising and incorrect intermediate steps during tree
search at test time. Experimental results demonstrate that combining tree-search
-based inference strategies with TVMs significantly improves the accuracy of LLM
s in mathematical problem-solving tasks, surpassing the performance of existing
verifiers.
****
Andreas C Mueller,Carlo A Curino,Raghu Ramakrishnan
MotherNet: Fast Training and Inference via Hyper-Network Transformers
Foundation models are transforming machine learning across many modalities, with
 in-context learning replacing classical model training. Recent work on tabular
data hints at a similar opportunity to build foundation models for classificatio
n for numerical data. However, existing meta-learning approaches can not compete
 with tree-based methods in terms of inference time. In this paper, we propose M
otherNet, a hypernetwork architecture trained on synthetic classification tasks
that, once prompted with a never-seen-before training set generates the weights
of a trained ``child'' neural-network by in-context learning using a single forw
ard pass. In contrast to most existing hypernetworks that are usually trained fo
r relatively constrained multi-task settings, MotherNet can create models for mu
lticlass classification on arbitrary tabular datasets without any dataset specif
ic gradient descent.
The child network generated by MotherNet outperforms neural networks trained usi
ng gradient descent on small datasets, and is competitive with predictions by Ta
bPFN and standard ML methods like Gradient Boosting. Unlike a direct application
 of TabPFN, MotherNet generated networks are highly efficient at inference time.

****
Shuo Liu,An Zhang,Guoqing Hu,Hong Qian,Tat-Seng Chua
Preference Diffusion for Recommendation
Recommender systems aim to predict personalized item rankings by modeling user p
reference distributions derived from historical behavior data. While diffusion m
odels (DMs) have recently gained attention for their ability to model complex di
stributions, current DM-based recommenders typically rely on traditional objecti
ves such as mean squared error (MSE) or standard recommendation objectives. Thes
e approaches are either suboptimal for personalized ranking tasks or fail to exp
loit the full generative potential of DMs. To address these limitations, we prop
ose \textbf{PreferDiff}, an optimization objective tailored for DM-based recomme
nders. PreferDiff reformulates the traditional Bayesian Personalized Ranking (BP
R) objective into a log-likelihood generative framework, enabling it to effectiv
ely capture user preferences by integrating multiple negative samples. To handle
 the intractability, we employ variational inference, minimizing the variational
 upper bound. Furthermore, we replace MSE with cosine error to improve alignment
 with recommendation tasks, and we balance generative learning and preference mo
deling to enhance the training stability of DMs. PreferDiff devises three appeal
ing properties. First, it is the first personalized ranking loss designed specif
ically for DM-based recommenders. Second, it improves ranking performance and ac
celerates convergence by effectively addressing hard negatives. Third, we establ
ish its theoretical connection to Direct Preference Optimization (DPO), demonstr
ating its potential to align user preferences within a generative modeling frame
work. Extensive experiments across six benchmarks validate PreferDiff's superior
 recommendation performance.
Our codes are available at \url{https://github.com/lswhim/PreferDiff}.
****
Yichi Zhang,Yici Yan,Alex Schwing,Zhizhen Zhao
Towards Hierarchical Rectified Flow
We formulate a hierarchical rectified flow to model data distributions. It hiera
rchically couples multiple ordinary differential equations (ODEs) and defines a
time-differentiable stochastic process that generates a data distribution from a
 known source distribution. Each ODE resembles the ODE that is solved in a class
ic rectified flow, but differs in its domain, i.e., location, velocity, accelera
tion, etc. Unlike the classic rectified flow formulation, which formulates a sin
gle ODE in the location domain and only captures the expected velocity field (su
fficient to capture a multi-modal data distribution), the hierarchical rectified
 flow formulation models the multi-modal random velocity field, acceleration fie
ld, etc., in their entirety. This more faithful modeling of the random velocity
field enables integration paths to intersect when the underlying ODE is solved d
uring data generation. Intersecting paths in turn lead to integration trajectori
es that are more straight than those obtained in the classic rectified flow form
ulation, where integration paths cannot intersect. This leads to modeling of dat
a distributions with fewer neural function evaluations. We empirically verify th
is on synthetic 1D and 2D data as well as MNIST, CIFAR-10, and ImageNet-32 data.
 Our code is available at: https://riccizz.github.io/HRF/.
****
Badr MOUFAD,Yazid Janati,Lisa Bedin,Alain Oliviero Durmus,randal douc,Eric Mouli
nes,Jimmy Olsson
Variational Diffusion Posterior Sampling with Midpoint Guidance
Diffusion models have recently shown considerable potential in solving Bayesian
inverse problems when used as priors. However, sampling from the resulting denoi
sing posterior distributions remains a challenge as it involves intractable term
s. To tackle this issue, state-of-the-art approaches formulate the problem as th
at of sampling from a surrogate diffusion model targeting the posterior and deco
mpose its scores into two terms: the prior score and an intractable guidance ter
m. While the former is replaced by the pre-trained score of the considered diffu
sion model, the guidance term has to be estimated. In this paper, we propose a n
ovel approach that utilises a decomposition of the transitions which, in contras
t to previous methods, allows a trade-off between the complexity of the intracta

ble guidance term and that of the prior transitions. We validate the proposed approach through extensive experiments on linear and nonlinear inverse problems, including challenging cases with latent diffusion models as priors, and demonstrate its effectiveness in reconstructing electrocardiogram (ECG) from partial measurements for accurate cardiac diagnosis.
****

Yaxuan Wang,Jiaheng Wei,Chris Yuhao Liu,Jinlong Pang,Quan Liu,Ankit Shah,Yujia Bao,Yang Liu,Wei Wei

LLM Unlearning via Loss Adjustment with Only Forget Data

Unlearning in Large Language Models (LLMs) is essential for ensuring ethical and responsible AI use, especially in addressing privacy leak, bias, safety, and evolving regulations. Existing approaches to LLM unlearning often rely on retain data or a reference LLM, yet they struggle to adequately balance unlearning performance with overall model utility. This challenge arises because leveraging explicit retain data or implicit knowledge of retain data from a reference LLM to fine-tune the model tends to blur the boundaries between the forgotten and retain data, as different queries often elicit similar responses. In this work, we propose eliminating the need to retain data or the reference LLM for response calibration in LLM unlearning. Recognizing that directly applying gradient ascent on the forget data often leads to optimization instability and poor performance, our method guides the LLM on what not to respond to, and importantly, how to respond, based on the forget data. Hence, we introduce Forget data only Loss AjustmenT (FLAT), a "flat" loss adjustment approach which addresses these issues by maximizing f-divergence between the available template answer and the forget answer only w.r.t. the forget data. The variational form of the defined f-divergence theoretically provides a way of loss adjustment by assigning different importance weights for the learning w.r.t. template responses and the forgetting of responses subject to unlearning. Empirical results demonstrate that our approach not only achieves superior unlearning performance compared to existing methods but also minimizes the impact on the model’s retained capabilities, ensuring high utility across diverse tasks, including copyrighted content unlearning on Harry Potter dataset and MUSE Benchmark, and entity unlearning on the TOFU dataset.
****

Hansi Yang,Quanming Yao,James Kwok

Curriculum-aware Training for Discriminating Molecular Property Prediction Models

Despite their wide application across various fields, current molecular property prediction models struggle with the challenge of activity cliff, which refers to the situation where molecules with similar chemical structures display remarkable different properties. This phenomenon hinders existing models' ability to learn distinctive representations for molecules with similar chemical structures, and results in inaccurate predictions on molecules with activity cliff. To address this limitation, we first present empirical evidence demonstrating the ineffectiveness of standard training pipelines on molecules with activity cliff. We propose a novel approach that reformulates molecular property prediction as a node classification problem, introducing two innovative tasks at both the node and edge levels to improve learning outcomes for these challenging molecules with activity cliff. Our method is versatile, allowing seamless integration with a variety of base models, whether pre-trained or randomly initialized. Extensive evaluation across different molecular property prediction datasets validate the effectiveness of our approach.
****

Hanlin Yang,Jian Yao,Weiming Liu,Qing Wang,Hanmin Qin,Kong hansheng,Kirk Tang,Jiechao Xiong,Chao Yu,Kai Li,Junliang Xing,Hongwu Chen,Juchao Zhuo,QIANG FU,Yang Wei,Haobo Fu

Diverse Policies Recovering via Pointwise Mutual Information Weighted Imitation Learning

Recovering a spectrum of diverse policies from a set of expert trajectories is an important research topic in imitation learning. After determining a latent style for a trajectory, previous diverse polices recovering methods usually employ

a vanilla behavioral cloning learning objective conditioned on the latent style, treating each state-action pair in the trajectory with equal importance. Based on an observation that in many scenarios, behavioral styles are often highly relevant with only a subset of state-action pairs, this paper presents a new principled method in diverse polices recovering. In particular, after inferring or assigning a latent style for a trajectory, we enhance the vanilla behavioral cloning by incorporating a weighting mechanism based on pointwise mutual information. This additional weighting reflects the significance of each state-action pair's contribution to learning the style, thus allowing our method to focus on state-action pairs most representative of that style.
We provide theoretical justifications for our new objective, and extensive empirical evaluations confirm the effectiveness of our method in recovering diverse polices from expert data.
****
Ryan McKenna
Scaling up the Banded Matrix Factorization Mechanism for Large Scale Differentially Private ML
Correlated noise mechanisms such as DP Matrix Factorization (DP-MF) have proven to be effective alternatives to DP-SGD in large-epsilon few-epoch training regimes. Significant work has been done to find the best correlated noise strategies, and the current state-of-the-art approach is DP-BandMF , which optimally balances the benefits of privacy amplification and noise correlation. Despite it's utility advantages, severe scalability limitations prevent this mechanism from handling large-scale training scenarios where the number of training iterations may be more than 10^4 and the number of model parameters may exceed 10^7. In this work, we present techniques to scale up DP-BandMF along these two dimensions, significantly extending it's reach and enabling it to effectively handle settings with over 10^6 training iterations and 10^9 model parameters, with no utility degradation at smaller scales.
****
Hanrong Ye,Haotian Zhang,Erik Daxberger,Lin Chen,Zongyu Lin,Yanghao Li,Bowen Zhang,Haoxuan You,Dan Xu,Zhe Gan,Jiasen Lu,Yinfei Yang
MMEgo: Towards Building Egocentric Multimodal LLMs for Video QA
This research aims to comprehensively explore building a multimodal foundation model for egocentric video understanding.
To achieve this goal, we work on three fronts.
First, as there is a lack of QA data for egocentric video understanding, we automatically generate 7M high-quality QA samples for egocentric videos ranging from 30 seconds to one hour long in Ego4D based on human-annotated data.
This is one of the largest egocentric QA datasets.
Second, we contribute a challenging egocentric QA benchmark with 629 videos and 7,026 questions to evaluate the models' ability in recognizing and memorizing visual details across videos of varying lengths. We introduce a new de-biasing evaluation method to help mitigate the unavoidable language bias present in the models being evaluated.
Third, we propose a specialized multimodal architecture featuring a novel ``Memory Pointer Prompting" mechanism. This design includes a global glimpse step to gain an overarching understanding of the entire video and identify key visual information, followed by a fallback step that utilizes the key visual information to generate responses. This enables the model to more effectively comprehend extended video content.
With the data, benchmark, and model, we build MM-Ego, an egocentric multimodal LLM that shows powerful performance on egocentric video understanding.
****
George Stoica,Pratik Ramesh,Boglarka Ecsedi,Leshem Choshen,Judy Hoffman
Model merging with SVD to tie the Knots
Recent model merging methods demonstrate that the parameters of fully-finetuned models specializing in distinct tasks can be combined into one model capable of solving all tasks without retraining. Yet, this success does not transfer well when merging LoRA finetuned models. We study this phenomenon and observe that the

weights of LoRA finetuned models showcase a lower degree of alignment compared to their fully-finetuned counterparts. We hypothesize that improving this alignment is key to obtaining better LoRA model merges, and propose KnOTS to address this problem. KnOTS uses the SVD to jointly transform the weights of different LoRA models into an aligned space, where existing merging methods can be applied. In addition, we introduce a new benchmark that explicitly evaluates whether merged models are general models. Notably, KnOTS consistently improves LoRA merging by up to 4.3% across several vision and language benchmarks, including our new setting. We release our code at: https://github.com/gstoica27/KnOTS.
****

Yang Jin,Zhicheng Sun,Ningyuan Li,Kun Xu,Kun Xu,Hao Jiang,Nan Zhuang,Quzhe Huang,Yang Song,Yadong MU,Zhouchen Lin
Pyramidal Flow Matching for Efficient Video Generative Modeling
Video generation requires modeling a vast spatiotemporal space, which demands significant computational resources and data usage. To reduce the complexity, the prevailing approaches employ a cascaded architecture to avoid direct training with full resolution latent. Despite reducing computational demands, the separate optimization of each sub-stage hinders knowledge sharing and sacrifices flexibility. This work introduces a unified pyramidal flow matching algorithm. It reinterprets the original denoising trajectory as a series of pyramid stages, where only the final stage operates at the full resolution, thereby enabling more efficient video generative modeling. Through our sophisticated design, the flows of different pyramid stages can be interlinked to maintain continuity. Moreover, we craft autoregressive video generation with a temporal pyramid to compress the full-resolution history. The entire framework can be optimized in an end-to-end manner and with a single unified Diffusion Transformer (DiT). Extensive experiments demonstrate that our method supports generating high-quality 5-second (up to 10-second) videos at 768p resolution and 24 FPS within 20.7k A100 GPU training hours. All code and models are open-sourced at https://pyramid-flow.github.io.
****

Qinghao Ye,Xianhan Zeng,Fu Li,Chunyuan Li,Haoqi Fan
Painting with Words: Elevating Detailed Image Captioning with Benchmark and Alignment Learning
Image captioning has long been a pivotal task in visual understanding, with recent advancements in vision-language models (VLMs) significantly enhancing the ability to generate detailed image captions. However, the evaluation of detailed image captioning remains underexplored due to outdated evaluation metrics and coarse annotations. In this paper, we introduce DeCapBench along with a novel metric, DCScore, specifically designed for detailed captioning tasks. DCScore evaluates hallucinations and fine-grained comprehensiveness by deconstructing responses into the smallest self-sufficient units, termed primitive information units, and assessing them individually. Our evaluation shows that DCScore aligns more closely with human judgment than other rule-based or model-based metrics. Concurrently, DeCapBench exhibits a high correlation with VLM arena results on descriptive tasks, surpassing existing benchmarks for vision-language models. Additionally, we present an automatic fine-grained feedback collection method, FeedQuill, for preference optimization based on our advanced metric, demonstrating robust generalization capabilities across auto-generated preference data. Extensive experiments on multiple VLMs demonstrate that our method not only significantly reduces hallucinations but also enhances performance across various benchmarks, achieving superior detail captioning performance while surpassing GPT-4o.
****

Ka-Hei Hui,Chao Liu,Xiaohui Zeng,Chi-Wing Fu,Arash Vahdat
Not-So-Optimal Transport Flows for 3D Point Cloud Generation
Learning generative models of 3D point clouds is one of the fundamental problems in 3D generative learning. One of the key properties of point clouds is their permutation invariance, i.e., changing the order of points in a point cloud does not change the shape they represent. In this paper, we analyze the recently proposed equivariant OT flows that learn permutation invariant generative models for point-based molecular data and we show that these models scale poorly on large

point clouds. Also, we observe learning (equivariant) OT flows is generally chal
lenging since straightening flow trajectories makes the learned flow model compl
ex at the beginning of the trajectory. To remedy these, we propose not-so-optima
l transport flow models that obtain an approximate OT by an offline OT precomput
ation, enabling an efficient construction of OT pairs for training. During train
ing, we can additionally construct a hybrid coupling by combining our approximat
e OT and independent coupling to make the target flow models easier to learn. In
 an extensive empirical study, we show that our proposed model outperforms prior
 diffusion- and flow -based approaches on a wide range of unconditional generati
on and shape completion on the ShapeNet benchmark.
****
Luran Wang,Chaoran Cheng,Yizhen Liao,Yanru Qu,Ge Liu
Training Free Guided Flow-Matching with Optimal Control
Controlled generation with pre-trained Diffusion and Flow Matching models has va
st applications. One strategy for guiding ODE-based generative models is through
 optimizing a target loss $R(x_1)$ while staying close to the prior distribution
. Along this line, some recent work showed the effectiveness of guiding flow mod
el by differentiating through its ODE sampling process. Despite the superior per
formance, the theoretical understanding of this line of methods is still prelimi
nary, leaving space for algorithm improvement. Moreover, existing methods predom
inately focus on Euclidean data manifold, and there is a compelling need for gui
ded flow methods on complex geometries such as SO(3), which prevails in high-sta
ke scientific applications like protein design. We present OC-Flow, a general an
d theoretically grounded training-free framework for guided flow matching using
optimal control. Building upon advances in optimal control theory, we develop ef
fective and practical algorithms for solving optimal control in guided ODE-based
 generation and provide a systematic theoretical analysis of the convergence gua
rantee in both Euclidean and SO(3). We show that existing backprop-through-ODE m
ethods can be interpreted as special cases of Euclidean OC-Flow. OC-Flow achieve
d superior performance in extensive experiments on text-guided image manipulatio
n, conditional molecule generation, and all-atom peptide design.
****
Hikaru Shindo,Quentin Delfosse,Devendra Singh Dhami,Kristian Kersting
BlendRL: A Framework for Merging Symbolic and Neural Policy Learning
Humans can leverage both symbolic reasoning and intuitive responses. In contrast
, reinforcement learning policies are typically encoded in either opaque systems
 like neural networks or symbolic systems that rely on predefined symbols and ru
les. This disjointed approach severely limits the agents’ capabilities, as they
often lack either the flexible low-level reaction characteristic of neural agent
s or the interpretable reasoning of symbolic agents.

To overcome this challenge, we introduce *BlendRL*, a neuro-symbolic RL framewor
k that harmoniously integrates both paradigms.
We empirically demonstrate that BlendRL agents outperform both neural and symbol
ic baselines in standard Atari environments, and showcase their robustness to en
vironmental changes. Additionally, we analyze the interaction between neural and
 symbolic policies, illustrating how their hybrid use helps agents overcome each
 other's limitations.
****
Ashwinee Panda,Xinyu Tang,Christopher A. Choquette-Choo,Milad Nasr,Prateek Mitta
l
Privacy Auditing of Large Language Models
Current techniques for privacy auditing of large language models (LLMs) have lim
ited efficacy---they rely on basic approaches to generate canaries which leads t
o weak membership inference attacks that in turn give loose lower bounds on the
empirical privacy leakage.
We develop canaries that are far more effective than those used in prior work un
der threat models that cover a range of realistic settings.
We demonstrate through extensive experiments on multiple families of fine-tuned
LLMs that our approach sets a new standard for detection of privacy leakage. For

measuring the memorization rate of non-privately trained LLMs, our designed canaries surpass prior approaches.
For example, on the Qwen2.5-0.5B model, our designed canaries achieve 49.6% TPR at 1% FPR, vastly surpassing the prior approach's 4.2% TPR at 1% FPR.
Our method can be used to provide a privacy audit of $\varepsilon \approx 1$ for a model trained with theoretical ε of 4. To the best of our knowledge, this is the first time that a privacy audit of LLM training has achieved non trivial auditing success in the setting where the attacker cannot train shadow models, insert gradient canaries, or access the model at every iteration.
****
Nils Wandel,Stefan Schulz,Reinhard Klein
Metamizer: A Versatile Neural Optimizer for Fast and Accurate Physics Simulations
Efficient physics simulations are essential for numerous applications, ranging from realistic cloth animations in video games, to analyzing pollutant dispersion in environmental sciences, to calculating vehicle drag coefficients in engineering applications. Unfortunately, analytical solutions to the underlying physical equations are rarely available, and numerical solutions are computationally demanding.
Latest developments in the field of physics-based Deep Learning have led to promising efficiency gains but still suffer from limited generalization capabilities across multiple different PDEs.

Thus, in this work, we introduce **Metamizer**, a novel neural optimizer that iteratively solves a wide range of physical systems without retraining by minimizing a physics-based loss function. To this end, our approach leverages a scale-invariant architecture that enhances gradient descent updates to accelerate convergence. Since the neural network itself acts as an optimizer, training this neural optimizer falls into the category of meta-optimization approaches.
We demonstrate that Metamizer achieves high accuracy across multiple PDEs after training on the Laplace, advection-diffusion and incompressible Navier-Stokes equation as well as on cloth simulations. Remarkably, the model also generalizes to PDEs that were not covered during training such as the Poisson, wave and Burgers equation.
****
Shuze Liu,Claire Chen,Shangtong Zhang
Doubly Optimal Policy Evaluation for Reinforcement Learning
Policy evaluation estimates the performance of a policy by (1) collecting data from the environment and (2) processing raw data into a meaningful estimate. Due to the sequential nature of reinforcement learning, any improper data-collecting policy or data-processing method substantially deteriorates the variance of evaluation results over long time steps. Thus, policy evaluation often suffers from large variance and requires massive data to achieve the desired accuracy. In this work, we design an optimal combination of data-collecting policy and data-processing baseline. Theoretically, we prove our doubly optimal policy evaluation method is unbiased and guaranteed to have lower variance than previously best-performing methods. Empirically, compared with previous works, we show our method reduces variance substantially and achieves superior empirical performance.
****
Xinyou Wang,Zaixiang Zheng,Fei YE,Dongyu Xue,Shujian Huang,Quanquan Gu
DPLM-2: A Multimodal Diffusion Protein Language Model
Proteins are essential macromolecules defined by their amino acid sequences, which determine their three-dimensional structures and, consequently, their functions in all living organisms. Therefore, generative protein modeling necessitates a multimodal approach to simultaneously model, understand, and generate both sequences and structures. However, existing methods typically use separate models for each modality, limiting their ability to capture the intricate relationships between sequence and structure. This results in suboptimal performance in tasks that requires joint understanding and generation of both modalities.

In this paper, we introduce DPLM-2, a multimodal protein foundation model that e
xtends discrete diffusion protein language model (DPLM) to accommodate both sequ
ences and structures.
To enable structural learning with the language model, 3D coordinates are conver
ted to discrete tokens using a lookup-free quantization-based tokenizer.
By training on both experimental and high-quality synthetic structures, DPLM-2 l
earns the joint distribution of sequence and structure, as well as their margina
ls and conditionals.
We also implement an efficient warm-up strategy to exploit the connection betwee
n large-scale evolutionary data and structural inductive biases from pre-trained
 sequence-based protein language models.
Empirical evaluation shows that DPLM-2 can simultaneously generate highly compat
ible amino acid sequences and their corresponding 3D structures eliminating the
need for a two-stage generation approach.
Moreover, DPLM-2 demonstrates competitive performance in various conditional gen
eration tasks, including folding, inverse folding, and scaffolding with multimod
al motif inputs.
****
Yoav Wald,Mark Goldstein,Yonathan Efroni,Wouter A.C. van Amsterdam,Rajesh Rangan
ath
Time After Time: Deep-Q Effect Estimation for Interventions on When and What to
do
Problems in fields such as healthcare, robotics, and finance requires reasoning
about the value both of what decision or action to take and when to take it. The
 prevailing hope is that artificial intelligence will support such decisions by
estimating the causal effect of policies such as how to treat patients or how to
 allocate resources over time. However, existing methods for estimating the effe
ct of a policy struggle with \emph{irregular time}. They either discretize time
, or disregard the effect of timing policies.

We present a new deep-Q algorithm that estimates the effect of both when and wha
t to do called Earliest Disagreement Q-Evaluation (EDQ). EDQ makes use of recurs
ion for the Q-function that is compatible with flexible sequence models, such as
 transformers. EDQ provides accurate estimates under standard assumptions. We va
lidate the approach through experiments on survival time and tumor growth tasks.
****
Yangsibo Huang,Daogao Liu,Lynn Chua,Badih Ghazi,Pritish Kamath,Ravi Kumar,Pasin
Manurangsi,Milad Nasr,Amer Sinha,Chiyuan Zhang
Unlearn and Burn: Adversarial Machine Unlearning Requests Destroy Model Accuracy
Machine unlearning algorithms, designed for selective removal of training data f
rom models, have emerged as a promising approach to growing privacy concerns. In
 this work, we expose a critical yet underexplored vulnerability in the deployme
nt of unlearning systems: the assumption that the data requested for removal is
always part of the original training set. We present a threat model where an att
acker can degrade model accuracy by submitting adversarial unlearning requests f
or data \textit{not} present in the training set. We propose white-box and black
-box attack algorithms and evaluate them through a case study on image classific
ation tasks using the CIFAR-10 and ImageNet datasets, targeting a family of wide
ly used unlearning methods. Our results show extremely poor test accuracy follow
ing the attack—3.6% on CIFAR-10 and 0.4% on ImageNet for white-box attacks, and
8.5% on CIFAR-10 and 1.3% on ImageNet for black-box attacks. Additionally, we ev
aluate various verification mechanisms to detect the legitimacy of unlearning re
quests and reveal the challenges in verification, as most of the mechanisms fail
 to detect stealthy attacks without severely impairing their ability to process
valid requests. These findings underscore the urgent need for research on more r
obust request verification methods and unlearning protocols, should the deployme
nt of machine unlearning systems become more prevalent in the future.
****
Ezekiel Williams,Alexandre Payeur,Avery Hee-Woon Ryoo,Thomas Jiralerspong,Matthe
w G Perich,Luca Mazzucato,Guillaume Lajoie

Expressivity of Neural Networks with Random Weights and Learned Biases
Landmark universal function approximation results for neural networks with trained weights and biases provided the impetus for the ubiquitous use of neural networks as learning models in neuroscience and Artificial Intelligence (AI). Recent work has extended these results to networks in which a smaller subset of weights (e.g., output weights) are tuned, leaving other parameters random. However, it remains an open question whether universal approximation holds when only biases are learned, despite evidence from neuroscience and AI that biases significantly shape neural responses. The current paper answers this question. We provide theoretical and numerical evidence demonstrating that feedforward neural networks with fixed random weights can approximate any continuous function on compact sets. We further show an analogous result for the approximation of dynamical systems with recurrent neural networks. Our findings are relevant to neuroscience, where they demonstrate the potential for behaviourally relevant changes in dynamics without modifying synaptic weights, as well as for AI, where they shed light on recent fine-tuning methods for large language models, like bias and prefix-based approaches.
****
Yuzhe YAO,Jun Chen,Zeyi Huang,Haonan Lin,Mengmeng Wang,Guang Dai,Jingdong Wang
Manifold Constraint Reduces Exposure Bias in Accelerated Diffusion Sampling
Diffusion models have demonstrated significant potential for generating high-quality images, audio, and videos. However, their iterative inference process entails substantial computational costs, limiting practical applications. Recently, researchers have introduced accelerated sampling methods that enable diffusion models to generate samples with far fewer timesteps than those used during training. Nonetheless, as the number of sampling steps decreases, the prediction errors significantly degrade the quality of generated outputs. Additionally, the exposure bias in diffusion models further amplifies these errors. To address these challenges, we leverage a manifold hypothesis to explore the exposure bias problem in depth. Based on this geometric perspective, we propose a manifold constraint that effectively reduces exposure bias during accelerated sampling of diffusion models. Notably, our method involves no additional training and requires only minimal hyperparameter tuning. Extensive experiments demonstrate the effectiveness of our approach, achieving a FID score of 15.60 with 10-step SDXL on MS-COCO, surpassing the baseline by a reduction of 2.57 in FID.
****
Xudong Gong,Feng Dawei,Kele Xu,weijia wang,Zhangjun Sun,Xing Zhou,Si Zheng,Bo Ding,Huaimin Wang
VVC-Gym: A Fixed-Wing UAV Reinforcement Learning Environment for Multi-Goal Long-Horizon Problems
Multi-goal long-horizon problems are prevalent in real-world applications. The additional goal space introduced by multi-goal problems intensifies the spatial complexity of exploration; meanwhile, the long interaction sequences in long-horizon problems exacerbate the temporal complexity of exploration. Addressing the great exploration challenge posed by multi-goal long-horizon problems depends not only on the design of algorithms but also on the design of environments and the availability of demonstrations to assist in training. To facilitate the above research, we propose a multi-goal long-horizon Reinforcement Learning (RL) environment based on realistic fixed-wing UAV's velocity vector control, named VVC-Gym, and generate multiple demonstration sets of various quality. Through experimentation, we analyze the impact of different environment designs on training, assess the quantity and quality of demonstrations and their influence on training, and assess the effectiveness of various RL algorithms, providing baselines on VVC-Gym and its corresponding demonstrations. The results suggest that VVC-Gym is suitable for studying: (1) the influence of environment designs on addressing multi-goal long-horizon problems with RL. (2) the assistance that demonstrations can provide in overcoming the exploration challenges of multi-goal long-horizon problems. (3) the RL algorithm designs with the least possible impact from environment designs on the efficiency and effectiveness of training.
****

Haowen Pan,Xiaozhi Wang,Yixin Cao,Zenglin Shi,Xun Yang,Juanzi Li,Meng Wang
Precise Localization of Memories: A Fine-grained Neuron-level Knowledge Editing Technique for LLMs
Knowledge editing aims to update outdated information in Large Language Models (LLMs). A representative line of study is locate-then-edit methods, which typically employ causal tracing to identify the modules responsible for recalling factual knowledge about entities. However, we find these methods are often sensitive only to changes in the subject entity, leaving them less effective at adapting to changes in relations. This limitation results in poor editing locality, which can lead to the persistence of irrelevant or inaccurate facts, ultimately compromising the reliability of LLMs. We believe this issue arises from the insufficient precision of knowledge localization. To address this, we propose a Fine-grained Neuron-level Knowledge Editing (FiNE) method that enhances editing locality without affecting overall success rates. By precisely identifying and modifying specific neurons within feed-forward networks, FiNE significantly improves knowledge localization and editing. Quantitative experiments demonstrate that FiNE efficiently achieves better overall performance compared to existing techniques, providing new insights into the localization and modification of knowledge within LLMs.
****
Mridul Gupta,Samyak Jain,Vansh Ramani,HARIPRASAD KODAMANA,Sayan Ranu
Bonsai: Gradient-free Graph Condensation for Node Classification
Graph condensation has emerged as a promising avenue to enable scalable training of GNNs by compressing the training dataset while preserving essential graph characteristics. Our study uncovers significant shortcomings in current graph condensation techniques. First, the majority of the algorithms paradoxically require training on the full dataset to perform condensation. Second, due to their gradient-emulating approach, these methods require fresh condensation for any change in hyperparameters or GNN architecture, limiting their flexibility and reusability. To address these challenges, we present Bonsai, a novel graph condensation method empowered by the observation that *computation trees* form the fundamental processing units of message-passing GNNs. Bonsai condenses datasets by encoding a careful selection of *exemplar* trees that maximize the representation of all computation trees in the training set. This unique approach imparts Bonsai as the first linear-time, model-agnostic graph condensation algorithm for node classification that outperforms existing baselines across 7 real-world datasets on accuracy, while being 22 times faster on average. Bonsai is grounded in rigorous mathematical guarantees on the adopted approximation strategies, making it robust to GNN architectures, datasets, and parameters.
****
Sarp Aykent,Tian Xia
GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks
Understanding complex three-dimensional (3D) structures of graphs is essential for accurately modeling various properties, yet many existing approaches struggle with fully capturing the intricate spatial relationships and symmetries inherent in such systems, especially in large-scale, dynamic molecular datasets. These methods often must balance trade-offs between expressiveness and computational efficiency, limiting their scalability. To address this gap, we propose a novel Geometric Tensor Network (GotenNet) that effectively models the geometric intricacies of 3D graphs while ensuring strict equivariance under the Euclidean group E(3). Our approach directly tackles the expressiveness-efficiency trade-off by leveraging effective geometric tensor representations without relying on irreducible representations or Clebsch-Gordan transforms, thereby reducing computational overhead. We introduce a unified structural embedding, incorporating geometry-aware tensor attention and hierarchical tensor refinement that iteratively updates edge representations through inner product operations on high-degree steerable features, allowing for flexible and efficient representations for various tasks. We evaluated models on QM9, rMD17, MD22, and Molecule3D datasets, where the proposed model consistently outperforms state-of-the-art methods in both scalar and high-degree property predictions, demonstrating exceptional robustness across d

iverse datasets, and establishes GotenNet as a versatile and scalable framework for 3D equivariant Graph Neural Networks.
****
Satoki Ishikawa,Makoto Yamada,Han Bao,Yuki Takezawa
PhiNets: Brain-inspired Non-contrastive Learning Based on Temporal Prediction Hypothesis
Predictive coding has been established as a promising neuroscientific theory to describe the mechanism of information processing in the retina or cortex.
This theory hypothesises that cortex predicts sensory inputs at various levels of abstraction to minimise prediction errors.
Inspired by predictive coding, Chen et al. (2024) proposed another theory, temporal prediction hypothesis, to claim that sequence memory residing in hippocampus has emerged through predicting input signals from the past sensory inputs.
Specifically, they supposed that the CA3 predictor in hippocampus creates synaptic delay between input signals, which is compensated by the following CA1 predictor.
Though recorded neural activities were replicated based on the temporal prediction hypothesis, its validity has not been fully explored.
In this work, we aim to explore the temporal prediction hypothesis from the perspective of self-supervised learning (SSL).
Specifically, we focus on non-contrastive learning, which generates two augmented views of an input image and predicts one from another.
Non-contrastive learning is intimately related to the temporal prediction hypothesis because the synaptic delay is implicitly created by StopGradient.
Building upon a popular non-contrastive learner, SimSiam, we propose PhiNet, an extension of SimSiam to have two predictors explicitly corresponding to the CA3 and CA1, respectively.
Through studying the PhiNet model, we discover two findings.
First, meaningful data representations emerge in PhiNet more stably than in SimSiam.
This is initially supported by our learning dynamics analysis: PhiNet is more robust to the representational collapse.
Second, PhiNet adapts more quickly to newly incoming patterns in online and continual learning scenarios.
For practitioners, we additionally propose an extension called X-PhiNet integrated with a momentum encoder, excelling in continual learning.
All in all, our work reveals that the temporal prediction hypothesis is a reasonable model in terms of the robustness and adaptivity.
****
Mingxi Lei,Chunwei Ma,Meng Ding,Yufan Zhou,Ziyun Huang,Jinhui Xu
TTVD: Towards a Geometric Framework for Test-Time Adaptation Based on Voronoi Diagram
Deep learning models often struggle with generalization when deploying on real-world data, due to the common distributional shift to the training data. Test-time adaptation (TTA) is an emerging scheme used at inference time to address this issue. In TTA, models are adapted online at the same time when making predictions to test data. Neighbor-based approaches have gained attention recently, where prototype embeddings provide location information to alleviate the feature shift between training and testing data. However, due to their inherit limitation of simplicity, they often struggle to learn useful patterns and encounter performance degradation. To confront this challenge, we study the TTA problem from a geometric point of view. We first reveal that the underlying structure of neighbor-based methods aligns with the Voronoi Diagram, a classical computational geometry model for space partitioning. Building on this observation, we propose the Test-Time adjustment by Voronoi Diagram guidance (TTVD), a novel framework that leverages the benefits of this geometric property. Specifically, we explore two key structures: 1) Cluster-induced Voronoi Diagram (CIVD): This integrates the joint contribution of self-supervision and entropy-based methods to provide richer information. 2) Power Diagram (PD): A generalized version of the Voronoi Diagram that refines partitions by assigning weights to each Voronoi cell. Our experiment

s under rigid, peer-reviewed settings on CIFAR-10-C, CIFAR-100-C, ImageNet-C, an d ImageNet-R shows that TTVD achieves remarkable improvements compared to state-of-the-art methods. Moreover, extensive experimental results also explore the ef fects of batch size and class imbalance, which are two scenarios commonly encoun tered in real-world applications. These analyses further validate the robustness and adaptability of our proposed framework.
****

Feng Liu,Faguo Wu,Xiao Zhang
Score-based free-form architectures for high-dimensional Fokker-Planck equations
Deep learning methods incorporate PDE residuals as the loss function for solving Fokker-Planck equations, and usually impose the proper normalization condition to avoid a trivial solution. However, soft constraints require careful balancing of multi-objective loss functions, and specific network architectures may limit representation capacity under hard constraints. In this paper, we propose a nov el framework: Fokker-Planck neural network (FPNN) that adopts a score PDE loss t o decouple the score learning and the density normalization into two stages. Our method allows free-form network architectures to model the unnormalized density and strictly satisfy normalization constraints by post-processing. We demonstra te the effectiveness on various high-dimensional steady-state Fokker-Planck (SFP) equations, achieving superior accuracy and over a 20$\times$ speedup compared to state-of-the-art methods. Without any labeled data, FPNNs achieve the mean ab solute percentage error (MAPE) of 11.36\%, 13.87\% and 12.72\% for 4D Ring, 6D U nimodal and 6D Multi-modal problems respectively, requiring only 256, 980, and 9 80 parameters. Experimental results highlights the potential as a universal fast solver for handling more than 20-dimensional SFP equations, with great gains in efficiency, accuracy, memory and computational resource usage.
****

Vaibhav Raj,Indradyumna Roy,Ashwin Ramachandran,Soumen Chakrabarti,Abir De
Charting the Design Space of Neural Graph Representations for Subgraph Matching
Subgraph matching is vital in knowledge graph (KG) question answering, molecule design, scene graph, code and circuit search, etc.
Neural methods have shown promising results for subgraph matching.
Our study of recent systems suggests refactoring them into a unified design spac e for graph matching networks.
Existing methods occupy only a few isolated patches in this space, which remains largely uncharted.
We undertake the first comprehensive exploration of this space, featuring such a xes as attention-based vs. soft permutation-based interaction between query and corpus graphs, aligning nodes vs. edges, and the form of the final scoring netwo rk that integrates neural representations of the graphs.
Our extensive experiments reveal that judicious and hitherto-unexplored combinat ions of choices in this space lead to large performance benefits.
Beyond better performance, our study uncovers valuable insights and establishes general design principles for neural graph representation and interaction, which may be of wider interest.
****

Ronghua Zheng,Hanru Bai,Weiyang Ding
KooNPro: A Variance-Aware Koopman Probabilistic Model Enhanced by Neural Process for Time Series Forecasting
The probabilistic forecasting of time series is a well-recognized challenge, par ticularly in disentangling correlations among interacting time series and addres sing the complexities of distribution modeling. By treating time series as tempo ral dynamics, we introduce **KooNPro**, a novel probabilistic time series foreca sting model that combines variance-aware deep **Koo**pman model with **N**eural **Pro**cess. KooNPro introduces a variance-aware continuous spectrum using Gauss ian distributions to capture complex temporal dynamics with improved stability. It further integrates the Neural Process to capture fine dynamics, enabling enha nced dynamics capture and prediction. Extensive experiments on nine real-world d atasets demonstrate that KooNPro consistently outperforms state-of-the-art basel ines. Ablation studies highlight the importance of the Neural Process component

and explore the impact of key hyperparameters. Overall, KooNPro presents a promi
sing novel approach for probabilistic time series forecasting.
****

The Viet Bui,Thanh Hong Nguyen,Tien Anh Mai

ComaDICE: Offline Cooperative Multi-Agent Reinforcement Learning with Stationary
 Distribution Shift Regularization

Offline reinforcement learning (RL) has garnered significant attention for its a
bility to learn effective policies from pre-collected datasets without the need
for further environmental interactions. While promising results have been demons
trated in single-agent settings, offline multi-agent reinforcement learning (MAR
L) presents additional challenges due to the large joint state-action space and
the complexity of multi-agent behaviors. A key issue in offline RL is the distri
butional shift, which arises when the target policy being optimized deviates fro
m the behavior policy that generated the data. This problem is exacerbated in MA
RL due to the interdependence between agents' local policies and the expansive j
oint state-action space. Prior approaches have primarily addressed this challeng
e by incorporating regularization in the space of either Q-functions or policies
. In this work, we propose a novel type of regularizer in the space of stationar
y distributions to address the distributional shift more effectively. Our algori
thm, ComaDICE, provides a principled framework for offline cooperative MARL to c
orrect the stationary distribution of the global policy, which is then leveraged
 to derive local policies for individual agents. Through extensive experiments o
n the offline multi-agent MuJoCo and StarCraft II benchmarks, we demonstrate tha
t ComaDICE achieves superior performance compared to state-of-the-art offline MA
RL methods across nearly all tasks.
****

Sujay Nagaraj,Walter Gerych,Sana Tonekaboni,Anna Goldenberg,Berk Ustun,Thomas Ha
rtvigsen

Learning under Temporal Label Noise

Many time series classification tasks, where labels vary over time, are affected
 by label noise that also varies over time. Such noise can cause label quality t
o improve, worsen, or periodically change over time. We first propose and formal
ize temporal label noise, an unstudied problem for sequential classification of
time series. In this setting, multiple labels are recorded over time while being
 corrupted by a time-dependent noise function. We first demonstrate the importan
ce of modeling the temporal nature of the label noise function and how existing
methods will consistently underperform. We then propose methods to train noise-t
olerant classifiers by estimating the temporal label noise function directly fro
m data. We show that our methods lead to state-of-the-art performance under dive
rse types of temporal label noise on real-world datasets.
****

Yikang Li,Yeqing Qiu,Yuxuan Chen,Zhouchen Lin

Affine Steerable Equivariant Layer for Canonicalization of Neural Networks

In the field of equivariant networks, achieving affine equivariance, particularl
y for general group representations, has long been a challenge.
In this paper, we propose the steerable EquivarLayer, a generalization of InvarL
ayer (Li et al., 2024), by building on the concept of equivariants beyond invari
ants.
The steerable EquivarLayer supports affine equivariance with arbitrary input and
 output representations, marking the first model to incorporate steerability int
o networks for the affine group.
To integrate it with canonicalization, a promising approach for making pre-train
ed models equivariant, we introduce a novel Det-Pooling module, expanding the ap
plicability of EquivarLayer and the range of groups suitable for canonicalizatio
n.
We conduct experiments on image classification tasks involving group transformat
ions to validate the steerable EquivarLayer in the role of a canonicalization fu
nction, demonstrating its effectiveness over data augmentation.
****

Andreas Opedal,Haruki Shirakami,Bernhard Schölkopf,Abulhair Saparov,Mrinmaya Sac

han
MathGAP: Out-of-Distribution Evaluation on Problems with Arbitrarily Complex Pro
ofs
Large language models (LLMs) can solve arithmetic word problems with high accura
cy, but little is known about how well they generalize to more complex problems.
This is difficult to study, as (i) much of the available evaluation data has al
ready been seen by the most capable models during training, and (ii) existing be
nchmarks do not capture how problem proofs may be arbitrarily complex in various
ways. In this paper, we present a data-generation framework for evaluating LLMs
on problems with arbitrarily complex arithmetic proofs, called MathGAP. MathGAP
generates problem statements and chain-of-thought reasoning traces according to
specifications about their arithmetic proof structure, enabling systematic stud
ies on easy-to-hard generalization with respect to complexity of proof trees. Us
ing MathGAP, we find that LLMs show a significant decrease in performance as pro
ofs get deeper and wider. This effect is more pronounced in complex, nonlinear p
roof structures, which are challenging even for the most capable models. The mod
els are also sensitive to simple changes in sentence ordering. However, they rem
ain capable of solving some complex problems, suggesting that reasoning generali
zation is noisy.
****
Can Jin,Ying Li,Mingyu Zhao,Shiyu Zhao,Zhenting Wang,Xiaoxiao He,Ligong Han,Tong
Che,Dimitris N. Metaxas
LoR-VP: Low-Rank Visual Prompting for Efficient Vision Model Adaptation
Visual prompting has gained popularity as a method for adapting pre-trained mode
ls to specific tasks, particularly in the realm of parameter-efficient tuning. H
owever, existing visual prompting techniques often pad the prompt parameters aro
und the image, limiting the interaction between the visual prompts and the origi
nal image to a small set of patches while neglecting the inductive bias present
in shared information across different patches. In this study, we conduct a thor
ough preliminary investigation to identify and address these limitations. We pro
pose a novel visual prompt design, introducing **Lo**w-**R**ank matrix multiplic
ation for **V**isual **P**rompting (LoR-VP), which enables shared and patch-spec
ific information across rows and columns of image pixels. Extensive experiments
across seven network architectures and four datasets demonstrate significant imp
rovements in both performance and efficiency compared to state-of-the-art visual
prompting methods, achieving up to $6\times$ faster training times, utilizing $
18\times$ fewer visual prompt parameters, and delivering a 3.1% improvement in p
erformance.
****
Ran Eisenberg,Jonathan Svirsky,Ofir Lindenbaum
COPER: Correlation-based Permutations for Multi-View Clustering
Combining data from different sources can improve data analysis tasks such as cl
ustering. However, most of the current multi-view clustering methods are limited
to specific domains or rely on a suboptimal and computationally intensive two-s
tage process of representation learning and clustering. We propose an end-to-end
deep learning-based multi-view clustering framework for general data types (suc
h as images and tables). Our approach involves generating meaningful fused repre
sentations using a novel permutation-based canonical correlation objective. We p
rovide a theoretical analysis showing how the learned embeddings approximate tho
se obtained by supervised linear discriminant analysis (LDA). Cluster assignment
s are learned by identifying consistent pseudo-labels across multiple views. Add
itionally, we establish a theoretical bound on the error caused by incorrect pse
udo-labels in the unsupervised representations compared to LDA. Extensive experi
ments on ten multi-view clustering benchmark datasets provide empirical evidence
for the effectiveness of the proposed model.
****
Menglong Zhang,Fuyuan Qian,Quanying Liu
Learning Task Belief Similarity with Latent Dynamics for Meta-Reinforcement Lear
ning
Meta-reinforcement learning requires utilizing prior task distribution informati

on obtained during exploration to rapidly adapt to unknown tasks. The efficiency of an agent's exploration hinges on accurately identifying the current task. Recent Bayes-Adaptive Deep RL approaches often rely on reconstructing the environment's reward signal, which is challenging in sparse reward settings, leading to suboptimal exploitation. Inspired by bisimulation metrics, which robustly extracts behavioral similarity in continuous MDPs, we propose SimBelief—a novel meta-RL framework via measuring similarity of task belief in Bayes-Adaptive MDP (BAMDP). SimBelief effectively extracts common features of similar task distributions, enabling efficient task identification and exploration in sparse reward environments. We introduce latent task belief metric to learn the common structure of similar tasks and incorporate it into the real task belief. By learning the latent dynamics across task distributions, we connect shared latent task belief features with specific task features, facilitating rapid task identification and adaptation. Our method outperforms state-of-the-art baselines on sparse reward MuJoCo and panda-gym tasks.
****

Hyungkyu Kang,Min-hwan Oh
Adversarial Policy Optimization for Offline Preference-based Reinforcement Learning
In this paper, we study offline preference-based reinforcement learning (PbRL), where learning is based on pre-collected preference feedback over pairs of trajectories. While offline PbRL has demonstrated remarkable empirical success, existing theoretical approaches face challenges in ensuring conservatism under uncertainty, requiring computationally intractable confidence set constructions. We address this limitation by proposing Adversarial Preference-based Policy Optimization (APPO), a computationally efficient algorithm for offline PbRL that guarantees sample complexity bounds without relying on explicit confidence sets. By framing PbRL as a two-player game between a policy and a model, our approach enforces conservatism in a tractable manner. Using standard assumptions on function approximation and bounded trajectory concentrability, we derive a sample complexity bound. To our knowledge, APPO is the first offline PbRL algorithm to offer both statistical efficiency and practical applicability. Experimental results on continuous control tasks demonstrate that APPO effectively learns from complex datasets, showing comparable performance with existing state-of-the-art methods.
****

Yongchao Chen,Harsh Jhamtani,Srinagesh Sharma,Chuchu Fan,Chi Wang
Steering Large Language Models between Code Execution and Textual Reasoning
While a lot of recent research focuses on enhancing the textual reasoning capabilities of Large Language Models (LLMs) by optimizing the multi-agent framework or reasoning chains, several benchmark tasks can be solved with 100\% success through direct coding, which is more scalable and avoids the computational overhead associated with textual iterating and searching. Textual reasoning has inherent limitations in solving tasks with challenges in math, logics, optimization, and searching, which is unlikely to be solved by simply scaling up the model and data size. The recently released OpenAI GPT Code Interpreter and multi-agent frameworks such as AutoGen have demonstrated remarkable proficiency of integrating code generation and execution to solve complex tasks using LLMs. However, based on our experiments on 7 existing popular methods for steering code/text generation in both single- and multi-turn settings with 14 tasks and 6 types of LLMs (including the new O1-preview), currently there is no optimal method to correctly steer LLMs to write code when needed. We discover some interesting patterns on when models use code vs. textual reasoning with the evolution to task complexity and model sizes, which even result in an astonishingly inverse scaling behavior. We also discover that results from LLM written code are not always better than using textual reasoning, even if the task could be solved through code. To mitigate the above issues, we propose three methods to better steer LLM code/text generation and achieve a notable improvement. The costs of token lengths and runtime are thoroughly discussed for all the methods. We believe the problem of steering LLM code/text generation is critical for future research and has much space for further improvement. Project Page, Datasets, and Codes are available at https://

yongchao98.github.io/CodeSteer/.
****

Qing Feng,Tianyi Ma,Ruihao Zhu

Satisficing Regret Minimization in Bandits

Motivated by the concept of satisficing in decision-making, we consider the problem of satisficing exploration in bandit optimization. In this setting, the learner aims at finding a satisficing arm whose mean reward exceeds a certain threshold. The performance is measured by satisficing regret, which is the cumulative deficit of the chosen arm's mean reward compared to the threshold. We propose $\texttt{SELECT}$, a general algorithmic template for Satisficing REgret Minimization via SampLing and LowEr Confidence bound Testing, that attains constant satisficing regret for a wide variety of bandit optimization problems in the realizable case (i.e., whenever a satisficing arm exists). Specifically, given a class of bandit optimization problems and a corresponding learning oracle with sub-linear (standard) regret upper bound, $\texttt{SELECT}$ iteratively makes use of the oracle to identify a potential satisficing arm. Then, it collects data samples from this arm, and continuously compares the lower confidence bound of the identified arm's mean reward against the threshold value to determine if it is a satisficing arm. As a complement, $\texttt{SELECT}$ also enjoys the same (standard) regret guarantee as the oracle in the non-realizable case. Finally, we conduct numerical experiments to validate the performance of $\texttt{SELECT}$ for several popular bandit optimization settings.
****

Menghua Wu,Russell Littman,Jacob Levine,Lin Qiu,Tommaso Biancalani,David Richmond,Jan-Christian Huetter

Contextualizing biological perturbation experiments through language

High-content perturbation experiments allow scientists to probe biomolecular systems at unprecedented resolution, but experimental and analysis costs pose significant barriers to widespread adoption. Machine learning has the potential to guide efficient exploration of the perturbation space and extract novel insights from these data. However, current approaches neglect the semantic richness of the relevant biology, and their objectives are misaligned with downstream biological analyses. In this paper, we hypothesize that large language models (LLMs) present a natural medium for representing complex biological relationships and rationalizing experimental outcomes. We propose PerturbQA, a benchmark for structured reasoning over perturbation experiments. Unlike current benchmarks that primarily interrogate existing knowledge, PerturbQA is inspired by open problems in perturbation modeling: prediction of differential expression and change of direction for unseen perturbations, and gene set enrichment. We evaluate state-of-the-art machine learning and statistical approaches for modeling perturbations, as well as standard LLM reasoning strategies, and we find that current methods perform poorly on PerturbQA. As a proof of feasibility, we introduce Summer (SUMMarize, retrievE, and answeR, a simple, domain-informed LLM framework that matches or exceeds the current state-of-the-art. Our code and data are publicly available at https://github.com/genentech/PerturbQA.
****

Wenhao Wang,Adam Dziedzic,Grace C. Kim,Michael Backes,Franziska Boenisch

Captured by Captions: On Memorization and its Mitigation in CLIP Models

Multi-modal models, such as CLIP, have demonstrated strong performance in aligning visual and textual representations, excelling in tasks like image retrieval and zero-shot classification. Despite this success, the mechanisms by which these models utilize training data, particularly the role of memorization, remain unclear. In uni-modal models, both supervised and self-supervised, memorization has been shown to be essential for generalization. However, it is not well understood how these findings would apply to CLIP, which incorporates elements from both supervised learning via captions that provide a supervisory signal similar to labels, and from self-supervised learning via the contrastive objective.

To bridge this gap in understanding, we propose a formal definition of memorization in CLIP (CLIPMem) and use it to quantify memorization in CLIP models. Our results indicate that CLIP’s memorization behavior falls between the supervised an

d self-supervised paradigms, with "mis-captioned" samples exhibiting highest lev
els of memorization.
Additionally, we find that the text encoder contributes more to memorization tha
n the image encoder, suggesting that mitigation strategies should focus on the t
ext domain.
Building on these insights, we propose multiple strategies to reduce memorizatio
n while at the same time improving utility---something that had not been shown b
efore for traditional learning paradigms where reducing memorization typically r
esults in utility decrease.
****

Hanqiao Ye,Yuzhou Liu,Yangdong Liu,Shuhan Shen
NeuralPlane: Structured 3D Reconstruction in Planar Primitives with Neural Field
s
3D maps assembled from planar primitives are compact and expressive in represent
ing man-made environments. In this paper, we present **NeuralPlane**, a novel ap
proach that explores **neural** fields for multi-view 3D **plane** reconstructio
n. Our method is centered upon the core idea of distilling geometric and semanti
c cues from inconsistent 2D plane observations into a unified 3D neural represen
tation, which unlocks the full leverage of plane attributes. It is accomplished
through several key designs, including: 1) a monocular module that generates geo
metrically smooth and semantically meaningful segments known as 2D plane observa
tions, 2) a plane-guided training procedure that implicitly learns accurate 3D g
eometry from the multi-view plane observations, and 3) a self-supervised feature
 field termed *Neural Coplanarity Field* that enables the modeling of scene sema
ntics alongside the geometry. Without relying on prior plane annotations, our me
thod achieves high-fidelity reconstruction comprising planar primitives that are
 not only crisp but also well-aligned with the semantic content. Comprehensive e
xperiments on ScanNetv2 and ScanNet++ demonstrate the superiority of our method
in both geometry and semantics.
****

Yichen Wu,Hongming Piao,Long-Kai Huang,Renzhen Wang,Wanhua Li,Hanspeter Pfister,
Deyu Meng,Kede Ma,Ying Wei
SD-LoRA: Scalable Decoupled Low-Rank Adaptation for Class Incremental Learning
Continual Learning (CL) with foundation models has recently emerged as a promisi
ng paradigm to exploit abundant knowledge acquired during pre-training for tackl
ing sequential tasks. However, existing prompt-based and Low-Rank Adaptation-bas
ed (LoRA-based) methods often require expanding a prompt/LoRA pool or retaining
samples of previous tasks, which poses significant scalability challenges as the
 number of tasks grows.
To address these limitations, we propose Scalable Decoupled LoRA (SD-LoRA) for c
lass incremental learning, which continually separates the learning of the magni
tude and direction of LoRA components without rehearsal. Our empirical and theor
etical analysis reveals that SD-LoRA tends to follow a low-loss trajectory and c
onverges to an overlapping low-loss region for all learned tasks, resulting in a
n excellent stability-plasticity trade-off. Building upon these insights, we int
roduce two variants of SD-LoRA with further improved parameter efficiency. All p
arameters of SD-LoRAs can be end-to-end optimized for CL objectives. Meanwhile,
they support efficient inference by allowing direct evaluation with the finally
trained model, obviating the need for component selection. Extensive experiments
 across multiple CL benchmarks and foundation models consistently validate the e
ffectiveness of SD-LoRA. The code is available at https://github.com/WuYichen-97
/SD-Lora-CL.
****

Ayan Sengupta,Siddhant Chaudhary,Tanmoy Chakraborty
You Only Prune Once: Designing Calibration-Free Model Compression With Policy Le
arning
The ever-increasing size of large language models (LLMs) presents significant ch
allenges for deployment due to their heavy computational and memory requirements
. Current model pruning techniques attempt to alleviate these issues by relying
heavily on external calibration datasets to determine which parameters to prune

or compress, thus limiting their flexibility and scalability across different co
mpression ratios. Moreover, these methods often cause severe performance degrada
tion, particularly in downstream tasks, when subjected to higher compression rat
es. In this paper, we propose *PruneNet*, a novel model compression method that
addresses these limitations by reformulating model pruning as a policy learning
process. PruneNet decouples the pruning process from the model architecture, eli
minating the need for calibration datasets. It learns a stochastic pruning polic
y to assess parameter importance solely based on intrinsic model properties whil
e preserving the spectral structure to minimize information loss. PruneNet can c
ompress the LLaMA-2-7B model in just 15 minutes, achieving over 80\% retention o
f its zero-shot performance with a 30\% compression ratio, outperforming existin
g methods that retain only 75\% performance. Furthermore, on complex multitask l
anguage understanding tasks, PruneNet demonstrates its robustness by preserving
up to 80\% performance of the original model, proving itself a superior alternat
ive to conventional structured compression techniques.
****

Tao Feng,Yihang Sun,Jiaxuan You

GraphEval: A Lightweight Graph-Based LLM Framework for Idea Evaluation

The powerful capabilities of Large Language Models (LLMs) have led to their grow
ing use in evaluating human-generated content, particularly in evaluating resear
ch ideas within academic settings. Existing solutions primarily rely on prompt-b
ased LLM methods or fine-tuned lightweight language models for idea evaluation.
However, these methods are often unstable and struggle to comprehend the complex
 semantic information embedded in the ideas, impeding their ability to perform h
igh-quality evaluations. To address the above challenges, we propose $\texttt{Gr
aphEval}$, a lightweight graph-based LLM framework for idea evaluation. Our insi
ght is that a complex idea can be broken down into comprehensible viewpoint node
s using prompts from small LLMs. These viewpoint nodes can then be linked togeth
er through edges created from LLM-based relation extraction and/or BERT similari
ty scores. The created viewpoint-graph can be used to conveniently propagate sco
res across view-nodes to improve the robustness of the idea evaluations. In part
icular, we propose two lightweight graph-based methods for idea evaluation: (1)
GraphEval-LP: a training-free label propagation algorithm that propagates evalua
tion scores from known view-nodes to unknown nodes; (2) GraphEval-GNN: a Graph N
eural Networks (GNN) that is trained to predict the evaluation scores given the
observed graph with minimal computation resources. Moreover, to overcome LLM's l
imitation in objectively assessing the novelty of ideas, we further propose a no
velty detection model to GraphEval-GNN to enhance its capability in judging idea
 novelty. Experiments on two datasets show $\texttt{GraphEval}$ improves F1 scor
es by at least 14% with low computation and API costs. Additionally, $\texttt{Gr
aphEval}$ can effectively detect plagiarized ideas.
****

Akhilan Boopathy,Sunshine Jiang,William Yue,Jaedong Hwang,Abhiram Iyer,Ila R Fie
te

Breaking Neural Network Scaling Laws with Modularity

Modular neural networks outperform nonmodular neural networks on tasks ranging f
rom visual question answering to robotics. These performance improvements are th
ought to be due to modular networks' superior ability to model the compositional
 and combinatorial structure of real-world problems. However, a theoretical expl
anation of how modularity improves generalizability, and how to leverage task mo
dularity while training networks remains elusive. Using recent theoretical progr
ess in explaining neural network generalization, we investigate how the amount o
f training data required to generalize on a task varies with the intrinsic dimen
sionality of a task's input. We show theoretically that when applied to modularl
y structured tasks, while nonmodular networks require an exponential number of s
amples with task dimensionality, modular networks' sample complexity is independ
ent of task dimensionality: modular networks can generalize in high dimensions.
We then develop a novel learning rule for modular networks to exploit this advan
tage and empirically show the improved generalization of the rule, both in- and
out-of-distribution, on high-dimensional, modular tasks.

****
Emanuele Zangrando,Sara Venturini,Francesco Rinaldi,Francesco Tudisco
dEBORA: Efficient Bilevel Optimization-based low-Rank Adaptation
Low-rank adaptation methods are a popular approach for parameter-efficient fine-tuning of large-scale neural networks. However, selecting the optimal rank for each layer remains a challenging problem that significantly affects both performance and efficiency. In this paper, we introduce a novel bilevel optimization strategy that simultaneously trains both matrix and tensor low-rank adapters, dynamically selecting the optimal rank for each layer. Our method avoids the use of implicit differentiation in the computation of the hypergradient, and integrates a stochastic away-step variant of the Frank-Wolfe algorithm, eliminating the need for projection and providing identifiability guarantees of the optimal rank structure. This results in a highly efficient and cost-effective training scheme that adaptively allocates the parameter budget across the network layers. On top of a detailed theoretical analysis of the method, we provide different numerical experiments showcasing its effectiveness.
****
Seiji Maekawa,Hayate Iso,Nikita Bhutani
Holistic Reasoning with Long-Context LMs: A Benchmark for Database Operations on Massive Textual Data
The rapid increase in textual information means we need more efficient methods to sift through, organize, and understand it all. While retrieval-augmented generation (RAG) models excel in accessing information from large document collections, they struggle with complex tasks that require aggregation and reasoning over information spanning across multiple documents--what we call \textit{holistic reasoning}. Long-context language models (LCLMs) have great potential for managing large-scale documents, but their holistic reasoning capabilities remain unclear. In this work, we introduce HoloBench, a novel framework that brings database reasoning operations into text-based contexts, making it easier to systematically evaluate how LCLMs handle holistic reasoning across large documents. Our approach adjusts key factors such as context length, information density, distribution of information, and query complexity to evaluate LCLMs comprehensively.

Our experiments show that the amount of information in the context has a bigger influence on LCLM performance than the actual context length. Furthermore, the complexity of queries affects performance more than the amount of information, particularly for different types of queries. Interestingly, queries that involve finding maximum or minimum values are easier for LCLMs and are less affected by context length, even though they pose challenges for RAG systems. However, tasks requiring the aggregation of multiple pieces of information show a noticeable drop in accuracy as context length increases. Additionally, we find that while grouping relevant information generally improves performance, the optimal positioning varies across models. Our findings surface both the advancements and the ongoing challenges in achieving a holistic understanding of long contexts. These can guide future developments in LCLMs and set the stage for creating more robust language models for real-world applications.
****
Viggo Moro,Luiz F. O. Chamon
Solving Differential Equations with Constrained Learning
(Partial) differential equations (PDEs) are fundamental tools for describing natural phenomena, making their solution crucial in science and engineering. While traditional methods, such as the finite element method, provide reliable solutions, their accuracy is often tied to the use of computationally intensive fine meshes. Moreover, they do not naturally account for measurements or prior solutions, and any change in the problem parameters requires results to be fully recomputed. Neural network-based approaches, such as physics-informed neural networks and neural operators, offer a mesh-free alternative by directly fitting those models to the PDE solution. They can also integrate prior knowledge and tackle entire families of PDEs by simply aggregating additional training losses. Nevertheless, they are highly sensitive to hyperparameters such as collocation points and

the weights associated with each loss. This paper addresses these challenges by developing a science-constrained learning (SCL) framework. It demonstrates that finding a (weak) solution of a PDE is equivalent to solving a constrained learning problem with worst-case losses. This explains the limitations of previous methods that minimize the expected value of aggregated losses. SCL also organically integrates structural constraints (e.g., invariances) and (partial) measurements or known solutions. The resulting constrained learning problems can be tackled using a practical algorithm that yields accurate solutions across a variety of PDEs, neural network architectures, and prior knowledge levels without extensive hyperparameter tuning and sometimes even at a lower computational cost.
****
Dan Greenstein,Elazar Gershuni,Ilan Ben-Bassat,Yaroslav Fyodorov,Ran Moshe,Fiana Raiber,Alex Shtoff,Oren Somekh,Nadav Hallak
A Stochastic Approach to the Subset Selection Problem via Mirror Descent
The subset selection problem is fundamental in machine learning and other fields of computer science.
We introduce a stochastic formulation for the minimum cost subset selection problem in a black box setting, in which only the subset metric value is available.
Subsequently, we can handle two-stage schemes, with an outer subset-selection component and an inner subset cost evaluation component. We propose formulating the subset selection problem in a stochastic manner by choosing subsets at random from a distribution whose parameters are learned. Two stochastic formulations are proposed.
The first explicitly restricts the subset's cardinality, and the second yields the desired cardinality in expectation.
The distribution is parameterized by a decision variable, which we optimize using Stochastic Mirror Descent.
Our choice of distributions yields constructive closed-form unbiased stochastic gradient formulas and convergence guarantees, including a rate with favorable dependency on the problem parameters.
Empirical evaluation of selecting a subset of layers in transfer learning complements our theoretical findings and demonstrates the potential benefits of our approach.
****
Guancheng Wan,Zitong Shi,Wenke Huang,Guibin Zhang,Dacheng Tao,Mang Ye
Energy-based Backdoor Defense Against Federated Graph Learning
Federated Graph Learning is rapidly evolving as a privacy-preserving collaborative approach. However, backdoor attacks are increasingly undermining federated systems by injecting carefully designed triggers that lead to the model making incorrect predictions. Trigger structures and injection locations in Federated Graph Learning are more diverse, making traditional federated defense methods less effective. In our work, we propose an effective Federated Graph Backdoor Defense using Topological Graph Energy (FedTGE). At the local client level, it injects distribution knowledge into the local model, assigning low energy to benign samples and high energy to the constructed malicious substitutes, and selects benign clients through clustering. At the global server level, the energy elements uploaded by each client are treated as new nodes to construct a global energy graph for energy propagation, making the selected clients' energy elements more similar and further adjusting the aggregation weights. Our method can handle high data heterogeneity, does not require a validation dataset, and is effective under both small and large malicious proportions. Extensive results on various settings of federated graph scenarios under backdoor attacks validate the effectiveness of this approach.
****
Xuerui Qiu,Malu Zhang,Jieyuan Zhang,Wenjie Wei,Honglin Cao,Junsheng Guo,Rui-Jie Zhu,Yimeng Shan,Yang Yang,Haizhou Li
Quantized Spike-driven Transformer
Spiking neural networks (SNNs) are emerging as a promising energy-efficient alternative to traditional artificial neural networks (ANNs) due to their spike-driven paradigm.

However, recent research in the SNN domain has mainly focused on enhancing accur
acy by designing large-scale Transformer structures, which typically rely on sub
stantial computational resources, limiting their deployment on resource-constrai
ned devices.
To overcome this challenge, we propose a quantized spike-driven Transformer base
line (QSD-Transformer), which achieves reduced resource demands by utilizing a l
ow bit-width parameter.
Regrettably, the QSD-Transformer often suffers from severe performance degradati
on.
In this paper, we first conduct empirical analysis and find that the bimodal dis
tribution of quantized spike-driven self-attention (Q-SDSA) leads to spike infor
mation distortion (SID) during quantization, causing significant performance deg
radation. To mitigate this issue, we take inspiration from mutual information en
tropy and propose a bi-level optimization strategy to rectify the information di
stribution in Q-SDSA.
Specifically, at the lower level, we introduce an information-enhanced LIF to re
ctify the information distribution in Q-SDSA.
At the upper level, we propose a fine-grained distillation scheme for the QSD-Tr
ansformer to align the distribution in Q-SDSA with that in the counterpart ANN.
By integrating the bi-level optimization strategy, the QSD-Transformer can attai
n enhanced energy efficiency without sacrificing its high-performance advantage.
We validate the QSD-Transformer on various visual tasks, and experimental result
s indicate that our method achieves state-of-the-art results in the SNN domain.
For instance, when compared to the prior SNN benchmark on ImageNet, the QSD-Tran
sformer achieves 80.3\% top-1 accuracy, accompanied by significant reductions of
 6.0$\times$ and 8.1$\times$ in power consumption and model size, respectively.
Code is available at https://github.com/bollossom/QSD-Transformer.
****
Renhao Wang,Kevin Frans,Pieter Abbeel,Sergey Levine,Alexei A Efros
Prioritized Generative Replay
Sample-efficient online reinforcement learning often uses replay buffers to stor
e experience for reuse when updating the value function.
However, uniform replay is inefficient, since certain classes of transitions can
 be more relevant to learning. While prioritization of more useful samples is he
lpful, this strategy can also lead to overfitting, as useful samples are likely
to be more rare.
In this work, we instead propose a prioritized, parametric version of an agent's
 memory, using generative models to capture online experience. This paradigm ena
bles (1) densification of past experience, with new generations that benefit fro
m the generative model's generalization capacity and (2) guidance via a family o
f ``relevance functions'' that push these generations towards more useful parts
of an agent's acquired history. We show this recipe can be instantiated using co
nditional diffusion models and simple relevance functions such as curiosity- or
value-based metrics.
Our approach consistently improves performance and sample efficiency in both sta
te- and pixel-based domains. We expose the mechanisms underlying these gains, sh
owing how guidance promotes diversity in our generated transitions and reduces o
verfitting. We also showcase how our approach can train policies with even highe
r update-to-data ratios than before, opening up avenues to better scale online R
L agents.
****
Maxime Méloux,Silviu Maniu,François Portet,Maxime Peyrard
Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiabl
e?
As AI systems are increasingly deployed in high-stakes applications, ensuring th
eir interpretability is essential. Mechanistic Interpretability (MI) aims to rev
erse-engineer neural networks by extracting human-understandable algorithms embe
dded within their structures to explain their behavior. This work systematically
 examines a fundamental question: for a fixed behavior to explain, and under the
 criteria that MI sets for itself, are we guaranteed a unique explanation? Drawi

ng an analogy with the concept of identifiability in statistics, which ensures t
he uniqueness of parameters inferred from data under specific modeling assumptio
ns, we speak about the identifiability of explanations produced by MI.

We identify two broad strategies to produce MI explanations: (i) "where-then-wha
t", which first identifies a subset of the network (a circuit) that replicates t
he model's behavior before deriving its interpretation, and (ii) "what-then-wher
e", which begins with candidate explanatory algorithms and searches in the activ
ation subspaces of the neural model where the candidate algorithm may be impleme
nted, relying on notions of causal alignment between the states of the candidate
 algorithm and the neural network.

We systematically test the identifiability of both strategies using simple tasks
 (learning Boolean functions) and multi-layer perceptrons small enough to allow
a complete enumeration of candidate explanations. Our experiments reveal overwhe
lming evidence of non-identifiability in all cases: multiple circuits can replic
ate model behavior, multiple interpretations can exist for a circuit, several al
gorithms can be causally aligned with the neural network, and a single algorithm
 can be causally aligned with different subspaces of the network.

We discuss whether the unicity intuition is necessary. One could adopt a pragmat
ic stance, requiring explanations only to meet predictive and/or manipulability
standards. However, if unicity is considered essential, e.g., to provide a sense
 of understanding, we also discuss less permissive criteria. Finally, we also re
fer to the inner interpretability framework that demands explanation to be valid
ated by multiple complementary criteria. This work aims to contribute constructi
vely to the ongoing effort to formalize what we expect from explanations in AI.
****
Weitai Kang,Mengxue Qu,Jyoti Kini,Yunchao Wei,Mubarak Shah,Yan Yan
Intent3D: 3D Object Detection in RGB-D Scans Based on Human Intention
In real-life scenarios, humans seek out objects in the 3D world to fulfill their
 daily needs or intentions. This inspires us to introduce 3D intention groundin
g, a new task in 3D object detection employing RGB-D, based on human intention,
 such as "I want something to support my back." Closely related, 3D visual groun
ding focuses on understanding human reference. To achieve detection based on hum
an intention, it relies on humans to observe the scene, reason out the target th
at aligns with their intention ("pillow" in this case), and finally provide a re
ference to the AI system, such as "A pillow on the couch". Instead, 3D intention
 grounding challenges AI agents to automatically observe, reason and detect the
desired target solely based on human intention. To tackle this challenge, we int
roduce the new Intent3D dataset, consisting of 44,990 intention texts associated
 with 209 fine-grained classes from 1,042 scenes of the ScanNet dataset. We also
 establish several baselines based on different language-based 3D object detecti
on models on our benchmark. Finally, we propose IntentNet, our unique approach,
designed to tackle this intention-based detection problem. It focuses on three k
ey aspects: intention understanding, reasoning to identify object candidates, an
d cascaded adaptive learning that leverages the intrinsic priority logic of diff
erent losses for multiple objective optimization.
****
Leon Hetzel,Johanna Sommer,Bastian Rieck,Fabian J Theis,Stephan Günnemann
MAGNet: Motif-Agnostic Generation of Molecules from Scaffolds
Recent advances in machine learning for molecules exhibit great potential for fa
cilitating drug discovery from in silico predictions.
Most models for molecule generation rely on the decomposition of molecules into
frequently occurring substructures (motifs), from which they generate novel comp
ounds.
While motif representations greatly aid in learning molecular distributions, suc
h methods fail to represent substructures beyond their known motif set, posing a
 fundamental limitation for discovering novel compounds.
To address this limitation and enhance structural expressivity, we propose to se

parate structure from features by abstracting motifs to scaffolds and, subsequen tly, allocating atom and bond types.
To this end, we introduce a novel factorisation of the molecules' data distribut ion that considers the entire molecular context and facilitates learning adequat e assignments of atoms and bonds to scaffolds. Complementary to this, we propose MAGNet, the first model to freely learn motifs. Importantly, we demonstrate tha t MAGNet's improved expressivity leads to molecules with more structural diversi ty and, at the same time, diverse atom and bond assignments.
****
Guangchen Lan,Dong-Jun Han,Abolfazl Hashemi,Vaneet Aggarwal,Christopher Brinton
Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algo rithm Design and Convergence Analysis
To improve the efficiency of reinforcement learning (RL), we propose a novel asy nchronous federated reinforcement learning (FedRL) framework termed AFedPG, whic h constructs a global model through collaboration among N agents using policy gradient (PG) updates. To address the challenge of lagged policies in asynchrono us settings, we design a delay-adaptive lookahead technique *specifically for Fe dRL* that can effectively handle heterogeneous arrival times of policy gradients . We analyze the theoretical global convergence bound of AFedPG, and characteriz e the advantage of the proposed algorithm in terms of both the sample complexity and time complexity. Specifically, our AFedPG method achieves $\mathcal{O}(\fra c{{\epsilon}^{-2.5}}{N})$ sample complexity for global convergence at each agent on average. Compared to the single agent setting with $\mathcal{O}(\epsilon^{-2 .5})$ sample complexity, it enjoys a linear speedup with respect to the number o f agents. Moreover, compared to synchronous FedPG, AFedPG improves the time comp lexity from $\mathcal{O}(\frac{t_{\max}}{N})$ to $\mathcal{O}({\sum_{i=1}^{N} \f rac{1}{t_{i}}})^{-1}$, where t_{i} denotes the time consumption in each iterat ion at agent i, and $t_{\max}$ is the largest one. The latter complexity $\mat hcal{O}({\sum_{i=1}^{N} \frac{1}{t_{i}}})^{-1}$ is always smaller than the forme r one, and this improvement becomes significant in large-scale federated setting s with heterogeneous computing powers ($t_{\max}\gg t_{\min}$). Finally, we empi rically verify the improved performance of AFedPG in four widely used MuJoCo env ironments with varying numbers of agents. We also demonstrate the advantages of AFedPG in various computing heterogeneity scenarios.
****
Qian Liu,Xiaosen Zheng,Niklas Muennighoff,Guangtao Zeng,Longxu Dou,Tianyu Pang,J ing Jiang,Min Lin
RegMix: Data Mixture as Regression for Language Model Pre-training
The data mixture for large language model pre-training significantly impacts per formance, yet how to determine an effective mixture remains unclear. We propose RegMix to automatically identify a high-performing data mixture by formulating i t as a regression task. RegMix trains many small models on diverse data mixtures , uses regression to predict performance of unseen mixtures, and applies the bes t predicted mixture to train a large-scale model with orders of magnitude more c ompute. To empirically validate RegMix, we train 512 models with 1M parameters f or 1B tokens to fit the regression model and predict the best data mixture. Usin g this mixture we train a 1B parameter model for 25B tokens (i.e. 1000× larger a nd 25× longer) which we find performs best among 64 candidate 1B parameter model s with other mixtures. Furthermore, RegMix consistently outperforms human select ion in experiments involving models up to 7B models trained on 100B tokens, whil e matching or exceeding DoReMi using just 10% of the computational resources. Ou r experiments also show that (1) Data mixtures significantly impact performance; (2) Web corpora rather than data perceived as high-quality like Wikipedia have the strongest positive correlation with downstream performance; (3) Domains inte ract in complex ways often contradicting common sense, thus automatic approaches like RegMix are needed; (4) Data mixture effects transcend scaling laws. Our co de is available at https://github.com/sail-sg/regmix.
****
Jae-Jun Lee,Sung Whan Yoon
Can One Modality Model Synergize Training of Other Modality Models?

Learning with multiple modalities has recently demonstrated significant gains in many domains by maximizing the shared information across modalities. However, the current approaches strongly rely on high-quality paired datasets, which allow co-training from the paired labels from different modalities. In this context, we raise a pivotal question: Can a model with one modality synergize the training of other models with the different modalities, even without the paired multimodal labels? Our answer is 'Yes'. As a figurative description, we argue that a writer, i.e., a language model, can promote the training of a painter, i.e., a visual model, even without the paired ground truth of text and image. We theoretically argue that a superior representation can be achieved by the synergy between two different modalities without paired supervision. As proofs of concept, we broadly confirm the considerable performance gains from the synergy among visual, language, and audio models. From a theoretical viewpoint, we first establish a mathematical foundation of the synergy between two different modality models, where each one is trained with its own modality. From a practical viewpoint, our work aims to broaden the scope of multimodal learning to encompass the synergistic usage of single-modality models, relieving a strong limitation of paired supervision. The code is available at https://github.com/johnjaejunlee95/synergistic-multimodal.

****

Shuangqi Li,Hieu Le,Jingyi Xu,Mathieu Salzmann

Enhancing Compositional Text-to-Image Generation with Reliable Random Seeds

Text-to-image diffusion models have demonstrated remarkable capability in generating realistic images from arbitrary text prompts. However, they often produce inconsistent results for compositional prompts such as "two dogs" or "a penguin on the right of a bowl". Understanding these inconsistencies is crucial for reliable image generation. In this paper, we highlight the significant role of initial noise in these inconsistencies, where certain noise patterns are more reliable for compositional prompts than others. Our analyses reveal that different initial random seeds tend to guide the model to place objects in distinct image areas, potentially adhering to specific patterns of camera angles and image composition associated with the seed. To improve the model's compositional ability, we propose a method for mining these reliable cases, resulting in a curated training set of generated images without requiring any manual annotation.

By fine-tuning text-to-image models on these generated images, we significantly enhance their compositional capabilities. For numerical composition, we observe relative increases of 29.3\% and 19.5\% for Stable Diffusion and PixArt-α, respectively. Spatial composition sees even larger gains, with 60.7\% for Stable Diffusion and 21.1\% for PixArt-α.

****

Tristan Tomilin,Meng Fang,Mykola Pechenizkiy

HASARD: A Benchmark for Vision-Based Safe Reinforcement Learning in Embodied Agents

Advancing safe autonomous systems through reinforcement learning (RL) requires robust benchmarks to evaluate performance, analyze methods, and assess agent competencies. Humans primarily rely on embodied visual perception to safely navigate and interact with their surroundings, making it a valuable capability for RL agents. However, existing vision-based 3D benchmarks only consider simple navigation tasks. To address this shortcoming, we introduce **HASARD**, a suite of diverse and complex tasks to **HA**rness **SA**fe **R**L with **D**oom, requiring strategic decision-making, comprehending spatial relationships, and predicting the short-term future. HASARD features three difficulty levels and two action spaces. An empirical evaluation of popular baseline methods demonstrates the benchmark's complexity, unique challenges, and reward-cost trade-offs. Visualizing agent navigation during training with top-down heatmaps provides insight into a method's learning process. Incrementally training across difficulty levels offers an implicit learning curriculum. HASARD is the first safe RL benchmark to exclusively target egocentric vision-based learning, offering a cost-effective and insightful way to explore the potential and boundaries of current and future safe RL methods. The environments and baseline implementations are open-sourced.

****
Nian Ran,Peng Xiao,Yue Wang,Wesley Shi,Jianxin Lin,Qi Meng,Richard Allmendinger
HR-Extreme: A High-Resolution Dataset for Extreme Weather Forecasting
The application of large deep learning models in weather forecasting has led to
significant advancements in the field, including higher-resolution forecasting a
nd
extended prediction periods exemplified by models such as Pangu and Fuxi. Despit
e
these successes, previous research has largely been characterized by the neglect
of extreme weather events, and the availability of datasets specifically curated
 for
such events remains limited. Given the critical importance of accurately forecas
ting
extreme weather, this study introduces a comprehensive dataset that incorporates
high-resolution extreme weather cases derived from the High-Resolution Rapid
Refresh (HRRR) data, a 3-km real-time dataset provided by NOAA. We also
evaluate the current state-of-the-art deep learning models and Numerical Weather
Prediction (NWP) systems on HR-Extreme, and provide a improved baseline
deep learning model called HR-Heim which has superior performance on both
general loss and HR-Extreme compared to others. Our results reveal that the
errors of extreme weather cases are significantly larger than overall forecast e
rror,
highlighting them as an crucial source of loss in weather prediction. These find
ings
underscore the necessity for future research to focus on improving the accuracy
of
extreme weather forecasts to enhance their practical utility
****
Ségolène Tiffany Martin,Anne Gagneux,Paul Hagemann,Gabriele Steidl
PnP-Flow: Plug-and-Play Image Restoration with Flow Matching
In this paper, we introduce Plug-and-Play (PnP) Flow Matching, an algorithm for
solving imaging inverse problems. PnP methods leverage the strength of pre-train
ed denoisers, often deep neural networks, by integrating them in optimization sc
hemes. While they achieve state-of-the-art performance on various inverse proble
ms in imaging, PnP approaches face inherent limitations on more generative tasks
 like inpainting. On the other hand, generative models such as Flow Matching pus
hed the boundary in image sampling yet lack a clear method for efficient use in
image restoration. We propose to combine the PnP framework with Flow Matching (F
M) by defining a time-dependent denoiser using a pre-trained FM model. Our algor
ithm alternates between gradient descent steps on the data-fidelity term, reproj
ections onto the learned FM path, and denoising. Notably, our method is computat
ionally efficient and memory-friendly, as it avoids backpropagation through ODEs
 and trace computations. We evaluate its performance on denoising, super-resolut
ion, deblurring, and inpainting tasks, demonstrating superior results compared t
o existing PnP algorithms and Flow Matching based state-of-the-art methods. Code
 available at https://github.com/annegnx/PnP-Flow.
****
Michael Aerni,Javier Rando,Edoardo Debenedetti,Nicholas Carlini,Daphne Ippolito,
Florian Tramèr
Measuring Non-Adversarial Reproduction of Training Data in Large Language Models
Large language models memorize parts of their training data. Memorizing short sn
ippets and facts is required to answer questions about the world and to be fluen
t in any language. But models have also been shown to reproduce long verbatim se
quences of memorized text when prompted by a motivated adversary. In this work,
we investigate an intermediate regime of memorization that we call non-adversari
al reproduction, where we quantify the overlap between model responses and pretr
aining data when responding to natural and benign prompts. For a variety of inno
cuous prompt categories (e.g., writing a letter or a tutorial), we show that up
to 15% of the text output by popular conversational language models overlaps wit
h snippets from the Internet. In worst cases, we find generations where 100% of

the content can be found exactly online. For the same tasks, we find that human-written text has far less overlap with Internet data. We further study whether prompting strategies can close this reproduction gap between models and humans. While appropriate prompting can reduce non-adversarial reproduction on average, we find that mitigating worst-case reproduction of training data requires stronger defenses—even for benign interactions.
****
Haoxiang Wang,Tao Yu,Hui Qiao,Qionghai Dai
Neural Fluid Simulation on Geometric Surfaces
Incompressible fluid on the surface is an interesting research area in the fluid simulation, which is the fundamental building block in visual effects, design of liquid crystal films, scientific analyses of atmospheric and oceanic phenomena, etc. The task brings two key challenges: the extension of the physical laws on 3D surfaces and the preservation of the energy and volume. Traditional methods rely on grids or meshes for spatial discretization, which leads to high memory consumption and a lack of robustness and adaptivity for various mesh qualities and representations. Many implicit representations based simulators like INSR are proposed for the storage efficiency and continuity, but they face challenges in the surface simulation and the energy dissipation. We propose a neural physical simulation framework on the surface with the implicit neural representation. Our method constructs a parameterized vector field with the exterior calculus and Closest Point Method on the surfaces, which guarantees the divergence-free property and enables the simulation on different surface representations (e.g. implicit neural represented surfaces). We further adopt a corresponding covariant derivative based advection process for surface flow dynamics and energy preservation. Our method shows higher accuracy, flexibility and memory-efficiency in the simulations of various surfaces with low energy dissipation. Numerical studies also highlight the potential of our framework across different practical applications such as vorticity shape generation and vector field Helmholtz decomposition.
****
Siddhartha Gairola,Moritz Böhle,Francesco Locatello,Bernt Schiele
How to Probe: Simple Yet Effective Techniques for Improving Post-hoc Explanations
Post-hoc importance attribution methods are a popular tool for “explaining” Deep Neural Networks (DNNs) and are inherently based on the assumption that the explanations can be applied independently of how the models were trained. Contrarily, in this work we bring forward empirical evidence that challenges this very notion. Surprisingly, we discover a strong dependency on and demonstrate that the training details of a pre-trained model’s classification layer (<10% of model parameters) play a crucial role, much more than the pre-training scheme itself. This is of high practical relevance: (1) as techniques for pre-training models are becoming increasingly diverse, understanding the interplay between these techniques and attribution methods is critical; (2) it sheds light on an important yet overlooked assumption of post-hoc attribution methods which can drastically impact model explanations and how they are interpreted eventually. With this finding we also present simple yet effective adjustments to the classification layers, that can significantly enhance the quality of model explanations. We validate our findings across several visual pre-training frameworks (fully-supervised, self-supervised, contrastive vision-language training) and analyse how they impact explanations for a wide range of attribution methods on a diverse set of evaluation metrics.
****
Chong Cheng,Gaochao Song,Yiyang Yao,Qinzheng Zhou,Gangjian Zhang,Hao Wang
Graph-Guided Scene Reconstruction from Images with 3D Gaussian Splatting
This paper investigates an open research challenge of reconstructing high-quality, large-scale 3D open scenes from images. It is observed existing methods have various limitations, such as requiring precise camera poses for input and dense viewpoints for supervision.
To perform effective and efficient 3D scene reconstruction, we propose a novel graph-guided 3D scene reconstruction framework, GraphGS. Specifically, given a se

t of images captured by RGB cameras on a scene, we first design a spatial prior-based scene structure estimation method. This is then used to create a camera graph that includes information about the camera topology. Further, we propose to apply the graph-guided multi-view consistency constraint and adaptive sampling strategy to the 3D Gaussian Splatting optimization process. This greatly alleviates the issue of Gaussian points overfitting to specific sparse viewpoints and expedites the 3D reconstruction process. We demonstrate GraphGS achieves high-fidelity 3D reconstruction from images, which presents state-of-the-art performance through quantitative and qualitative evaluation across multiple datasets.
****

Werner Zellinger

Binary Losses for Density Ratio Estimation

Estimating the ratio of two probability densities from a finite number of observations is a central machine learning problem. A common approach is to construct estimators using binary classifiers that distinguish observations from the two densities. However, the accuracy of these estimators depends on the choice of the binary loss function, raising the question of which loss function to choose based on desired error properties. For example, traditional loss functions, such as logistic or boosting loss, prioritize accurate estimation of small density ratio values over large ones, even though the latter are more critical in many applications.

In this work, we start with prescribed error measures in a class of Bregman divergences and characterize all loss functions that result in density ratio estimators with small error. Our characterization extends results on composite binary losses from Reid & Williamson (2010) and their connection to density ratio estimation as identified by Menon & Ong (2016). As a result, we obtain a simple recipe for constructing loss functions with certain properties, such as those that prioritize an accurate estimation of large density ratio values. Our novel loss functions outperform related approaches for resolving parameter choice issues of 11 deep domain adaptation algorithms in average performance across 484 real-world tasks including sensor signals, texts, and images.
****

Kaiwen Zheng,Guande He,Jianfei Chen,Fan Bao,Jun Zhu

Elucidating the Preconditioning in Consistency Distillation

Consistency distillation is a prevalent way for accelerating diffusion models adopted in consistency (trajectory) models, in which a student model is trained to traverse backward on the probability flow (PF) ordinary differential equation (ODE) trajectory determined by the teacher model. Preconditioning is a vital technique for stabilizing consistency distillation, by linear combining the input data and the network output with pre-defined coefficients as the consistency function. It imposes the boundary condition of consistency functions without restricting the form and expressiveness of the neural network. However, previous preconditionings are hand-crafted and may be suboptimal choices. In this work, we offer the first theoretical insights into the preconditioning in consistency distillation, by elucidating its design criteria and the connection to the teacher ODE trajectory. Based on these analyses, we further propose a principled way dubbed \textit{Analytic-Precond} to analytically optimize the preconditioning according to the consistency gap (defined as the gap between the teacher denoiser and the optimal student denoiser) on a generalized teacher ODE. We demonstrate that Analytic-Precond can facilitate the learning of trajectory jumpers, enhance the alignment of the student trajectory with the teacher's, and achieve $2\times$ to $3\times$ training acceleration of consistency trajectory models in multi-step generation across various datasets.
****

Zihan Ding,Jiahui Fu,Si Liu,Hongyu Li,Siheng Chen,Hongsheng Li,Shifeng Zhang,Xu Zhou

Point Cluster: A Compact Message Unit for Communication-Efficient Collaborative Perception

The objective of the collaborative perception task is to enhance the individual

agent's perception capability through message communication among neighboring ag
ents. A central challenge lies in optimizing the inherent trade-off between perc
eption ability and communication cost. To tackle this bottleneck issue, we argue
 that a good message unit should encapsulate both semantic and structural inform
ation in a sparse format, a feature not present in prior approaches. In this pap
er, we innovatively propose a compact message unit, namely point cluster, whose
core idea is to represent potential objects efficiently with explicitly decouple
d low-level structure information and high-level semantic information. Building
upon this new message unit, we propose a comprehensive framework CPPC for commun
ication-efficient collaborative perception. The core principle of CPPC is twofol
d: first, through strategical point sampling, structure information can be well
preserved with a few key points, which can significantly reduce communication co
st; second, the sequence format of point clusters enables efficient message aggr
egation by set matching and merging, thereby eliminating unnecessary computation
 generated when aligning squared BEV maps, especially for long-range collaborati
on. To handle time latency and pose errors encountered in real-world scenarios,
we also carefully design parameter-free solutions that can adapt to different no
isy levels without finetuning. Experiments on two widely recognized collaborativ
e perception benchmarks showcase the superior performance of our method compared
 to the previous state-of-the-art approaches.
****

Zian Li,Xiyuan Wang,Shijia Kang,Muhan Zhang
On the Completeness of Invariant Geometric Deep Learning Models
Invariant models, one important class of geometric deep learning models, are cap
able of generating meaningful geometric representations by leveraging informativ
e geometric features in point clouds. These models are characterized by their si
mplicity, good experimental results and computational efficiency. However, their
 theoretical expressive power still remains unclear, restricting a deeper unders
tanding of the potential of such models. In this work, we concentrate on charact
erizing the theoretical expressiveness of a wide range of invariant models under
 fully-connected conditions. We first rigorously characterize the expressivene
ss of the most classic invariant model, message-passing neural networks incorpor
ating distance (DisGNN), restricting its unidentifiable cases to be only highly
symmetric point clouds. We then prove that GeoNGNN, the geometric counterpart of
 one of the simplest subgraph graph neural networks, can effectively break these
 corner cases' symmetry and thus achieve E(3)-completeness. By leveraging GeoNGN
N as a theoretical tool, we further prove that: 1) most subgraph GNNs developed
in traditional graph learning can be seamlessly extended to geometric scenarios
with E(3)-completeness; 2) DimeNet, GemNet and SphereNet, three well-established
 invariant models, are also all capable of achieving E(3)-completeness. Our theo
retical results fill the gap in the expressive power of invariant models, contri
buting to a rigorous and comprehensive understanding of their capabilities.
****

Prithaj Banerjee,Harish Guruprasad Ramaswamy,Mahesh Lorik Yadav,Chandra Shekar L
akshminarayanan
Deep Networks Learn Features From Local Discontinuities in the Label Function
Deep neural networks outperform kernel machines on several datasets due to featu
re learning that happens during gradient descent training. In this paper, we ana
lyze the mechanism through which feature learning happens and use a notion of fe
atures that corresponds to discontinuities in the true label function. We hypoth
esize that the core feature learning mechanism is label function discontinuities
 attracting model function discontinuities during training. To test this hypothe
sis, we perform experiments on classification data where the true label function
 is given by an oblique decision tree. This setup allows easy enumeration of lab
el function discontinuities, while still remaining intractable for static kernel
/linear methods. We then design/construct a novel deep architecture called a Dee
p Linearly Gated Network (DLGN), whose discontinuities in the input space can be
 easily enumerated. In this setup, we provide supporting evidence demonstrating
 the movement of model function discontinuities towards the label function disco
ntinuities during training. The easy enumerability of discontinuities in the DLG

N also enables greater mechanistic interpretability. We demonstrate this by extr
acting the parameters of a high-accuracy decision tree from the parameters of a
DLGN. We also show that the DLGN is competitive with ReLU networks and other tre
e-learning algorithms on several real-world tabular datasets.
****
Yan Scholten,Stephan Günnemann,Leo Schwinn
A Probabilistic Perspective on Unlearning and Alignment for Large Language Model
s
Comprehensive evaluation of Large Language Models (LLMs) is an open research pro
blem. Existing evaluations rely on deterministic point estimates generated via g
reedy decoding. However, we find that deterministic evaluations fail to capture
the whole output distribution of a model, yielding inaccurate estimations of mod
el capabilities. This is particularly problematic in critical contexts such as u
nlearning and alignment, where precise model evaluations are crucial. To remedy
this, we introduce the first formal probabilistic evaluation framework for LLMs.
 Namely, we propose novel metrics with high probability guarantees concerning th
e output distribution of a model. Our metrics are application-independent and al
low practitioners to make more reliable estimates about model capabilities befor
e deployment. Our experimental analysis reveals that deterministic evaluations f
alsely indicate successful unlearning and alignment, whereas our probabilistic e
valuations better capture model capabilities. We show how to overcome challenges
 associated with probabilistic outputs in a case study on unlearning by introduc
ing (1) a novel loss based on entropy optimization, and (2) adaptive temperature
 scaling. We demonstrate that our approach significantly enhances unlearning in
probabilistic settings on recent benchmarks. Overall, our proposed shift from po
int estimates to probabilistic evaluations of output distributions represents an
 important step toward comprehensive evaluations of LLMs.
****
Jack Sandberg,Niklas Åkerblom,Morteza Haghir Chehreghani
Bayesian Analysis of Combinatorial Gaussian Process Bandits
We consider the combinatorial volatile Gaussian process (GP) semi-bandit problem
. Each round, an agent is provided a set of available base arms and must select
a subset of them to maximize the long-term cumulative reward. We study the Bayes
ian setting and provide novel Bayesian cumulative regret bounds for three GP-bas
ed algorithms: GP-UCB, GP-BayesUCB and GP-TS. Our bounds extend previous results
 for GP-UCB and GP-TS to the \emph{infinite}, \emph{volatile} and \emph{combinat
orial} setting, and to the best of our knowledge, we provide the first regret bo
und for GP-BayesUCB. Volatile arms encompass other widely considered bandit prob
lems such as contextual bandits.
Furthermore, we employ our framework to address the challenging real-world probl
em of online energy-efficient navigation, where we demonstrate its effectiveness
 compared to the alternatives.
****
Qingyang Zhang,Yatao Bian,Xinke Kong,Peilin Zhao,Changqing Zhang
COME: Test-time Adaption by Conservatively Minimizing Entropy
Machine learning models must continuously self-adjust themselves for novel data
distribution in the open world. As the predominant principle, entropy minimizati
on (EM) has been proven to be a simple yet effective cornerstone in existing tes
t-time adaption (TTA) methods. While unfortunately its fatal limitation (i.e., o
verconfidence) tends to result in model collapse. For this issue, we propose to
\textbf{\texttt{Co}}nservatively \textbf{\texttt{M}}inimize the \textbf{\texttt{
E}}ntropy (\texttt{COME}), which is a simple drop-in replacement of traditional
EM to elegantly address the limitation. In essence, \texttt{COME} explicitly mod
els the uncertainty by characterizing a Dirichlet prior distribution over model
predictions during TTA. By doing so, \texttt{COME} naturally regularizes the mod
el to favor conservative confidence on unreliable samples. Theoretically, we pro
vide a preliminary analysis to reveal the ability of \texttt{COME} in enhancing
the optimization stability by introducing a data-adaptive lower bound on the ent
ropy. Empirically, our method achieves state-of-the-art performance on commonly
used benchmarks, showing significant improvements in terms of classification acc

uracy and uncertainty estimation under various settings including standard, life -long and open-world TTA, i.e., up to 34.5% improvement on accuracy and 15.1% on false positive rate. Our code is available at: \href{https://github.com/B lueWhaleLab/COME}{https://github.com/BlueWhaleLab/COME}.
****
Chenxi Wang,Xiang Chen,Ningyu Zhang,Bozhong Tian,Haoming Xu,Shumin Deng,Huajun C hen
MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation
Multimodal Large Language Models (MLLMs) frequently exhibit hallucination phenom ena, but the underlying reasons remain poorly understood. In this paper, we pres ent an empirical analysis and find that, although MLLMs incorrectly generate the objects in the final output, they are actually able to recognize visual objects in the preceding layers. We speculate that this may be due to the strong knowle dge priors of the language model suppressing the visual information, leading to hallucinations. Motivated by this, we propose a novel dynamic correction decodin g method for MLLMs DeCo, which adaptively selects the appropriate preceding laye rs and proportionally integrates knowledge into the final layer to adjust the ou tput logits. Note that DeCo is model agnostic and can be seamlessly incorporated with various classic decoding strategies and applied to different MLLMs. We eva luate DeCo on widely-used benchmarks, demonstrating that it can reduce hallucina tion rates by a large margin compared to baselines, highlighting its potential t o mitigate hallucinations. Code is available at https://github.com/zjunlp/DeCo.
****
Nan Jiang,Chengxiao Wang,Kevin Liu,Xiangzhe Xu,Lin Tan,Xiangyu Zhang,Petr Babkin
Nova: Generative Language Models for Assembly Code with Hierarchical Attention a nd Contrastive Learning
Binary code analysis is the foundation of crucial tasks in the security domain; thus building effective binary analysis techniques is more important than ever. Large language models (LLMs) although have brought impressive improvement to sou rce code tasks, do not directly generalize to assembly code due to the unique ch allenges of assembly: (1) the low information density of assembly and (2) the di verse optimizations in assembly code. To overcome these challenges, this work pr oposes a hierarchical attention mechanism that builds attention summaries to cap ture the semantics more effectively and designs contrastive learning objectives to train LLMs to learn assembly optimization. Equipped with these techniques, th is work develops Nova, a generative LLM for assembly code. Nova outperforms exis ting techniques on binary code decompilation by up to 14.84 -- 21.58% higher Pas s@1 and Pass@10, and outperforms the latest binary code similarity detection tec hniques by up to 6.17% Recall@1, showing promising abilities on both assembly ge neration and understanding tasks.
****
Chuanyang Zheng
IFORMER: INTEGRATING CONVNET AND TRANSFORMER FOR MOBILE APPLICATION
We present a new family of mobile hybrid vision networks, called iFormer, with a focus on optimizing latency and accuracy on mobile applications. iFormer effect ively integrates the fast local representation capacity of convolution with the efficient global modeling ability of self-attention. The local interactions are derived from transforming a standard convolutional network, \textit{i.e.}, ConvN eXt, to design a more lightweight mobile network. Our newly introduced mobile mo dulation attention removes memory-intensive operations in MHA and employs an eff icient modulation mechanism to boost dynamic global representational capacity. W e conduct comprehensive experiments demonstrating that iFormer outperforms exist ing lightweight networks across various tasks. Notably, iFormer achieves an impr essive Top-1 accuracy of 80.4% on ImageNet-1k with a latency of only 1.10 ms on an iPhone 13, surpassing the recently proposed MobileNetV4 under similar latency constraints. Additionally, our method shows significant improvements in downstr eam tasks, including COCO object detection, instance segmentation, and ADE20k se mantic segmentation, while still maintaining low latency on mobile devices for h igh-resolution inputs in these scenarios. Code and models are available at: http s://github.com/ChuanyangZheng/iFormer.

****
Sungyoon Kim,Aaron Mishkin,Mert Pilanci
Exploring The Loss Landscape Of Regularized Neural Networks Via Convex Duality
We discuss several aspects of the loss landscape of regularized neural networks: the structure of stationary points, connectivity of optimal solutions, path with non-increasing loss to arbitrary global optimum, and the nonuniqueness of optimal solutions, by casting the problem into an equivalent convex problem and considering its dual. Starting from two-layer neural networks with scalar output, we first characterize the solution set of the convex problem using its dual and further characterize all stationary points. With the characterization, we show that the topology of the global optima goes through a phase transition as the width of the network changes, and construct counterexamples where the problem may have a continuum of optimal solutions. Finally, we show that the solution set characterization and connectivity results can be extended to different architectures, including two layer vector-valued neural networks and parallel three-layer neural networks.
****
Byungjai Kim,Chanho Ahn,Wissam J. Baddar,Kikyung Kim,HUIJIN LEE,Saehyun Ahn,Seungju Han,Sungjoo Suh,Eunho Yang
Test-Time Ensemble via Linear Mode Connectivity: A Path to Better Adaptation
Test-time adaptation updates pretrained models on the fly to handle distribution shifts in test data. While existing research has focused on stable optimization during adaptation, less attention has been given to enhancing model representations for adaptation capability. To address this gap, we propose Test-Time Ensemble (TTE) grounded in the intriguing property of linear mode connectivity. TTE leverages ensemble strategies during adaptation: 1) adaptively averaging the parameter weights of assorted test-time adapted models and 2) incorporating dropout to further promote representation diversity. These strategies encapsulate model diversity into a single model, avoiding computational burden associated with managing multiple models. Besides, we propose a robust knowledge distillation scheme to prevent model collapse, ensuring stable optimization and preserving the ensemble benefits during adaptation. Notably, TTE integrates seamlessly with existing TTA approaches, advancing their adaptation capabilities. In extensive experiments, integration with TTE consistently outperformed baseline models across various challenging scenarios, demonstrating its effectiveness and general applicability.
****
Xinchen Zhang,Ling Yang,Guohao Li,YaQi Cai,xie jiake,Yong Tang,Yujiu Yang,Mengdi Wang,Bin CUI
IterComp: Iterative Composition-Aware Feedback Learning from Model Gallery for Text-to-Image Generation
Advanced diffusion models like Stable Diffusion 3, Omost, and FLUX have made notable strides in compositional text-to-image generation. However, these methods typically exhibit distinct strengths for compositional generation, with some excelling in handling attribute binding and others in spatial relationships. This disparity highlights the need for an approach that can leverage the complementary strengths of various models to comprehensively improve the composition capability. To this end, we introduce IterComp, a novel framework that aggregates composition-aware model preferences from multiple models and employs an iterative feedback learning approach to enhance compositional generation. Specifically, we curate a gallery of six powerful open-source diffusion models and evaluate their three key compositional metrics: attribute binding, spatial relationships, and non-spatial relationships. Based on these metrics, we develop a composition-aware model preference dataset comprising numerous image-rank pairs to train composition-aware reward models. Then, we propose an iterative feedback learning method to enhance compositionality in a closed-loop manner, enabling the progressive self-refinement of both the base diffusion model and reward models over multiple iterations. Detailed theoretical proof demonstrates the effectiveness of this method. Extensive experiments demonstrate our significant superiority over previous methods, particularly in multi-category object composition and complex semantic al

ignment. IterComp opens new research avenues in reward feedback learning for dif
fusion models and compositional generation. Code: https://github.com/YangLing081
8/IterComp
****
David Xing Wu,Anant Sahai
Provable weak-to-strong generalization via benign overfitting
The classic teacher-student model in machine learning posits that a strong teach
er supervises a weak student to improve the student's capabilities.
    We instead consider the inverted situation, where a weak teacher supervises
a strong student with imperfect pseudolabels.
    This paradigm was recently brought forth by \citet{burns2023weak} and termed
 \emph{weak-to-strong generalization}.
    We theoretically investigate weak-to-strong generalization for binary and mu
ltilabel classification in a stylized overparameterized spiked covariance model
with Gaussian covariates where the weak teacher's pseudolabels are asymptoticall
y like random guessing.
    Under these assumptions, we provably identify two asymptotic phases of the s
trong student's generalization after weak supervision: (1) successful generaliza
tion and (2) random guessing.
    Our techniques should eventually extend to weak-to-strong multiclass classif
ication.
    Towards doing so, we prove a tight lower tail inequality for the maximum of
correlated Gaussians, which may be of independent interest.
    Understanding the multilabel setting reinforces the value of using logits fo
r weak supervision when they are available.
****
Yonatan Sverdlov,Ido Springer,Nadav Dym
REVISITING MULTI-PERMUTATION EQUIVARIANCE THROUGH THE LENS OF IRREDUCIBLE REPRES
ENTATIONS
This paper explores the characterization of equivariant linear layers for repres
entations of permutations and related groups. Unlike traditional approaches,
which address these problems using parameter-sharing, we consider an alternative
methodology based on irreducible representations and Schur’s lemma. Using this
methodology, we obtain an alternative derivation for existing models like DeepSe
ts,
2-IGN graph equivariant networks, and Deep Weight Space (DWS) networks. The
derivation for DWS networks is significantly simpler than that of previous resul
ts.
Next, we extend our approach to unaligned symmetric sets, where equivariance
to the wreath product of groups is required. Previous works have addressed this
problem in a rather restrictive setting, in which almost all wreath equivariant
layers
are Siamese. In contrast, we give a full characterization of layers in this case
 and
show that there is a vast number of additional non-Siamese layers in some settin
gs.
We also show empirically that these additional non-Siamese layers can improve
performance in tasks like graph anomaly detection, weight space alignment, and
learning Wasserstein distances.
****
Jiaming Zhuo,Yuwei Liu,Yintong Lu,Ziyi Ma,Kun Fu,Chuan Wang,Yuanfang Guo,Zhen Wa
ng,Xiaochun Cao,Liang Yang
DUALFormer: Dual Graph Transformer
Graph Transformers (GTs), adept at capturing the locality and globality of graph
s, have shown promising potential in node classification tasks. Most state-of-th
e-art GTs succeed through integrating local Graph Neural Networks (GNNs) with th
eir global Self-Attention (SA) modules to enhance structural awareness. Nonethel
ess, this architecture faces limitations arising from scalability challenges and
 the trade-off between capturing local and global information. On the one hand,
the quadratic complexity associated with the SA modules poses a significant chal

lenge for many GTs, particularly when scaling them to large-scale graphs. Numero
us GTs necessitated a compromise, relinquishing certain aspects of their express
ivity to garner computational efficiency. On the other hand, GTs face challenges
 in maintaining detailed local structural information while capturing long-range
 dependencies. As a result, they typically require significant computational cos
ts to balance the local and global expressivity. To address these limitations, t
his paper introduces a novel GT architecture, dubbed DUALFormer, featuring a dua
l-dimensional design of its GNN and SA modules. Leveraging approximation theory
from Linearized Transformers and treating the query as the surrogate representat
ion of node features, DUALFormer \emph{efficiently} performs the computationally
 intensive global SA module on feature dimensions. Furthermore, by such a separa
tion of local and global modules into dual dimensions, DUALFormer achieves a nat
ural balance between local and global expressivity. In theory, DUALFormer can re
duce intra-class variance, thereby enhancing the discriminability of node repres
entations. Extensive experiments on eleven real-world datasets demonstrate its e
ffectiveness and efficiency over existing state-of-the-art GTs.
****

Katie Matton,Robert Ness,John Guttag,Emre Kiciman

Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations

Large language models (LLMs) are capable of generating *plausible* explanations
of how they arrived at an answer to a question. However, these explanations can
misrepresent the model's "reasoning" process, i.e., they can be *unfaithful*. Th
is, in turn, can lead to over-trust and misuse. We introduce a new approach for
measuring the faithfulness of LLM explanations. First, we provide a rigorous def
inition of faithfulness. Since LLM explanations mimic human explanations, they o
ften reference high-level *concepts* in the input question that purportedly infl
uenced the model. We define faithfulness in terms of the difference between the
set of concepts that the LLM's *explanations imply* are influential and the set
that *truly* are. Second, we present a novel method for estimating faithfulness
that is based on: (1) using an auxiliary LLM to modify the values of concepts wi
thin model inputs to create realistic counterfactuals, and (2) using a hierarchi
cal Bayesian model to quantify the causal effects of concepts at both the exampl
e- and dataset-level. Our experiments show that our method can be used to quanti
fy and discover interpretable patterns of unfaithfulness. On a social bias task,
 we uncover cases where LLM explanations hide the influence of social bias. On a
 medical question answering task, we uncover cases where LLM explanations provid
e misleading claims about which pieces of evidence influenced the model's decisi
ons.
****

Yalong Jiang

Local Patterns Generalize Better for Novel Anomalies

Video anomaly detection (VAD) aims to identify novel actions or events which are
 unseen during training. Existing mainstream VAD techniques typically focus on t
he global patterns with redundant details and struggle to generalize to unseen s
amples. In this paper, we propose a framework that identifies the local patterns
 which generalize to novel samples and models the dynamics of local patterns. Th
e capability of extracting spatial local patterns is achieved through a two-stag
e process involving image-text alignment and cross-modality attention. Generaliz
able representations are built by focusing on semantically relevant components w
hich can be recombined to capture the essence of novel anomalies, reducing unnec
essary visual data variances. To enhance local patterns with temporal clues, we
propose a State Machine Module (SMM) that utilizes earlier high-resolution textu
al tokens to guide the generation of precise captions for subsequent low-resolut
ion observations. Furthermore, temporal motion estimation complements spatial lo
cal patterns to detect anomalies characterized by novel spatial distributions or
 distinctive dynamics. Extensive experiments on popular benchmark datasets demon
strate the achievement of state-of-the-art performance. Code is available at htt
ps://github.com/AllenYLJiang/Local-Patterns-Generalize-Better/.
****

Jinbiao Chen,Jiahai Wang,Zhiguang Cao,Yaoxin Wu

Neural Multi-Objective Combinatorial Optimization via Graph-Image Multimodal Fusion

Existing neural multi-objective combinatorial optimization (MOCO) methods still exhibit an optimality gap since they fail to fully exploit the intrinsic features of problem instances. A significant factor contributing to this shortfall is their reliance solely on graph-modal information. To overcome this, we propose a novel graph-image multimodal fusion (GIMF) framework that enhances neural MOCO methods by integrating graph and image information of the problem instances. Our GIMF framework comprises three key components: (1) a constructed coordinate image to better represent the spatial structure of the problem instance, (2) a problem-size adaptive resolution strategy during the image construction process to improve the cross-size generalization of the model, and (3) a multimodal fusion mechanism with modality-specific bottlenecks to efficiently couple graph and image information. We demonstrate the versatility of our GIMF by implementing it with two state-of-the-art neural MOCO backbones. Experimental results on classic MOCO problems show that our GIMF significantly outperforms state-of-the-art neural MOCO methods and exhibits superior generalization capability.

****

Harish Babu Manogaran,M. Maruf,Arka Daw,Kazi Sajeed Mehrab,Caleb Patrick Charpentier,Josef Uyeda,Wasila M Dahdul,Matthew J Thompson,Elizabeth G Campolongo,Kaiya L Provost,Wei-Lun Chao,Tanya Berger-Wolf,Paula Mabee,Hilmar Lapp,Anuj Karpatne

What Do You See in Common? Learning Hierarchical Prototypes over Tree-of-Life to Discover Evolutionary Traits

A grand challenge in biology is to discover evolutionary traits---features of organisms common to a group of species with a shared ancestor in the tree of life (also referred to as phylogenetic tree). With the growing availability of image repositories in biology, there is a tremendous opportunity to discover evolutionary traits directly from images in the form of a hierarchy of prototypes. However, current prototype-based methods are mostly designed to operate over a flat structure of classes and face several challenges in discovering hierarchical prototypes, including the issue of learning over-specific prototypes at internal nodes. To overcome these challenges, we introduce the framework of Hierarchy aligned Commonality through Prototypical Networks (HComP-Net). The key novelties in HComP-Net include a novel over-specificity loss to avoid learning over-specific prototypes, a novel discriminative loss to ensure prototypes at an internal node are absent in the contrasting set of species with different ancestry, and a novel masking module to allow for the exclusion of over-specific prototypes at higher levels of the tree without hampering classification performance. We empirically show that HComP-Net learns prototypes that are accurate, semantically consistent, and generalizable to unseen species in comparison to baselines. Our code is publicly accessible at Imageomics Institute Github site: https://github.com/Imageomics/HComPNet.

****

Zhiwei Xu,Zhiyu Ni,Yixin Wang,Wei Hu

Let Me Grok for You: Accelerating Grokking via Embedding Transfer from a Weaker Model

''Grokking'' is a phenomenon where a neural network first memorizes training data and generalizes poorly, but then suddenly transitions to near-perfect generalization after prolonged training. While intriguing, this delayed generalization phenomenon compromises predictability and efficiency. Ideally, models should generalize directly without delay. To this end, this paper proposes GrokTransfer, a simple and principled method for accelerating grokking in training neural networks, based on the key observation that data embedding plays a crucial role in determining whether generalization is delayed. GrokTransfer first trains a smaller, weaker model to reach a nontrivial (but far from optimal) test performance. Then, the learned input embedding from this weaker model is extracted and used to initialize the embedding in the target, stronger model. We rigorously prove that, on a synthetic XOR task where delayed generalization always occurs in normal training, GrokTransfer enables the target model to generalize directly without delay. Moreover, we demonstrate that, across empirical studies o

f different tasks, GrokTransfer effectively reshapes the training dynamics and e
liminates delayed generalization, for both fully-connected neural networks and T
ransformers.
****

Ziwei Huang,Jianan Zhou,Zhiguang Cao,Yixin XU
Rethinking Light Decoder-based Solvers for Vehicle Routing Problems
Light decoder-based solvers have gained popularity for solving vehicle routing p
roblems (VRPs) due to their efficiency and ease of integration with reinforcemen
t learning algorithms. However, they often struggle with generalization to large
r problem instances or different VRP variants. This paper revisits light decoder
-based approaches, analyzing the implications of their reliance on static embedd
ings and the inherent challenges that arise. Specifically, we demonstrate that
in the light decoder paradigm, the encoder is implicitly tasked with capturing i
nformation for all potential decision scenarios during solution construction wit
hin a single set of embeddings, resulting in high information density. Furtherm
ore, our empirical analysis reveals that the overly simplistic decoder struggles
to effectively utilize this dense information, particularly as task complexity
increases, which limits generalization to out-of-distribution (OOD) settings. B
uilding on these insights, we show that enhancing the decoder capacity, with a s
imple addition of identity mapping and a feed-forward layer, can considerably al
leviate the generalization issue. Experimentally, our method significantly enhan
ces the OOD generalization of light decoder-based approaches on large-scale inst
ances and complex VRP variants, narrowing the gap with the heavy decoder paradig
m. Our code is available at: https://github.com/ziweileonhuang/reld-nco.
****

Yuchen Sun,Kejun Huang
Global Identifiability of Overcomplete Dictionary Learning via L1 and Volume Min
imization
We propose a novel formulation for dictionary learning with an overcomplete dict
ionary, i.e., when the number of atoms is larger than the dimension of the dicti
onary. The proposed formulation consists of a weighted sum of ℓ_1 norms of
the rows of the sparse coefficient matrix plus the log of the matrix volume of t
he dictionary matrix. The main contribution of this work is to show that this no
vel formulation guarantees global identifiability of the overcomplete dictionary
, under a mild condition that the sparse coefficient matrix satisfies a strong s
cattering condition in the hypercube. Furthermore, if every column of the coeffi
cient matrix is sparse and the dictionary guarantees ℓ_1 recovery, then the
coefficient matrix is identifiable as well. This is a major breakthrough for no
t only dictionary learning but also general matrix factorization models as ident
ifiability is guaranteed even when the latent dimension is higher than the ambie
nt dimension. We also provide a probabilistic analysis and show that if the spar
se coefficient matrix is generated from the widely adopted sparse-Gaussian model
, then the $m\times k$ overcomplete dictionary is globally identifiable if the s
ample size is bigger than a constant times $(k^2/m)\log(k^2/m)$ with overwhelmin
g probability. Finally, we propose an algorithm based on alternating minimizatio
n to solve the new proposed formulation.
****

Zicheng Zhang,Haoning Wu,Chunyi Li,Yingjie Zhou,Wei Sun,Xiongkuo Min,Zijian Chen
,Xiaohong Liu,Weisi Lin,Guangtao Zhai
A-Bench: Are LMMs Masters at Evaluating AI-generated Images?
How to accurately and efficiently assess AI-generated images (AIGIs) remains a c
ritical challenge for generative models. Given the high costs and extensive time
commitments required for user studies, many researchers have turned towards emp
loying large multi-modal models (LMMs) as AIGI evaluators, the precision and val
idity of which are still questionable. Furthermore, traditional benchmarks often
utilize mostly natural-captured content rather than AIGIs to test the abilities
of LMMs, leading to a noticeable gap for AIGIs. Therefore, we introduce **A-Ben
ch** in this paper, a benchmark designed to diagnose *whether LMMs are masters a
t evaluating AIGIs*. Specifically, **A-Bench** is organized under two key princi
ples: 1) Emphasizing both high-level semantic understanding and low-level visual

quality perception to address the intricate demands of AIGIs. 2) Various genera
tive models are utilized for AIGI creation, and various LMMs are employed for ev
aluation, which ensures a comprehensive validation scope. Ultimately, 2,864 AIGI
s from 16 text-to-image models are sampled, each paired with question-answers an
notated by human experts. We hope that **A-Bench** will significantly enhance th
e evaluation process and promote the generation quality for AIGIs.
****

Pinzheng Wang,Zecheng Tang,Keyan Zhou,Juntao Li,Qiaoming Zhu,Min Zhang
Revealing and Mitigating Over-Attention in Knowledge Editing
Large Language Models~(LLMs) have demonstrated superior performance across a wid
e range of tasks, but they still exhibit undesirable errors due to incorrect kno
wledge learned from the training data. To avoid this, knowledge editing methods
emerged to precisely edit the specific model knowledge via efficiently modifying
a very small percentage of parameters. However, those methods can lead to the p
roblem of **Specificity Failure**, where the existing knowledge and capabilities
are severely degraded due to editing.
Our preliminary indicates that Specificity Failure primarily stems from the mode
l's attention heads assigning excessive attention scores to entities related to
the edited knowledge, thereby unduly focusing on specific snippets within the co
ntext, which we denote as the **Attention Drift** phenomenon.
To mitigate such Attention Drift issue, we introduce a simple yet effective meth
od **S**elective **A**ttention **D**rift **R**estriction(**SADR**), which introd
uces an additional regularization term during the knowledge editing process to r
estrict changes in the attention weight distribution, thereby preventing undue f
ocus on the edited entity.
Experiments on five frequently-used strong LLMs demonstrate the effectiveness of
our method, where SADR can significantly mitigate Specificity Failure in the pr
edominant knowledge editing tasks.
****

Milong Ren,ZaiKai He,Haicang Zhang
Multi-objective antibody design with constrained preference optimization
Antibody design is crucial for developing therapies against diseases such as can
cer and viral infections. Recent deep generative models have significantly advan
ced computational antibody design, particularly in enhancing binding affinity to
target antigens. However, beyond binding affinity, antibodies should exhibit ot
her favorable biophysical properties such as non-antigen binding specificity and
low self-association, which are important for antibody developability and clini
cal safety. To address this challenge, we propose AbNovo, a framework that lever
ages constrained preference optimization for multi-objective antibody design. Fi
rst, we pre-train an antigen-conditioned generative model for antibody structure
and sequence co-design. Then, we fine-tune the model using binding affinity as
a reward while enforcing explicit constraints on other biophysical properties. S
pecifically, we model the physical binding energy with continuous rewards rather
than pairwise preferences and explore a primal-and-dual approach for constraine
d optimization. Additionally, we incorporate a structure-aware protein language
model to mitigate the issue of limited training data. Evaluated on independent t
est sets, AbNovo outperforms existing methods in metrics of binding affinity suc
h as Rosetta binding energy and evolutionary plausibility, as well as in metrics
for other biophysical properties like stability and specificity.
****

Bhavya Vasudeva,Deqing Fu,Tianyi Zhou,Elliott Kau,Youqi Huang,Vatsal Sharan
Transformers Learn Low Sensitivity Functions: Investigations and Implications
Transformers achieve state-of-the-art accuracy and robustness across many tasks,
but an understanding of their inductive biases and how those biases differ from
other neural network architectures remains elusive. In this work, we identify t
he sensitivity of the model to token-wise random perturbations in the input as a
unified metric which explains the inductive bias of transformers across differe
nt data modalities and distinguishes them from other architectures. We show that
transformers have lower sensitivity than MLPs, CNNs, ConvMixers and LSTMs, acr
oss both vision and language tasks. We also show that this low-sensitivity bias

has important implications: i) lower sensitivity correlates with improved robust
ness; it can also be used as an efficient intervention to further improve the ro
bustness of transformers; ii) it corresponds to flatter minima in the loss lands
cape; and iii) it can serve as a progress measure for grokking. We support these
 findings with theoretical results showing (weak) spectral bias of transformers
in the NTK regime, and improved robustness due to the lower sensitivity.
****
Xiandong Zou,Mingzhu Shen,Christos-Savvas Bouganis,Yiren Zhao
Cached Multi-Lora Composition for Multi-Concept Image Generation
Low-Rank Adaptation (LoRA) has emerged as a widely adopted technique in text-to-
image models, enabling precise rendering of multiple distinct elements, such as
characters and styles, in multi-concept image generation. However, current appro
aches face significant challenges when composing these LoRAs for multi-concept i
mage generation, particularly as the number of LoRAs increases, resulting in dim
inished generated image quality.
In this paper, we initially investigate the role of LoRAs in the denoising proce
ss through the lens of the Fourier frequency domain.
Based on the hypothesis that applying multiple LoRAs could lead to "semantic con
flicts", we have conducted empirical experiments and find that certain LoRAs amp
lify high-frequency features such as edges and textures, whereas others mainly f
ocus on low-frequency elements, including the overall structure and smooth color
 gradients.
Building on these insights, we devise a frequency domain based sequencing strate
gy to determine the optimal order in which LoRAs should be integrated during inf
erence. This strategy offers a methodical and generalizable solution compared to
 the naive integration commonly found in existing LoRA fusion techniques.
To fully leverage our proposed LoRA order sequence determination method in multi
-LoRA composition tasks, we introduce a novel, training-free framework, Cached M
ulti-LoRA (CMLoRA), designed to efficiently integrate multiple LoRAs while maint
aining cohesive image generation.
With its flexible backbone for multi-LoRA fusion and a non-uniform caching strat
egy tailored to individual LoRAs, CMLoRA has the potential to reduce semantic co
nflicts in LoRA composition and improve computational efficiency.
Our experimental evaluations demonstrate that CMLoRA outperforms state-of-the-ar
t training-free LoRA fusion methods by a significant margin -- it achieves an av
erage improvement of 2.19% in CLIPScore, and $11.25%$% in MLLM win rate compar
ed to LoraHub, LoRA Composite, and LoRA Switch.
****
Nate Gillman,Daksh Aggarwal,Michael Freeman,Chen Sun
Fourier Head: Helping Large Language Models Learn Complex Probability Distributi
ons
As the quality of large language models has improved, there has been increased i
nterest in using them to model non-linguistic tokens. For example, the Decision
Transformer recasts agentic decision making as a sequence modeling problem, usin
g a decoder-only LLM to model the distribution over the discrete action space fo
r an Atari agent. However, when adapting LLMs to non-linguistic domains, it rema
ins unclear if softmax over discrete bins captures the continuous structure of t
he tokens and the potentially complex distributions needed for high quality toke
n generation. We introduce a neural network layer, constructed using Fourier ser
ies, which we can easily substitute for any linear layer if we want the outputs
to have a more continuous structure. We perform extensive analysis on synthetic
datasets, as well as on large-scale decision making and time series forecasting
tasks. We also provide theoretical evidence that this layer can better learn sig
nal from data while ignoring high-frequency noise. All of our results support th
e effectiveness of our proposed Fourier head in scenarios where the underlying d
ata distribution has a natural continuous structure. For example, the Fourier he
ad improves a Decision Transformer agent's returns across four benchmark Atari g
ames by as much as 377\%, and increases a state-of-the-art times series foundati
on model's forecasting performance by 3.5\% across 20 benchmarks unseen during t
raining.

We release our implementation at https://nategillman.com/fourier-head
****
Micha■ Bortkiewicz,W■adys■aw Pa■ucki,Vivek Myers,Tadeusz Dziarmaga,Tomasz Arczewski,■ukasz Kuci■ski,Benjamin Eysenbach
Accelerating Goal-Conditioned Reinforcement Learning Algorithms and Research
Self-supervision has the potential to transform reinforcement learning (RL), paralleling the breakthroughs it has enabled in other areas of machine learning. While self-supervised learning in other domains aims to find patterns in a fixed dataset, self-supervised goal-conditioned reinforcement learning (GCRL) agents discover *new* behaviors by learning from the goals achieved during unstructured interaction with the environment. However, these methods have failed to see similar success, both due to a lack of data from slow environment simulations as well as a lack of stable algorithms. We take a step toward addressing both of these issues by releasing a high-performance codebase and benchmark (`JaxGCRL`) for self-supervised GCRL, enabling researchers to train agents for millions of environment steps in minutes on a single GPU. By utilizing GPU-accelerated replay buffers, environments, and a stable contrastive RL algorithm, we reduce training time by up to $22\times$. Additionally, we assess key design choices in contrastive RL, identifying those that most effectively stabilize and enhance training performance. With this approach, we provide a foundation for future research in self-supervised GCRL, enabling researchers to quickly iterate on new ideas and evaluate them in diverse and challenging environments. Code: https://anonymous.4open.science/r/JaxGCRL-2316/README.md
****
Junjie Yin,Alexander M Rush
Compute-Constrained Data Selection
Data selection can reduce the amount of training data needed to finetune LLMs; however, the efficacy of data selection scales directly with its compute. Motivated by the practical challenge of compute-constrained finetuning, we consider the setting in which both the cost of selecting data and training are budgeted for. We first formalize the problem of data selection with a cost-aware utility function,
and model the data selection problem as trading off initial-selection cost for training gain. We run a comprehensive sweep of experiments across multiple tasks, varying compute budget by scaling finetuning tokens, model sizes, and data selection compute. Interestingly we find that many powerful data selection methods are almost never compute-optimal, and that cheaper data selection alternatives dominate both from a theoretical and empirical perspective. For compute-optimal training, we find that perplexity and gradient data selection require training-to-selection model size ratios of 5x and 10x, respectively.
****
Yuchen Shen,Chenhao Zhang,Sijie Fu,Chenghui Zhou,Newell Washburn,Barnabas Poczos
Chemistry-Inspired Diffusion with Non-Differentiable Guidance
Recent advances in diffusion models have shown remarkable potential in the conditional generation of novel molecules. These models can be guided in two ways: (i) explicitly, through additional features representing the condition, or (ii) implicitly, using a property predictor. However, training property predictors or conditional diffusion models requires an abundance of labeled data and is inherently challenging in real-world applications. We propose a novel approach that attenuates the limitations of acquiring large labeled datasets by leveraging domain knowledge from quantum chemistry as a non-differentiable oracle to guide an unconditional diffusion model. Instead of relying on neural networks, the oracle provides accurate guidance in the form of estimated gradients, allowing the diffusion process to sample from a conditional distribution specified by quantum chemistry. We show that this results in more precise conditional generation of novel and stable molecular structures. Our experiments demonstrate that our method: (1) significantly reduces atomic forces, enhancing the validity of generated molecules when used for stability optimization; (2) is compatible with both explicit and implicit guidance in diffusion models, enabling joint optimization of molecu

lar properties and stability; and (3) generalizes effectively to molecular optim
ization tasks beyond stability optimization. Our implementation is available at
https://github.com/A-Chicharito-S/ChemGuide.
****
David Lüdke,Enric Rabasseda Raventós,Marcel Kollovieh,Stephan Günnemann
Unlocking Point Processes through Point Set Diffusion
Point processes model the distribution of random point sets in mathematical spac
es, such as spatial and temporal domains, with applications in fields like seism
ology, neuroscience, and economics.
Existing statistical and machine learning models for point processes are predomi
nantly constrained by their reliance on the characteristic intensity function, i
ntroducing an inherent trade-off between efficiency and flexibility.
In this paper, we introduce Point Set Diffusion, a diffusion-based latent variab
le model that can represent arbitrary point processes on general metric spaces w
ithout relying on the intensity function.
By directly learning to stochastically interpolate between noise and data point
sets, our approach effectively captures the distribution of point processes and
enables efficient, parallel sampling and flexible generation for complex conditi
onal tasks.
Experiments on synthetic and real-world datasets demonstrate that Point Set Diff
usion achieves state-of-the-art performance in unconditional and conditional gen
eration of spatial and spatiotemporal point processes while providing up to orde
rs of magnitude faster sampling.
****
Ryan D'Orazio,Danilo Vucetic,Zichu Liu,Junhyung Lyle Kim,Ioannis Mitliagkas,Gaut
hier Gidel
Solving hidden monotone variational inequalities with surrogate losses
Deep learning has proven to be effective in a wide variety of loss minimization
problems.
However, many applications of interest, like minimizing projected Bellman error
and min-max optimization, cannot be modelled as minimizing a scalar loss functio
n but instead correspond to solving a variational inequality (VI) problem.
This difference in setting has caused many practical challenges as naive gradien
t-based approaches from supervised learning tend to diverge and cycle in the VI
case.
In this work, we propose a principled surrogate-based approach compatible with d
eep learning to solve VIs.
We show that our surrogate-based approach has three main benefits: (1) under ass
umptions that are realistic in practice (when hidden monotone structure is prese
nt, interpolation, and sufficient optimization of the surrogates), it guarantees
 convergence, (2) it provides a unifying perspective of existing methods, and (3
) is amenable to existing deep learning optimizers like ADAM.
Experimentally, we demonstrate our surrogate-based approach is effective in min-
max optimization and minimizing projected Bellman error. Furthermore, in the dee
p reinforcement learning case, we propose a novel variant of TD(0) which is more
 compute and sample efficient.
****
Hatef Otroshi Shahreza,Sébastien Marcel
HyperFace: Generating Synthetic Face Recognition Datasets by Exploring Face Embe
dding Hypersphere
Face recognition datasets are often collected by crawling Internet and without i
ndividuals' consents, raising ethical and privacy concerns. Generating syntheti
c datasets for training face recognition models has emerged as a promising alter
native. However, the generation of synthetic datasets remains challenging as it
 entails adequate inter-class and intra-class variations. While advances in gene
rative models have made it easier to increase intra-class variations in face dat
asets (such as pose, illumination, etc.), generating sufficient inter-class vari
ation is still a difficult task. In this paper, we formulate the dataset generat
ion as a packing problem on the embedding space (represented on a hypersphere) o
f a face recognition model and propose a new synthetic dataset generation approa

ch, called HyperFace. We formalize our packing problem as an optimization proble
m and solve it with a gradient descent-based approach. Then, we use a conditiona
l face generator model to synthesize face images from the optimized embeddings.
We use our generated datasets to train face recognition models and evaluate the
trained models on several benchmarking real datasets. Our experimental results s
how that models trained with HyperFace achieve state-of-the-art performance in t
raining face recognition using synthetic datasets. Project page: https://www.idi
ap.ch/paper/hyperface
****

Julian D Parker,Anton Smirnov,Jordi Pons,CJ Carr,Zack Zukowski,Zach Evans,Xubo L
iu

Scaling Transformers for Low-Bitrate High-Quality Speech Coding

The tokenization of audio with neural audio codec models is a vital part of mode
rn AI pipelines for the generation or understanding of speech, alone or in a mul
timodal context. Traditionally such tokenization models have concentrated on low
 parameter-count architectures using only components with strong inductive biase
s. In this work we show that by applying a transformer architecture with large p
arameter count to this problem, and applying a flexible Finite Scalar Quantizati
on (FSQ) based bottleneck, it is possible to reach state-of-the-art speech quali
ty at extremely low bit-rates of 400 or 700 bits-per-second. The trained mod
els strongly out-perform existing baselines in both objective and subjective tes
ts.
****

Baran Hashemi,Roderic Guigo Corominas,Alessandro Giacchetto

Can Transformers Do Enumerative Geometry?

We introduce a Transformer-based approach to computational enumerative geometry,
 specifically targeting the computation of ψ-class intersection numbers on
the moduli space of curves. Traditional methods for calculating these numbers su
ffer from factorial computational complexity, making them impractical to use. By
 reformulating the problem as a continuous optimization task, we compute interse
ction numbers across a wide value range from 10^{-45} to 10^{45}. To capture
 the recursive nature inherent in these intersection numbers, we propose the Dyn
amic Range Activator (DRA), a new activation function that enhances the Transfor
mer's ability to model recursive patterns and handle severe heteroscedasticity.
Given precision requirements for computing the intersections, we quantify the un
certainty of the predictions using Conformal Prediction with a dynamic sliding w
indow adaptive to the partitions of equivalent number of marked points. To the b
est of our knowledge, there has been no prior work on modeling recursive functio
ns with such a high-variance and factorial growth. Beyond simply computing inter
section numbers, we explore the enumerative "world-model" of Transformers. Our i
nterpretability analysis reveals that the network is implicitly modeling the Vir
asoro constraints in a purely data-driven manner. Moreover, through abductive hy
pothesis testing, probing, and causal inference, we uncover evidence of an emerg
ent internal representation of the the large-genus asymptotic of ψ-class in
tersection numbers. These findings suggest that the network internalizes the par
ameters of the asymptotic closed-form and the polynomiality phenomenon of ψ
-class intersection numbers in a non-linear manner.
****

Marius Memmel,Jacob Berg,Bingqing Chen,Abhishek Gupta,Jonathan Francis

STRAP: Robot Sub-Trajectory Retrieval for Augmented Policy Learning

Robot learning is witnessing a significant increase in the size, diversity, and
complexity of pre-collected datasets, mirroring trends in domains such as natura
l language processing and computer vision. Many robot learning methods treat suc
h datasets as multi-task expert data and learn a multi-task, generalist policy b
y training broadly across them. Notably, while these generalist policies can imp
rove the average performance across many tasks, the performance of generalist po
licies on any one task is often suboptimal due to negative transfer between part
itions of the data, compared to task-specific specialist policies. In this work,
 we argue for the paradigm of training policies during deployment given the scen
arios they encounter: rather than deploying pre-trained policies to unseen probl

ems in a zero-shot manner, we non-parametrically retrieve and train models direc
tly on relevant data at test time. Furthermore, we show that many robotics tasks
 share considerable amounts of low-level behaviors and that retrieval at the "su
b"-trajectory granularity enables significantly improved data utilization, gener
alization, and robustness in adapting policies to novel problems. In contrast, e
xisting full-trajectory retrieval methods tend to underutilize the data and miss
 out on shared cross-task content. This work proposes STRAP, a technique for lev
eraging pre-trained vision foundation models and dynamic time warping to retriev
e sub-sequences of trajectories from large training corpora in a robust fashion.
 STRAP outperforms both prior retrieval algorithms and multi-task learning metho
ds in simulated and real experiments, showing the ability to scale to much large
r offline datasets in the real world as well as the ability to learn robust cont
rol policies with just a handful of real-world demonstrations.
****
Alex Kipnis,Konstantinos Voudouris,Luca M. Schulze Buschoff,Eric Schulz
metabench - A Sparse Benchmark of Reasoning and Knowledge in Large Language Mode
ls
Large Language Models (LLMs) vary in their abilities on a range of tasks. Initia
tives such as the Open LLM Leaderboard aim to quantify these differences with se
veral large benchmarks (sets of test items to which an LLM can respond either co
rrectly or incorrectly).
   However, high correlations within and between benchmark scores suggest that (
1) there exists a small set of common underlying abilities that these benchmarks
 measure, and (2) items tap into redundant information and the benchmarks may th
us be considerably compressed.
   We use data from n > 5000 LLMs to identify the most informative items of six
benchmarks, ARC, GSM8K, HellaSwag, MMLU, TruthfulQA and WinoGrande (with d = 28,
632 items in total). From them we distill a sparse benchmark, metabench, that ha
s less than 3% of the original size of all six benchmarks combined. This new spa
rse benchmark goes beyond point scores by yielding estimators of the underlying
benchmark-specific abilities.
   We show that these estimators (1) can be used to reconstruct each original in
dividual benchmark score with, on average, 1.24% root mean square error (RMSE),
(2) reconstruct the original total score with 0.58% RMSE, and (3) have a single
underlying common factor whose Spearman correlation with the total score is r =
0.94.
****
Zihan Pengmei,Zhengyuan Shen,Zichen Wang,Marcus D. Collins,Huzefa Rangwala
Pushing the Limits of All-Atom Geometric Graph Neural Networks: Pre-Training, Sc
aling, and Zero-Shot Transfer
The ability to construct transferable descriptors for molecular and biological s
ystems has broad applications in drug discovery, molecular dynamics, and protein
 analysis. Geometric graph neural networks (Geom-GNNs) utilizing all-atom inform
ation have revolutionized atomistic simulations by enabling the prediction of in
teratomic potentials and molecular properties. Despite these advances, the appli
cation of all-atom Geom-GNNs in protein modeling remains limited due to computat
ional constraints. In this work, we first demonstrate the potential of pre-train
ed Geom-GNNs as zero-shot transfer learners, effectively modeling protein system
s with all-atom granularity. Through extensive experimentation to evaluate their
 expressive power, we characterize the scaling behaviors of Geom-GNNs across sel
f-supervised, supervised, and unsupervised setups. Interestingly, we find that G
eom-GNNs deviate from conventional power-law scaling observed in other domains,
with no predictable scaling principles for molecular representation learning. Fu
rthermore, we show how pre-trained graph embeddings can be directly used for ana
lysis and synergize with other architectures to enhance expressive power for pro
tein modeling.
****
Yue Hu,Yuzhu Cai,Yaxin Du,Xinyu Zhu,Xiangrui Liu,Zijie Yu,Yuchen Hou,Shuo Tang,S
iheng Chen
Self-Evolving Multi-Agent Collaboration Networks for Software Development

LLM-driven multi-agent collaboration (MAC) systems have demonstrated impressive capabilities in automatic software development at the function level. However, their heavy reliance on human design limits their adaptability to the diverse demands of real-world software development.
To address this limitation, we introduce EvoMAC, a novel self-evolving paradigm for MAC networks. Inspired by traditional neural network training, EvoMAC obtains text-based environmental feedback by verifying the MAC network's output against a target proxy and leverages a novel textual backpropagation to update the network.
To extend coding capabilities beyond function-level tasks to more challenging software-level development, we further propose RSD-Bench, a requirement-oriented software development benchmark, which features complex and diverse software requirements along with automatic evaluation of requirement correctness.
Our experiments show that:
i) The automatic requirement-aware evaluation in RSD-Bench closely aligns with human evaluations, validating its reliability as a software-level coding benchmark.
ii) EvoMAC outperforms previous SOTA methods on both the software-level RSD-Bench and the function-level HumanEval benchmarks, reflecting its superior coding capabilities.
****
Yifan Yang,Gang Chen,Hui Ma,Cong Zhang,Zhiguang Cao,Mengjie Zhang
Graph Assisted Offline-Online Deep Reinforcement Learning for Dynamic Workflow Scheduling
Dynamic workflow scheduling (DWS) in cloud computing presents substantial challenges due to heterogeneous machine configurations, unpredictable workflow arrivals/patterns, and constantly evolving environments. However, existing research often assumes homogeneous setups and static conditions, limiting flexibility and adaptability in real-world scenarios. In this paper, we propose a novel *Graph assisted Offline-Online Deep Reinforcement Learning* (GOODRL) approach to building an effective and efficient scheduling agent for DWS. Our approach features three key innovations: (1) a *task-specific* graph representation and a *Graph Attention Actor Network* that enable the agent to dynamically assign focused tasks to heterogeneous machines while explicitly considering the future impact of each machine on these tasks; (2) a *system-oriented* graph representation and a *Graph Attention Critic Network* that facilitate efficient processing of new information and understanding its impact on the current state, crucial for managing unpredictable workflow arrivals/patterns in real-time; and (3) an *offline-online* method that utilizes imitation learning for effective offline training and applies gradient control and decoupled high-frequency critic training techniques during online learning to sustain the agent’s robust performance in rapidly changing environments. Experimental results demonstrate that GOODRL significantly outperforms several state-of-the-art algorithms, achieving substantially lower mean flowtime and high adaptability in various online and offline scenarios.
****
Hyun Kyu Lee,Sung Whan Yoon
Flat Reward in Policy Parameter Space Implies Robust Reinforcement Learning
Investigating flat minima on loss surfaces in parameter space is well-documented in the supervised learning context, highlighting its advantages for model generalization. However, limited attention has been paid to the reinforcement learning (RL) context, where the impact of flatter reward landscapes in policy parameter space remains largely unexplored. Beyond merely extrapolating from supervised learning, which suggests a link between flat reward landscapes and enhanced generalization, we aim to formally connect the flatness of the reward surface to the robustness of RL models. In policy models where a deep neural network determines actions, flatter reward landscapes in response to parameter perturbations lead to consistent rewards even when actions are perturbed. Moreover, robustness to action perturbations further enhances robustness against other variations, such as changes in state transition probabilities and reward functions. We extensively simulate various RL environments, confirming the consistent benefits of flatte

r reward landscapes in enhancing the robustness of RL under diverse conditions, including action selection, transition dynamics, and reward functions. The code for these experiments is available at https://github.com/HK-05/flatreward-RRL.
****

Kaya Stechly,Karthik Valmeekam,Subbarao Kambhampati
On the self-verification limitations of large language models on reasoning and p lanning tasks
There has been considerable divergence of opinion on the reasoning abilities of Large Language Models (LLMs).
While the initial optimism that reasoning might emerge automatically with scale has been tempered thanks to a slew of counterexamples--ranging from multiplicati on to simple planning--there persists a wide spread belief that LLMs can self-cr itique and improve their own solutions in an iterative fashion.
This belief seemingly rests on the assumption that verification of correctness s hould be easier than generation--a rather classical argument from computational complexity--which should be irrelevant to LLMs to the extent that what they are doing is approximate retrieval.
In this paper, we set out to systematically investigate the effectiveness of ite rative prompting in the context of reasoning and planning.
We present a principled empirical study of the performance of GPT-4 in three dom ains: Game of 24, Graph Coloring, and STRIPS planning.
We experiment both with the model critiquing its own answers and with an externa l correct reasoner verifying proposed solutions.
In each case, we analyze whether the content of criticisms actually affects bott om line performance, and whether we can ablate elements of the augmented system without losing performance. We observe significant performance collapse with self-critique and significant performance gains with sound external verific ation.
We also note that merely re-prompting with a sound verifier maintains most of th e benefits of more involved setups.
****

Lile Cai,Chuan-Sheng Foo,Xun Xu,ZAIWANG GU,Jun Cheng,Xulei Yang
Evidential Learning-based Certainty Estimation for Robust Dense Feature Matching
Dense feature matching methods aim to estimate a dense correspondence field betw een images. Inaccurate correspondence can occur due to the presence of unmatchab le region, necessitating the need for certainty measurement. This is typically a ddressed by training a binary classifier to decide whether each predicted corres pondence is reliable. However, deep neural network-based classifiers can be vuln erable to image corruptions or perturbations, making it difficult to obtain reli able matching pairs in corrupted scenario. In this work, we propose an evidentia l deep learning framework to enhance the robustness of dense matching against co rruptions. We modify the certainty prediction branch in dense matching models to generate appropriate belief masses and compute the certainty score by taking ex pectation over the resulting Dirichlet distribution. We evaluate our method on a wide range of benchmarks and show that our method leads to improved robustness against common corruptions and adversarial attacks, achieving up to 10.1\% impro vement under severe corruptions.
****

Rui Hu,Yifan Zhang,Zhuoran Li,Longbo Huang
Beyond Squared Error: Exploring Loss Design for Enhanced Training of Generative Flow Networks
Generative Flow Networks (GFlowNets) are a novel class of generative models desi gned to sample from unnormalized distributions and have found applications in va rious important tasks, attracting great research interest in their training algo rithms. In general, GFlowNets are trained by fitting the forward flow to the bac kward flow on sampled training objects. Prior work focused on the choice of trai ning objects, parameterizations, sampling and resampling strategies, and backwar d policies, aiming to enhance credit assignment, exploration, or exploitation of the training process. However, the choice of regression loss, which can highly influence the exploration and exploitation behavior of the under-training policy

, has been overlooked. Due to the lack of theoretical understanding for choosing an appropriate regression loss, most existing algorithms train the flow network by minimizing the squared error of the forward and backward flows in log-space, i.e., using the quadratic regression loss. In this work, we rigorously prove that distinct regression losses correspond to specific divergence measures, enabling us to design and analyze regression losses according to the desired properties of the corresponding divergence measures. Specifically, we examine two key properties: zero-forcing and zero-avoiding, where the former promotes exploitation and higher rewards, and the latter encourages exploration and enhances diversity. Based on our theoretical framework, we propose three novel regression losses, namely, Shifted-Cosh, Linex(1/2), and Linex(1). We evaluate them across three benchmarks: hyper-grid, bit-sequence generation, and molecule generation. Our proposed losses are compatible with most existing training algorithms, and significantly improve the performances of the algorithms concerning convergence speed, sample diversity, and robustness.
****

Zhekai Du,Yinjie Min,Jingjing Li,Ke Lu,Changliang Zou,Liuhua Peng,Tingjin Chu,Mingming Gong

LoCA: Location-Aware Cosine Adaptation for Parameter-Efficient Fine-Tuning

Low-rank adaptation (LoRA) has become a prevalent method for adapting pre-trained large language models to downstream tasks. However, the simple low-rank decomposition form may constrain the optimization flexibility. To address this limitation, we introduce Location-aware Cosine Adaptation (LoCA), a novel frequency-domain parameter-efficient fine-tuning method based on inverse Discrete Cosine Transform (iDCT) with selective locations of learnable components. We begin with a comprehensive theoretical comparison between frequency-domain and low-rank decompositions for fine-tuning pre-trained large models. Our analysis reveals that frequency-domain approximation with carefully selected frequency components can surpass the expressivity of traditional low-rank-based methods. Furthermore, we demonstrate that iDCT offers a more efficient implementation compared to inverse Discrete Fourier Transform (iDFT), allowing for better selection and tuning of frequency components while maintaining equivalent expressivity to the optimal iDFT-based adaptation. By employing finite-difference approximation to estimate gradients for discrete locations of learnable coefficients on the DCT spectrum, LoCA dynamically selects the most informative frequency components during training. Experiments on diverse language and vision fine-tuning tasks demonstrate that LoCA offers enhanced parameter efficiency while maintains computational feasibility comparable to low-rank-based methods.
****

Bing Sun,Jun Sun,Wei Zhao

Democratic Training Against Universal Adversarial Perturbations

Despite their advances and success, real-world deep neural networks are known to be vulnerable to adversarial attacks. Universal adversarial perturbation, an input-agnostic attack, poses a serious threat for them to be deployed in security-sensitive systems. In this case, a single universal adversarial perturbation deceives the model on a range of clean inputs without requiring input-specific optimization, which makes it particularly threatening. In this work, we observe that universal adversarial perturbations usually lead to abnormal entropy spectrum in hidden layers, which suggests that the prediction is dominated by a small number of ``feature'' in such cases (rather than democratically by many features). Inspired by this, we propose an efficient yet effective defense method for mitigating UAPs called \emph{Democratic Training} by performing entropy-based model enhancement to suppress the effect of the universal adversarial perturbations in a given model. \emph{Democratic Training} is evaluated with 7 neural networks trained on 5 benchmark datasets and 5 types of state-of-the-art universal adversarial attack methods. The results show that it effectively reduces the attack success rate, improves model robustness and preserves the model accuracy on clean samples.
****

Severi Rissanen,Markus Heinonen,Arno Solin

Free Hunch: Denoiser Covariance Estimation for Diffusion Models Without Extra Costs

The covariance for clean data given a noisy observation is an important quantity in many training-free guided generation methods for diffusion models. Current methods require heavy test-time computation, altering the standard diffusion training process or denoiser architecture, or making heavy approximations. We propose a new framework that sidesteps these issues by using covariance information that is available for free from training data and the curvature of the generative trajectory, which is linked to the covariance through the second-order Tweedie's formula. We integrate these sources of information using (i) a novel method to transfer covariance estimates across noise levels and (ii) low-rank updates in a given noise level. We validate the method on linear inverse problems, where it outperforms recent baselines, especially with fewer diffusion steps.

****

Yukun Chen,Shuo Shao,Enhao Huang,Yiming Li,Pin-Yu Chen,Zhan Qin,Kui Ren

REFINE: Inversion-Free Backdoor Defense via Model Reprogramming

Backdoor attacks on deep neural networks (DNNs) have emerged as a significant security threat, allowing adversaries to implant hidden malicious behaviors during the model training phase. Pre-processing-based defense, which is one of the most important defense paradigms, typically focuses on input transformations or backdoor trigger inversion (BTI) to deactivate or eliminate embedded backdoor triggers during the inference process. However, these methods suffer from inherent limitations: transformation-based defenses often fail to balance model utility and defense performance, while BTI-based defenses struggle to accurately reconstruct trigger patterns without prior knowledge. In this paper, we propose REFINE, an inversion-free backdoor defense method based on model reprogramming. REFINE consists of two key components: \textbf{(1)} an input transformation module that disrupts both benign and backdoor patterns, generating new benign features; and \textbf{(2)} an output remapping module that redefines the model's output domain to guide the input transformations effectively. By further integrating supervised contrastive loss, REFINE enhances the defense capabilities while maintaining model utility. Extensive experiments on various benchmark datasets demonstrate the effectiveness of our REFINE and its resistance to potential adaptive attacks.

****

Ziyu Tang,Weicai Ye,Yifan Wang,Di Huang,Hujun Bao,Tong He,Guofeng Zhang

ND-SDF: Learning Normal Deflection Fields for High-Fidelity Indoor Reconstruction

Neural implicit reconstruction via volume rendering has demonstrated its effectiveness in recovering dense 3D surfaces. However, it is non-trivial to simultaneously recover meticulous geometry and preserve smoothness across regions with differing characteristics. To address this issue, previous methods typically employ geometric priors, which are often constrained by the performance of the prior models. In this paper, we propose ND-SDF, which learns a Normal Deflection field to represent the angular deviation between the scene normal and the prior normal. Unlike previous methods that uniformly apply geometric priors on all samples, introducing significant bias in accuracy, our proposed normal deflection field dynamically learns and adapts the utilization of samples based on their specific characteristics, thereby improving both the accuracy and effectiveness of the model. Our method not only obtains smooth weakly textured regions such as walls and floors but also preserves the geometric details of complex structures. In addition, we introduce a novel ray sampling strategy based on the deflection angle to facilitate the unbiased rendering process, which significantly improves the quality and accuracy of intricate surfaces, especially on thin structures. Consistent improvements on various challenging datasets demonstrate the superiority of our method.

****

Yuxing Liu,Rui Pan,Tong Zhang

AdaGrad under Anisotropic Smoothness

Adaptive gradient methods have been widely adopted in training large-scale deep neural networks, especially large foundation models. Despite the huge success in

practice, their theoretical advantages over classical gradient methods with uni
form step sizes across all coordinates (e.g. SGD) have not been fully understood
, especially in the large batch-size setting commonly used in practice. This is
because the only theoretical result that can demonstrate this benefit was obtain
ed in the original paper of Adagrad for convex nonsmooth objective functions, wh
ich is insufficient for large batch algorithms. In this work, we attempt to reso
lve this gap between theory and practice by proposing a novel anisotropic genera
lized smoothness assumption and providing corresponding analysis of Adagrad. It
is shown that under anisotropic smoothness and noise conditions, AdaGrad can ach
ieve faster convergence guarantees in terms of better dimensional dependence tha
n algorithms with uniform step sizes across all coordinates. Experiments in logi
stic regression and instruction following fine-tuning tasks provide strong evide
nce to support our novel assumption and theoretical analysis.
****
Yuang Peng,Yuxin Cui,Haomiao Tang,Zekun Qi,Runpei Dong,Jing Bai,Chunrui Han,Zhen
g Ge,Xiangyu Zhang,Shu-Tao Xia
DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation
Personalized image generation holds great promise in assisting humans in everyda
y work and life due to its impressive function in creatively generating personal
ized content. However, current evaluations either are automated but misalign wit
h humans or require human evaluations that are time-consuming and expensive. In
this work, we present DreamBench++, a human-aligned benchmark that advanced mult
imodal GPT models automate. Specifically, we systematically design the prompts t
o let GPT be both human-aligned and self-aligned, empowered with task reinforcem
ent. Further, we construct a comprehensive dataset comprising diverse images and
 prompts. By benchmarking 7 modern generative models, we demonstrate that \dream
bench results in significantly more human-aligned evaluation, helping boost the
community with innovative findings.
****
Charlie Victor Snell,Jaehoon Lee,Kelvin Xu,Aviral Kumar
Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Param
eters for Reasoning
Enabling LLMs to improve their outputs by using more test-time compute is a crit
ical step towards building self-improving agents that can operate on open-ended
natural language. In this paper, we scale up inference-time computation in LLMs,
 with a focus on answering: if an LLM is allowed to use a fixed but non-trivial
amount of inference-time compute, how much can it improve its performance on a c
hallenging prompt? Answering this question has implications not only on performa
nce, but also on the future of LLM pretraining and how to tradeoff inference-tim
e and pre-training compute. Little research has attempted to understand the scal
ing behaviors of test-time inference methods, with current work largely providin
g negative results for a number of these strategies. In this work, we analyze tw
o primary mechanisms to scale test-time computation: (1) searching against dense
, process-based verifier reward models (PRMs); and (2) updating the model's dist
ribution over a response adaptively, given the prompt at test time. We find that
 in both cases, the effectiveness of different approaches to scaling test-time c
ompute critically varies depending on the difficulty of the prompt. This observa
tion motivates applying a "compute-optimal" scaling strategy, which acts to, as
effectively as possible, allocate test-time compute per prompt in an adaptive ma
nner. Using this compute-optimal strategy, we can improve the efficiency of test
-time compute scaling for math reasoning problems by more than 4x compared to a
best-of-N baseline. Additionally, in a FLOPs-matched evaluation, we find that on
 problems where a smaller base model attains somewhat non-trivial success rates,
 test-time compute can be used to outperform a 14x larger model.
****
Abbas Abdolmaleki,Bilal Piot,Bobak Shahriari,Jost Tobias Springenberg,Tim Hertwe
ck,Michael Bloesch,Rishabh Joshi,Thomas Lampe,Junhyuk Oh,Nicolas Heess,Jonas Buc
hli,Martin Riedmiller
Learning from negative feedback, or positive feedback or both
Existing preference optimization methods often assume scenarios where paired pre

ference feedback (preferred/positive vs. dis-preferred/negative examples) is ava
ilable. This requirement limits their applicability in scenarios where only unpa
ired feedback—for example, either positive or negative— is available. To address
 this, we introduce a novel approach that decouples learning from positive and n
egative feedback. This decoupling enables control over the influence of each fee
dback type and, importantly, allows learning even when only one feedback type is
 present. A key contribution is demonstrating stable learning from negative feed
back alone, a capability not well-addressed by current methods. Our approach bui
lds upon the probabilistic framework introduced in (Dayan and Hinton, 1997), whi
ch uses expectation-maximization (EM) to directly optimize the probability of po
sitive outcomes (as opposed to classic expected reward maximization). We address
 a key limitation in current EM-based methods: they solely maximize the likeliho
od of positive examples, while neglecting negative ones. We show how to extend E
M algorithms to explicitly incorporate negative examples, leading to a theoretic
ally grounded algorithm that offers an intuitive and versatile way to learn from
 both positive and negative feedback. We evaluate our approach for training lang
uage models based on human feedback as well as training policies for sequential
decision-making problems, where learned value functions are available.
****

Rishub Tamirisa,Bhrugu Bharathi,Long Phan,Andy Zhou,Alice Gatti,Tarun Suresh,Max
well Lin,Justin Wang,Rowan Wang,Ron Arel,Andy Zou,Dawn Song,Bo Li,Dan Hendrycks,
Mantas Mazeika

Tamper-Resistant Safeguards for Open-Weight LLMs

Rapid advances in the capabilities of large language models (LLMs) have raised w
idespread concerns regarding their potential for malicious use. Open-weight LLMs
 present unique challenges, as existing safeguards lack robustness to tampering
attacks that modify model weights. For example, recent works have demonstrated t
hat refusal and unlearning safeguards can be trivially removed with a few steps
of fine-tuning. These vulnerabilities necessitate new approaches for enabling th
e safe release of open-weight LLMs. We develop a method, called TAR, for buildin
g tamper-resistant safeguards into open-weight LLMs such that adversaries cannot
 remove the safeguards even after hundreds of steps of fine-tuning. In extensive
 evaluations and red teaming analyses, we find that our method greatly improves
tamper-resistance while preserving benign capabilities. Our results demonstrate
that progress on tamper-resistance is possible, opening up a promising new avenu
e to improve the safety and security of open-weight LLMs.
****

Gen Li,Yuling Yan

O(d/T) Convergence Theory for Diffusion Probabilistic Models under Minimal Assum
ptions

Score-based diffusion models, which generate new data by learning to reverse a d
iffusion process that perturbs data from the target distribution into noise, hav
e achieved remarkable success across various generative tasks. Despite their sup
erior empirical performance, existing theoretical guarantees are often constrain
ed by stringent assumptions or suboptimal convergence rates. In this paper, we e
stablish a fast convergence theory for the denoising diffusion probabilistic mod
el (DDPM), a widely used SDE-based sampler, under minimal assumptions. Our analy
sis shows that, provided ℓ_{2}-accurate estimates of the score functions, t
he total variation distance between the target and generated distributions is up
per bounded by $O(d/T)$ (ignoring logarithmic factors), where d is the data di
mensionality and T is the number of steps. This result holds for any target di
stribution with finite first-order moment. To our knowledge, this improves upon
existing convergence theory for the DDPM sampler, while imposing minimal assumpt
ions on the target data distribution and score estimates. This is achieved throu
gh a novel set of analytical tools that provides a fine-grained characterization
 of how the error propagates at each step of the reverse process.
****

Ziteng Wang,Jun Zhu,Jianfei Chen

ReMoE: Fully Differentiable Mixture-of-Experts with ReLU Routing

Sparsely activated Mixture-of-Experts (MoE) models are widely adopted to scale u

p model capacity without increasing the computation budget. However, vanilla Top K routers are trained in a discontinuous, non-differentiable way, limiting their performance and scalability.
To address this issue, we propose ReMoE, a fully differentiable MoE architecture that offers a simple yet effective drop-in replacement for the conventional Top K+Softmax routing, utilizing ReLU as the router instead. We further propose methods to regulate the router's sparsity while balancing the load among experts. ReMoE’s continuous nature enables efficient dynamic allocation of computation across tokens and layers, while also exhibiting domain specialization. Our experiments demonstrate that ReMoE consistently outperforms vanilla TopK-routed MoE across various model sizes, expert counts, and levels of granularity. Furthermore, ReMoE exhibits superior scalability with respect to the number of experts, surpassing traditional MoE architectures. The implementation based on Megatron-LM is available at https://github.com/thu-ml/ReMoE.
****
Jeremy Carleton,Prathik Vijaykumar,Divyanshu Saxena,Dheeraj Narasimha,Srinivas Shakkottai,Aditya Akella
CONGO: Compressive Online Gradient Optimization
We address the challenge of zeroth-order online convex optimization where the objective function's gradient exhibits sparsity, indicating that only a small number of dimensions possess non-zero gradients. Our aim is to leverage this sparsity to obtain useful estimates of the objective function's gradient even when the only information available is a limited number of function samples. Our motivation stems from the optimization of large-scale queueing networks that process time-sensitive jobs. Here, a job must be processed by potentially many queues in sequence to produce an output, and the service time at any queue is a function of the resources allocated to that queue. Since resources are costly, the end-to-end latency for jobs must be balanced with the overall cost of the resources used. While the number of queues is substantial, the latency function primarily reacts to resource changes in only a few, rendering the gradient sparse. We tackle this problem by introducing the Compressive Online Gradient Optimization framework which allows compressive sensing methods previously applied to stochastic optimization to achieve regret bounds with an optimal dependence on the time horizon without the full problem dimension appearing in the bound. For specific algorithms, we reduce the samples required per gradient estimate to scale with the gradient's sparsity factor rather than its full dimensionality. Numerical simulations and real-world microservices benchmarks demonstrate CONGO's superiority over gradient descent approaches that do not account for sparsity.
****
Bohan Tang,Zexi Liu,Keyue Jiang,Siheng Chen,Xiaowen Dong
Training-Free Message Passing for Learning on Hypergraphs
Hypergraphs are crucial for modelling higher-order interactions in real-world data. Hypergraph neural networks (HNNs) effectively utilise these structures by message passing to generate informative node features for various downstream tasks like node classification. However, the message passing module in existing HNNs typically requires a computationally intensive training process, which limits their practical use. To tackle this challenge, we propose an alternative approach by decoupling the usage of hypergraph structural information from the model learning stage. This leads to a novel training-free message passing module, named TF-MP-Module, which can be precomputed in the data preprocessing stage, thereby reducing the computational burden. We refer to the hypergraph neural network equipped with our TF-MP-Module as TF-HNN. We theoretically support the efficiency and effectiveness of TF-HNN by showing that: 1) It is more training-efficient compared to existing HNNs; 2) It utilises as much information as existing HNNs for node feature generation; and 3) It is robust against the oversmoothing issue while using long-range interactions. Experiments based on seven real-world hypergraph benchmarks in node classification and hyperlink prediction show that, compared to state-of-the-art HNNs, TF-HNN exhibits both competitive performance and superior training efficiency. Specifically, on the large-scale benchmark, Trivago, TF-HNN outperforms the node classification accuracy of the best baseline by 10% wi

th just 1% of the training time of that baseline.
****
Hao Wang,Lichen Pan,Yuan Shen,Zhichao Chen,Degui Yang,Yifei Yang,Sen Zhang,Xingg
ao Liu,Haoxuan Li,Dacheng Tao
FreDF: Learning to Forecast in the Frequency Domain
Time series modeling presents unique challenges due to autocorrelation in both h
istorical data and future sequences. While current research predominantly addres
ses autocorrelation within historical data, the correlations among future labels
 are often overlooked. Specifically, modern forecasting models primarily adhere
to the Direct Forecast (DF) paradigm, generating multi-step forecasts independen
tly and disregarding label correlations over time. In this work, we demonstrate
that the learning objective of DF is biased in the presence of label correlation
. To address this issue, we propose the Frequency-enhanced Direct Forecast (FreD
F), which mitigates label correlation by learning to forecast in the frequency d
omain, thereby reducing estimation bias. Our experiments show that FreDF signifi
cantly outperforms existing state-of-the-art methods and is compatible with a va
riety of forecast models. Code is available at https://github.com/Master-PLC/Fre
DF.
****
Themistoklis Haris
kNN Attention Demystified: A Theoretical Exploration for Scalable Transformers
Despite their power, Transformers face challenges with long sequences due to the
 quadratic complexity of self-attention. To address this limitation, methods lik
e k-Nearest-Neighbor (kNN) attention have been introduced [Roy et al., 2017]
, enabling each token to attend to only its k closest tokens. While kNN atte
ntion has shown empirical success in making Transformers more efficient, its exa
ct approximation guarantees have not been theoretically analyzed. In this work,
we establish a theoretical framework for kNN attention, reformulating self-att
ention as expectations over softmax distributions and leveraging lazy Gumbel sam
pling [Mussmann et al., 2017] with kNN indices for efficient approximation. Bu
ilding on this framework, we also propose novel sub-quadratic algorithms that ap
proximate self-attention gradients by leveraging efficient sampling techniques,
such as Markov Chain-based estimation. Finally, we demonstrate the practical eff
ectiveness of these algorithms through empirical experiments, showcasing their b
enefits in both training and inference.
****
Hanmo Liu,Shimin Di,Jialiang Wang,Zhili Wang,Jiachuan Wang,Xiaofang Zhou,Lei Che
n
Structuring Benchmark into Knowledge Graphs to Assist Large Language Models in R
etrieving and Designing Models
In recent years, the design and transfer of neural network models have been wide
ly studied due to their exceptional performance and capabilities. However, the c
omplex nature of datasets and the vast architecture space pose significant chall
enges for both manual and automated algorithms in creating high-performance mode
ls. Inspired by researchers who design, train, and document the performance of v
arious models across different datasets, this paper introduces a novel schema th
at transforms the benchmark data into a Knowledge Benchmark Graph (KBG), which p
rimarily stores the facts in the form of performance(data, model). Constructing
the KBG facilitates the structured storage of design knowledge, aiding subsequen
t model design and transfer. However, it is a non-trivial task to retrieve or de
sign suitable neural networks based on the KBG, as real-world data are often off
 the records. To tackle this challenge, we propose transferring existing models
stored in KBG by establishing correlations between unseen and previously seen da
tasets. Given that measuring dataset similarity is a complex and open-ended issu
e, we explore the potential for evaluating the correctness of the similarity fun
ction. Then, we further integrate the KBG with Large Language Models (LLMs), ass
isting LLMs to think and retrieve existing model knowledge in a manner akin to h
umans when designing or transferring models. We demonstrate our method specifica
lly in the context of Graph Neural Network (GNN) architecture design, constructi
ng a KBG (with 26,206 models, 211,669 performance records, and 2,540,064 facts)

and validating the effectiveness of leveraging the KBG to promote GNN architectu
re design.
****
Evan Z Wang,Federico Cassano,Catherine Wu,Yunfeng Bai,William Song,Vaskar Nath,Z
iwen Han,Sean M. Hendryx,Summer Yue,Hugh Zhang
Planning in Natural Language Improves LLM Search for Code Generation
While scaling training compute has led to remarkable improvements in large langu
age models (LLMs), scaling inference compute only recently began to yield analog
ous gains. We hypothesize that a core missing component is a lack of diverse LLM
 outputs, leading to inefficient search due to models repeatedly sampling highly
 similar, yet incorrect generations. We empirically demonstrate that this lack o
f diversity can be mitigated by searching over candidate plans for solving a pro
blem in natural language. Based on this insight, we propose PlanSearch, a novel
search algorithm which shows strong results across HumanEval+, MBPP+, and LiveCo
deBench (a contamination-free benchmark for competitive coding). PlanSearch gene
rates a diverse set of observations about the problem and uses these observation
s to construct plans for solving the problem. By searching over plans in natural
 language rather than directly over code solutions, PlanSearch explores a signif
icantly more diverse range of potential solutions compared to baseline search me
thods. Using PlanSearch on top of Claude 3.5 Sonnet achieves a pass@200 of 77.0%
 on LiveCodeBench, outperforming both the best pass-rate achieved without any se
arch (pass@1 = 41.4%) and using standard repeated sampling on top of existing no
n-search models (pass@200 = 60.6%). Finally, we show that, across all models, se
arch algorithms, and benchmarks analyzed, we can accurately predict performance
gains from search as a function of the diversity over generated ideas.
****
Fan Yao,Yuwei Cheng,Ermin Wei,Haifeng Xu
Single-agent Poisoning Attacks Suffice to Ruin Multi-Agent Learning
We investigate the robustness of multi-agent learning in strongly monotone games
 with bandit feedback. While previous research has developed learning algorithms
 that achieve last-iterate convergence to the unique Nash equilibrium (NE) at a
polynomial rate, we demonstrate that all such algorithms are vulnerable to adver
saries capable of poisoning even a single agent's utility observations. Specific
ally, we propose an attacking strategy such that for any given time horizon T,
 the adversary can mislead any multi-agent learning algorithm to converge to a p
oint other than the unique NE with a corruption budget that grows sublinearly in
 T. To further understand the inherent robustness of these algorithms, we char
acterize the fundamental trade-off between convergence speed and the maximum tol
erable total utility corruptions for two example algorithms, including the state
-of-the-art one. Our theoretical and empirical results reveal an intrinsic effic
iency-robustness trade-off: the faster an algorithm converges, the more vulnerab
le it becomes to utility poisoning attacks. To the best of our knowledge, this i
s the first work to identify and characterize such a trade-off in the context of
 multi-agent learning.
****
Shicheng Xu,Liang Pang,Yunchang Zhu,Huawei Shen,Xueqi Cheng
Cross-Modal Safety Mechanism Transfer in Large Vision-Language Models
Vision-language alignment in Large Vision-Language Models (LVLMs) successfully e
nables LLMs to understand visual input. However, we find that existing vision-la
nguage alignment methods fail to transfer the existing safety mechanism for text
 in LLMs to vision, which leads to vulnerabilities in toxic image. To explore th
e cause of this problem, we give the insightful explanation of where and how the
 safety mechanism of LVLMs operates and conduct comparative analysis between tex
t and vision. We find that the hidden states at the specific transformer layers
play a crucial role in the successful activation of safety mechanism, while the
vision-language alignment at hidden states level in current methods is insuffici
ent. This results in a semantic shift for input images compared to text in hidde
n states, therefore misleads the safety mechanism. To address this, we propose a
 novel Text-Guided vision-language Alignment method (TGA) for LVLMs. TGA retriev
es the texts related to input vision and uses them to guide the projection of vi

sion into the hidden states space in LLMs. Experiments show that \textbf{TGA} no
t only successfully transfers the safety mechanism for text in basic LLMs to vis
ion in vision-language alignment for LVLMs without any safety fine-tuning on the
 visual modality but also maintains the general performance on various vision ta
sks (Safe and Good). Code is in supplemental material and will be released on Gi
tHub after acceptance.
****
Simran Kaur,Simon Park,Anirudh Goyal,Sanjeev Arora
Instruct-SkillMix: A Powerful Pipeline for LLM Instruction Tuning
We introduce INSTRUCT-SKILLMIX, an automated approach for creating diverse, high
 quality SFT data for instruction-following. The pipeline involves two stages, e
ach leveraging an existing powerful LLM: (1) Skill extraction: uses the LLM to e
xtract core “skills” for instruction-following by directly prompting the model.
This is inspired by “LLM metacognition” of (Didolkar et al., 2024); (2) Data gen
eration: uses the powerful LLM to generate (instruction, response) data that
exhibit a randomly chosen pair of these skills. Here, the use of random skill co
mbinations promotes diversity and difficulty. The estimated cost of creating the
 dataset is under $600.

Vanilla SFT (i.e., no PPO, DPO, or RL methods) on data generated from INSTRUCT-S
KILLMIX leads to strong gains on instruction following benchmarks such as Alpaca
Eval 2.0, MT-Bench, and WildBench. With just 4K examples, LLaMA-3-8B-Base achiev
es 42.76% length-controlled win rate on AlpacaEval 2.0, a level similar to front
ier models like Claude 3 Opus and LLaMA-3.1-405B-Instruct. Ablation studies also
 suggest plausible reasons for why creating open instruction-tuning datasets via
 naive crowd-sourcing has proved difficult. In our dataset,adding 20% low qualit
y answers (“shirkers”) causes a noticeable degradation in performance.

The INSTRUCT-SKILLMIX pipeline seems flexible and adaptable to other settings.
****
Shannon How Shi Qi,Jagmohan Chauhan,Geoff V. Merrett,Jonathon Hare
FedTMOS: Efficient One-Shot Federated Learning with Tsetlin Machine
One-Shot Federated Learning (OFL) is a promising approach that reduce communicat
ion to a single round, minimizing latency and resource consumption. However, exi
sting OFL methods often rely on Knowledge Distillation, which introduce server-s
ide training, increasing latency. While neuron matching and model fusion techniq
ues bypass server-side training, they struggle with alignment when heterogeneous
 data is present. To address these challenges, we proposed One-Shot Federated Le
arning with Tsetlin Machine (FedTMOS), a novel data-free OFL framework built upo
n the low-complexity and class-adaptive properties of the Tsetlin Machine. FedTM
OS first clusters then reassigns class-specific weights to form models using an
inter-class maximization approach, efficiently generating balanced server models
 without requiring additional training. Our extensive experiments demonstrate th
at FedTMOS significantly outperforms its ensemble counterpart by an average of $
6.16$%, and the leading state-of-the-art OFL baselines by 7.22% across various
 OFL settings. Moreover, FedTMOS achieves at least a $2.3\times$ reduction in up
load communication costs and a $75\times$ reduction in server latency compared t
o methods requiring server-side training. These results establish FedTMOS as a h
ighly efficient and practical solution for OFL scenarios.
****
Junqi Shi,Zhujia Chen,Hanfei Li,Qi Zhao,Ming Lu,Tong Chen,Zhan Ma
On Quantizing Neural Representation for Variable-Rate Video Coding
This work introduces NeuroQuant, a novel post-training quantization (PTQ) approa
ch tailored to non-generalized Implicit Neural Representations for variable-rate
 Video Coding (INR-VC). Unlike existing methods that require extensive weight re
training for each target bitrate, we hypothesize that variable-rate coding can b
e achieved by adjusting quantization parameters (QPs) of pre-trained weights. Ou
r study reveals that traditional quantization methods, which assume inter-layer
independence, are ineffective for non-generalized INR-VC models due to significa
nt dependencies across layers. To address this, we redefine variable-rate INR-VC

as a mixed-precision quantization problem and establish a theoretical framework for sensitivity criteria aimed at simplified, fine-grained rate control. Additionally, we propose network-wise calibration and channel-wise quantization strategies to minimize quantization-induced errors, arriving at a unified formula for representation-oriented PTQ calibration. Our experimental evaluations demonstrate that NeuroQuant significantly outperforms existing techniques in varying bitwidth quantization and compression efficiency, accelerating encoding by up to eight times and enabling quantization down to INT2 with minimal reconstruction loss. This work introduces variable-rate INR-VC for the first time and lays a theoretical foundation for future research in rate-distortion optimization, advancing the field of video coding technology. The materials
will be available at https://github.com/Eric-qi/NeuroQuant.
****

Bahare Fatemi,Mehran Kazemi,Anton Tsitsulin,Karishma Malkan,Jinyeong Yim,John Palowitch,Sungyong Seo,Jonathan Halcrow,Bryan Perozzi
Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning
Large language models (LLMs) have showcased remarkable reasoning capabilities, yet they remain susceptible to errors, particularly in temporal reasoning tasks involving complex temporal logic. Existing research has explored LLM performance on temporal reasoning using diverse datasets and benchmarks. However, these studies often rely on real-world data that LLMs may have encountered during pre-training or employ anonymization techniques that can inadvertently introduce factual inconsistencies. In this work, we address these limitations by introducing novel synthetic datasets specifically designed to assess LLM temporal reasoning abilities in various scenarios. The diversity of question types across these datasets enables systematic investigation into the impact of the problem structure, size, question type, fact order, and other factors on LLM performance. Our findings provide valuable insights into the strengths and weaknesses of current LLMs in temporal reasoning tasks. To foster further research in this area, we will open-source the datasets and evaluation framework used in our experiments.
****

Zhuohua Li,Maoli Liu,Xiangxiang Dai,John C.S. Lui
Demystifying Online Clustering of Bandits: Enhanced Exploration Under Stochastic and Smoothed Adversarial Contexts
The contextual multi-armed bandit (MAB) problem is crucial in sequential decision-making. A line of research, known as online clustering of bandits, extends contextual MAB by grouping similar users into clusters, utilizing shared features to improve learning efficiency. However, existing algorithms, which rely on the upper confidence bound (UCB) strategy, struggle to gather adequate statistical information to accurately identify unknown user clusters. As a result, their theoretical analyses require several strong assumptions about the "diversity" of contexts generated by the environment, leading to impractical settings, complicated analyses, and poor practical performance. Removing these assumptions has been a long-standing open problem in the clustering of bandits literature. In this work, we provide two partial solutions. First, we introduce an additional exploration phase to accelerate the identification of clusters. We integrate this general strategy into both graph-based and set-based algorithms and propose two new algorithms, UniCLUB and UniSCLUB. Remarkably, our algorithms require substantially weaker assumptions and simpler theoretical analyses while achieving superior cumulative regret compared to previous studies. Second, inspired by the smoothed analysis framework, we propose a more practical setting that eliminates the requirement for i.i.d. context generation used in previous studies, thus enhancing the performance of existing algorithms for online clustering of bandits. Extensive evaluations on both synthetic and real-world datasets demonstrate that our proposed algorithms outperform existing approaches.
****

Peiran Dong,Haowei Li,Song Guo
Durable Quantization Conditioned Misalignment Attack on Large Language Models
As large language models (LLMs) are increasingly deployed on resource-constrained edge devices, quantization techniques have been widely adopted to reduce model

size and computational requirements. However, this process can expose models to new vulnerabilities. In this work, we introduce the Quantization Conditioned Misalignment (Q-Misalign) attack, a novel threat in which safety misalignment remains dormant in a full-precision LLM but becomes exploitable post-quantization. We demonstrate that our Q-Misalign attack effectively bypasses safety mechanisms and enables the generation of harmful content in quantized models while maintaining full-precision performance. Furthermore, we propose a contrastive task vector-based approach to enhance attack durability, ensuring that vulnerabilities persist even after downstream fine-tuning. Experimental results show that Q-Misalign attack significantly increases jailbreak success rates in quantized models, while preserving model utility and safety alignment in full precision. Our findings highlight a critical gap in current LLM safety measures and call for more robust defenses in quantization-aware scenarios.
****
Alexander Levine,Peter Stone,Amy Zhang
Learning a Fast Mixing Exogenous Block MDP using a Single Trajectory
In order to train agents that can quickly adapt to new objectives or reward functions, efficient unsupervised representation learning in sequential decision-making environments can be important. Frameworks such as the Exogenous Block Markov Decision Process (Ex-BMDP) have been proposed to formalize this representation-learning problem (Efroni et al., 2022b). In the Ex-BMDP framework, the agent's high-dimensional observations of the environment have two latent factors: a controllable factor, which evolves deterministically within a small state space according to the agent's actions, and an exogenous factor, which represents time-correlated noise, and can be highly complex. The goal of the representation learning problem is to learn an encoder that maps from observations into the controllable latent space, as well as the dynamics of this space. Efroni et al. (2022b) has shown that this is possible with a sample complexity that depends only on the size of the controllable latent space, and not on the size of the noise factor. However, this prior work has focused on the episodic setting, where the controllable latent state resets to a specific start state after a finite horizon.

By contrast, if the agent can only interact with the environment in a single continuous trajectory, prior works have not established sample-complexity bounds. We propose STEEL, the first provably sample-efficient algorithm for learning the controllable dynamics of an Ex-BMDP from a single trajectory, in the function approximation setting. STEEL has a sample complexity that depends only on the sizes of the controllable latent space and the encoder function class, and (at worst linearly) on the mixing time of the exogenous noise factor. We prove that STEEL is correct and sample-efficient, and demonstrate STEEL on two toy problems. Code is available at: https://github.com/midi-lab/steel.
****
Aliyah R. Hsu,Georgia Zhou,Yeshwanth Cherapanamjeri,Yaxuan Huang,Anobel Odisho,Peter R. Carroll,Bin Yu
Efficient Automated Circuit Discovery in Transformers using Contextual Decomposition
Automated mechanistic interpretation research has attracted great interest due to its potential to scale explanations of neural network internals to large models. Existing automated circuit discovery work relies on activation patching or its approximations to identify subgraphs in models for specific tasks (circuits). They often suffer from slow runtime, approximation errors, and specific requirements of metrics, such as non-zero gradients.
In this work, we introduce contextual decomposition for transformers (CD-T) to build interpretable circuits in large language models. CD-T can produce circuits at any level of abstraction and is the first to efficiently produce circuits as fine-grained as attention heads at specific sequence positions.
CD-T is compatible to all transformer types, and requires no training or manually-crafted examples.
CD-T consists of a set of mathematical equations to isolate contribution of model features. Through recursively computing contribution of all nodes in a computa

tional graph of a model using CD-T followed by pruning, we are able to reduce circuit discovery runtime from hours to seconds compared to state-of-the-art baselines.
On three standard circuit evaluation datasets (indirect object identification, greater-than comparisons, and docstring completion),
we demonstrate that CD-T outperforms ACDC and EAP by better recovering the manual circuits with an average of 97% ROC AUC under low runtimes.
In addition, we provide evidence that faithfulness of CD-T circuits is not due to random chance by showing our circuits are 80% more faithful than random circuits of up to 60% of the original model size.
Finally, we show CD-T circuits are able to perfectly replicate original models' behavior(faithfulness = 1) using fewer nodes than the baselines for all tasks.
Our results underscore the great promise of CD-T for efficient automated mechanistic interpretability, paving the way for new insights into the workings of large language models.
****
Tanya Chowdhury,Yair Zick,James Allan
RankSHAP: Shapley Value Based Feature Attributions for Learning to Rank
Numerous works propose post-hoc, model-agnostic explanations for learning to rank, focusing on ordering entities by their relevance to a query through feature attribution methods. However, these attributions often weakly correlate or contradict each other, confusing end users. We adopt an axiomatic game-theoretic approach, popular in the feature attribution community, to identify a set of fundamental axioms that every ranking-based feature attribution method should satisfy.
We then introduce Rank-SHAP, extending classical Shapley values to ranking. We evaluate the RankSHAP framework through extensive experiments on two datasets, multiple ranking methods and evaluation metrics. Additionally, a user study confirms RankSHAP’s alignment with human intuition. We also perform an axiomatic analysis of existing rank attribution algorithms to determine their compliance with our proposed axioms. Ultimately, our aim is to equip practitioners with a set of axiomatically backed feature attribution methods for studying IR ranking models, that ensure generality as well as consistency.
****
Quentin Bouniot,Pavlo Mozharovskyi,Florence d'Alché-Buc
Tailoring Mixup to Data for Calibration
Among all data augmentation techniques proposed so far, linear interpolation of training samples, also called Mixup, has found to be effective for a large panel of applications.
Along with improved predictive performance, Mixup is also a good technique for improving calibration.
However, mixing data carelessly can lead to manifold mismatch, i.e., synthetic data lying outside original class manifolds, which can deteriorate calibration.
In this work, we show that the likelihood of assigning a wrong label with mixup increases with the distance between data to mix.
To this end, we propose to dynamically change the underlying distributions of interpolation coefficients
depending on the similarity between samples to mix, and define a flexible framework to do so without losing in diversity. We provide extensive experiments for classification and regression tasks, showing that our proposed method improves predictive performance
and calibration of models, while being much more efficient.
****
Jiaxing Xu,Yongqiang Chen,Xia Dong,Mengcheng Lan,Tiancheng HUANG,Qingtian Bian,James Cheng,Yiping Ke
BrainOOD: Out-of-distribution Generalizable Brain Network Analysis
In neuroscience, identifying distinct patterns linked to neurological disorders, such as Alzheimer's and Autism, is critical for early diagnosis and effective intervention. Graph Neural Networks (GNNs) have shown promising in analyzing brain networks, but there are two major challenges in using GNNs: (1) distribution s

hifts in multi-site brain network data, leading to poor Out-of-Distribution (OOD) generalization, and (2) limited interpretability in identifying key brain regi ons critical to neurological disorders. Existing graph OOD methods, while effect ive in other domains, struggle with the unique characteristics of brain networks . To bridge these gaps, we introduce BrainOOD, a novel framework tailored for b rain networks that enhances GNNs' OOD generalization and interpretability. Brain OOD framework consists of a feature selector and a structure extractor, which in corporates various auxiliary losses including an improved Graph Information Bott leneck (GIB) objective to recover causal subgraphs. By aligning structure select ion across brain networks and filtering noisy features, BrainOOD offers reliable interpretations of critical brain regions. Our approach outperforms 16 existing methods and improves generalization to OOD subjects by up to 8.5%. Case studies highlight the scientific validity of the patterns extracted, which aligns with the findings in known neuroscience literature. We also propose the first OOD bra in network benchmark, which provides a foundation for future research in this fi eld. Our code is available at https://github.com/AngusMonroe/BrainOOD.
****

Van Nguyen,Tingmin Wu,Xingliang YUAN,Marthie Grobler,Surya Nepal,Carsten Rudolph
AI2TALE: An Innovative Information Theory-based Approach for Learning to Localiz e Phishing Attacks
Phishing attacks remain a significant challenge for detection, explanation, and defense, despite over a decade of research on both technical and non-technical s olutions. AI-based phishing detection methods are among the most effective appro aches for defeating phishing attacks, providing predictions on the vulnerability label (i.e., phishing or benign) of data. However, they often lack intrinsic ex plainability, failing to identify the specific information that triggers the cla ssification. To this end, we propose AI2TALE, an innovative deep learning-based approach for email (the most common phishing medium) phishing attack localizatio n. Our method aims to not only predict the vulnerability label of the email data but also provide the capability to automatically learn and identify the most im portant and phishing-relevant information (i.e., sentences) in the phishing emai l data, offering useful and concise explanations for the identified vulnerabilit y.

Extensive experiments on seven diverse real-world email datasets demonstrate the capability and effectiveness of our method in selecting crucial information, en abling accurate detection and offering useful and concise explanations (via the most important and phishing-relevant information triggering the classification) for the vulnerability of phishing emails. Notably, our approach outperforms stat e-of-the-art baselines by 1.5% to 3.5% on average in Label-Accuracy and Cognitiv e-True-Positive metrics under a weakly supervised setting, where only vulnerabil ity labels are used without requiring ground truth phishing information.
****

Jingyuan Zhang,Yiyang Duan,Shuaicheng Niu,YANG CAO,Wei Yang Bryan Lim
Enhancing Federated Domain Adaptation with Multi-Domain Prototype-Based Federate d Fine-Tuning
Federated Domain Adaptation (FDA) is a Federated Learning (FL) scenario where mo dels are trained across multiple clients with unique data domains but a shared c ategory space, without transmitting private data. The primary challenge in FDA i s data heterogeneity, which causes significant divergences in gradient updates w hen using conventional averaging-based aggregation methods, reducing the efficac y of the global model. This further undermines both in-domain and out-of-domain performance (within the same federated system but outside the local client), whi ch is critical in certain business applications. To address this, we propose a n ovel framework called \textbf{M}ulti-domain \textbf{P}rototype-based \textbf{F}e derated Fine-\textbf{T}uning (MPFT). MPFT fine-tunes a pre-trained model using m ulti-domain prototypes, i.e., several pretrained representations enriched with d omain-specific information from category-specific local data. This enables super vised learning on the server to create a globally optimized adapter that is subs equently distributed to local clients, without the intrusion of data privacy. Em

pirical results show that MPFT significantly improves both in-domain and out-of-domain accuracy over conventional methods, enhancing knowledge preservation and adaptation in FDA. Notably, MPFT achieves convergence within a single communication round, greatly reducing computation and communication costs. To ensure privacy, MPFT applies differential privacy to protect the prototypes. Additionally, we develop a prototype-based feature space hijacking attack to evaluate robustness, confirming that raw data samples remain unrecoverable even after extensive training epochs. The complete implementation of MPFL is available at \url{https://anonymous.4open.science/r/DomainFL/}.
****

Kristian Georgiev,Roy Rinberg,Sung Min Park,Shivam Garg,Andrew Ilyas,Aleksander Madry,Seth Neel

Machine Unlearning via Simulated Oracle Matching

Machine unlearning---efficiently removing the effect of a small "forget set" of training data on a pre-trained machine learning model---has recently attracted significant research interest. Despite this interest, however, recent work shows that existing machine unlearning techniques do not hold up to thorough evaluation in non-convex settings. In this work, we introduce a new machine unlearning technique that exhibits strong empirical performance even in such challenging settings. Our starting point is the perspective that the goal of unlearning is to produce a model whose outputs are *statistically indistinguishable* from those of a model re-trained on all but the forget set. This perspective naturally suggests a reduction from the unlearning problem to that of *data attribution, where the goal is to predict the effect of changing the training set on a model's outputs. Thus motivated, we propose the following meta-algorithm, which we call Datamodel Matching (DMM): given a trained model, we (a) use data attribution to *predict* the output of the model if it were re-trained on all but the forget set points; then (b) *fine-tune* the pre-trained model to match these predicted outputs. In a simple convex setting, we show how this approach provably outperforms a variety of iterative unlearning algorithms. Empirically, we use a combination of existing evaluations and a new metric based on the KL-divergence to show that even in non-convex settings, DMM achieves strong unlearning performance relative to existing algorithms. An added benefit of DMM is that it is a meta-algorithm, in the sense that future advances in data attribution translate directly into better unlearning algorithms, pointing to a clear direction for future progress in unlearning.
****

Parsa Vahidi,Omid G. Sani,Maryam Shanechi

BRAID: Input-driven Nonlinear Dynamical Modeling of Neural-Behavioral Data

Neural populations exhibit complex recurrent structures that drive behavior, while continuously receiving and integrating external inputs from sensory stimuli, upstream regions, and neurostimulation. However, neural populations are often modeled as autonomous dynamical systems, with little consideration given to the influence of external inputs that shape the population activity and behavioral outcomes. Here, we introduce BRAID, a deep learning framework that models nonlinear neural dynamics underlying behavior while explicitly incorporating any measured external inputs. Our method disentangles intrinsic recurrent neural population dynamics from the effects of inputs by including a forecasting objective within input-driven recurrent neural networks. BRAID further prioritizes the learning of intrinsic dynamics that are related to a behavior of interest by using a multi-stage optimization scheme. We validate BRAID with nonlinear simulations, showing that it can accurately learn the intrinsic dynamics shared between neural and behavioral modalities. We then apply BRAID to motor cortical activity recorded during a motor task and demonstrate that our method more accurately fits the neural-behavioral data by incorporating measured sensory stimuli into the model and improves the forecasting of neural-behavioral data compared with various baseline methods, whether input-driven or not.
****

Aohan Zeng,Zhengxiao Du,Mingdao Liu,Lei Zhang,shengmin jiang,Yuxiao Dong,Jie Tang

Scaling Speech-Text Pre-training with Synthetic Interleaved Data
Speech language models (SpeechLMs) accept speech input and produce speech output , allowing for more natural human-computer interaction compared to text-based la rge language models (LLMs).
Traditional approaches for developing SpeechLMs are constrained by the limited a vailability of unsupervised speech data and parallel speech-text data, which are significantly less abundant compared to text pre-training data, thereby limitin g their scalability as LLMs.
We propose a novel approach to scaling speech-text pre-training by leveraging la rge-scale synthetic interleaved data derived from text corpora, eliminating the need for parallel speech-text datasets.
Our method efficiently constructs speech-text interleaved data by sampling text spans from existing text corpora and synthesizing corresponding speech spans usi ng a text-to-token model, bypassing the need to generate actual speech.
We also employ a supervised speech tokenizer derived from an automatic speech re cognition (ASR) model by incorporating a vector-quantized bottleneck into the e ncoder. This supervised training approach results in discrete speech tokens with strong semantic preservation even at lower sampling rates (e.g. 12.5Hz), while still maintaining speech reconstruction quality.
Starting from a pre-trained language model and scaling our pre-training to 1 tri llion tokens (with 600B synthetic interleaved speech-text data), we achieve stat e-of-the-art performance in both speech language modeling and spoken question an swering, improving performance on spoken questions tasks from the previous SOTA of 13\% (Moshi) to 31\%.
We further demonstrate that by fine-tuning the pre-trained model with speech dia logue data, we can develop an end-to-end spoken chatbot that achieves competitiv e performance comparable to existing baselines in both conversational abilities and speech quality, even operating exclusively in the speech domain.
****
Jakub Vrabel,Ori Shem-Ur,Yaron Oz,David Krueger
Input Space Mode Connectivity in Deep Neural Networks
We extend the concept of loss landscape mode connectivity to the input space of deep neural networks. Mode connectivity was originally studied within parameter space, where it describes the existence of low-loss paths between different solu tions (loss minimizers) obtained through gradient descent. We present theoretica l and empirical evidence of its presence in the input space of deep networks, th ereby highlighting the broader nature of the phenomenon. We observe that differe nt input images with similar predictions are generally connected, and for traine d models, the path tends to be simple, with only a small deviation from being a linear path. Our methodology utilizes real, interpolated, and synthetic inputs c reated using the input optimization technique for feature visualization. We conj ecture that input space mode connectivity in high-dimensional spaces is a geomet ric effect that takes place even in untrained models and can be explained throug h percolation theory. We exploit mode connectivity to obtain new insights about adversarial examples and demonstrate its potential for adversarial detection. Ad ditionally, we discuss applications for the interpretability of deep networks.
****
Zirui Zhao,Hanze Dong,Amrita Saha,Caiming Xiong,Doyen Sahoo
Automatic Curriculum Expert Iteration for Reliable LLM Reasoning
Hallucinations (i.e., generating plausible but inaccurate content) and laziness (i.e. excessive refusals or defaulting to "I don't know") persist as major chall enges in LLM reasoning. Current efforts to reduce hallucinations primarily focus on factual errors in knowledge-grounded tasks, often neglecting hallucinations related to faulty reasoning. Meanwhile, some approaches render LLMs overly conse rvative, limiting their problem-solving capabilities. To mitigate hallucination and laziness in reasoning tasks, we propose Automatic Curriculum Expert Iteratio n (Auto-CEI) to enhance LLM reasoning and align responses to the model’s capabil ities--assertively answering within its limits and declining when tasks exceed t hem. In our method, Expert Iteration explores the reasoning trajectories near th e LLM policy, guiding incorrect paths back on track to reduce compounding errors

and improve robustness; it also promotes appropriate "I don't know" responses a
fter sufficient reasoning attempts. The curriculum automatically adjusts rewards
, incentivizing extended reasoning before acknowledging incapability, thereby pu
shing the limits of LLM reasoning and aligning its behaviour with these limits.
We compare Auto-CEI with various SOTA baselines across logical reasoning, mathem
atics, and planning tasks, where Auto-CEI achieves superior alignment by effecti
vely balancing assertiveness and conservativeness.
****

Hamid Kalantari,Russell Greiner,Pouria Ramazi

Extendable and Iterative Structure Learning Strategy for Bayesian Networks

Learning the structure of Bayesian networks is a fundamental yet computationally
intensive task, especially as the number of variables grows. Traditional algori
thms require retraining from scratch when new variables are introduced, making t
hem impractical for dynamic or large-scale applications. In this paper, we propo
se an extendable structure learning strategy that efficiently incorporates a new
variable Y into an existing Bayesian network graph $\mathcal{G}$ over variabl
es $\mathcal{X}$, resulting in an updated P-map graph $\bar{\mathcal{G}}$ on $\b
ar{\mathcal{X}} = \mathcal{X} \cup \{Y\}$. By leveraging the information encoded
in $\mathcal{G}$, our method significantly reduces computational overhead compa
red to learning $\bar{\mathcal{G}}$ from scratch. Empirical evaluations demonstr
ate runtime reductions of up to 1300x without compromising accuracy. Building on
this approach, we introduce a novel iterative paradigm for structure learning o
ver $\mathcal{X}$. Starting with a small subset $\mathcal{U} \subset \mathcal{X}
$, we iteratively add the remaining variables using our extendable algorithms to
construct a P-map graph over the full set. This method offers runtime advantage
s comparable to common algorithms while maintaining similar accuracy. Our contri
butions provide a scalable solution for Bayesian network structure learning, ena
bling efficient model updates in real-time and high-dimensional settings.
****

Chenyu Wang,Sharut Gupta,Xinyi Zhang,Sana Tonekaboni,Stefanie Jegelka,Tommi Jaak
kola,Caroline Uhler

An Information Criterion for Controlled Disentanglement of Multimodal Data

Multimodal representation learning seeks to relate and decompose information inh
erent in multiple modalities. By disentangling modality-specific information fro
m information that is shared across modalities, we can improve interpretability
and robustness and enable downstream tasks such as the generation of counterfact
ual outcomes. Separating the two types of information is challenging since they
are often deeply entangled in many real-world applications. We propose $\textbf{
Disentangled}$ $\textbf{S}$elf-$\textbf{S}$upervised $\textbf{L}$earning (Disent
angledSSL), a novel self-supervised approach for learning disentangled represent
ations. We present a comprehensive analysis of the optimality of each disentangl
ed representation, particularly focusing on the scenario not covered in prior wo
rk where the so-called $\textit{Minimum Necessary Information}$ (MNI) point is n
ot attainable. We demonstrate that \algo successfully learns shared and modality
-specific features on multiple synthetic and real-world datasets and consistentl
y outperforms baselines on various downstream tasks, including prediction tasks
for vision-language data, as well as molecule-phenotype retrieval tasks for biol
ogical data.
****

Yiding Wang,Yuxuan Chen,Fangwei Zhong,Long Ma,Yizhou Wang

Simulating Human-like Daily Activities with Desire-driven Autonomy

Desires motivate humans to interact autonomously with the complex world. In cont
rast, current AI agents require explicit task specifications, such as instructio
ns or reward functions, which constrain their autonomy and behavioral diversity.
In this paper, we introduce a Desire-driven Autonomous Agent (D2A) that can ena
ble a large language model (LLM) to autonomously propose and select tasks, motiv
ated by satisfying its multi-dimensional desires. Specifically, the motivational
framework of D2A is mainly constructed by a dynamic $Value\ System$, inspired b
y the Theory of Needs. It incorporates an understanding of human-like desires, s
uch as the need for social interaction, personal fulfillment, and self-care. At

each step, the agent evaluates the value of its current state, proposes a set of candidate activities, and selects the one that best aligns with its intrinsic motivations. We conduct experiments on Concordia, a text-based simulator, to demonstrate that our agent generates coherent, contextually relevant daily activities while exhibiting variability and adaptability similar to human behavior. A comparative analysis with other LLM-based agents demonstrates that our approach significantly enhances the rationality of the simulated activities.
****
Zhixin Li,Yuheng Jia
ConMix: Contrastive Mixup at Representation Level for Long-tailed Deep Clustering
Deep clustering has made remarkable progress in recent years. However, most existing deep clustering methods assume that distributions of different clusters are balanced or roughly balanced, which are not consistent with the common long-tailed distributions in reality. In nature, the datasets often follow long-tailed distributions, leading to biased models being trained with significant performance drop. Despite the widespread proposal of many long-tailed learning approaches with supervision information, research on long-tailed deep clustering remains almost uncharted. Unaware of the data distribution and sample labels, long-tailed deep clustering is highly challenging. To tackle this problem, we propose a novel contrastive mixup method for long-tailed deep clustering, named ConMix. The proposed method makes innovations to mixup representations in contrastive learning to enhance deep clustering in long-tailed scenarios. Neural networks trained with ConMix can learn more discriminative representations, thus achieve better long-tailed deep clustering performance. We theoretically prove that ConMix works through re-balancing loss for classes with different long-tailed degree. We evaluate our method on widely used benchmark datasets with different imbalance ratios, suggesting it outperforms many state-of-the-art deep clustering approaches. The code is available at https://github.com/LZX-001/ConMix.
****
Shaofeng Zhang,Qiang Zhou,Sitong Wu,Haoru Tan,Zhibin Wang,Jinfa Huang,Junchi Yan
CR2PQ: Continuous Relative Rotary Positional Query for Dense Visual Representation Learning
Dense visual contrastive learning (DRL) shows promise for learning localized information in dense prediction tasks, but struggles with establishing pixel/patch correspondence across different views (cross-contrasting). Existing methods primarily rely on self-contrasting the same view with variations, limiting input variance and hindering downstream performance. This paper delves into the mechanisms of self-contrasting and cross-contrasting, identifying the crux of the issue: transforming discrete positional embeddings to continuous representations. To address the correspondence problem, we propose a Continuous Relative Rotary Positional Query ({\mname}), enabling patch-level representation learning. Our extensive experiments on standard datasets demonstrate state-of-the-art (SOTA) results. Compared to the previous SOTA method (PQCL), our approach achieves significant improvements on COCO: with 300 epochs of pretraining, {\mname} obtains \textbf{3.4\%} mAP^{bb} and \textbf{2.1\%} mAP^{mk} improvements for detection and segmentation tasks, respectively. Furthermore, {\mname} exhibits faster convergence, achieving \textbf{10.4\%} mAP^{bb} and \textbf{7.9\%} mAP^{mk} improvements over SOTA with just 40 epochs of pretraining.
****
Shuai Zhang,Junfeng Fang,Xuqiang Li,hongxin xiang,ALAN XIA,Ye Wei,Wenjie Du,Yang Wang
Iterative Substructure Extraction for Molecular Relational Learning with Interactive Graph Information Bottleneck
Molecular relational learning (MRL) seeks to understand the interactions between molecules, a pivotal task in domains such as drug discovery and materials science. Recently, extracting core substructures and modeling their interactions have emerged as mainstream approaches within machine learning-assisted methods. However, these methods still exhibit some limitations, such as insufficient consideration of molecular interactions or capturing substructures that include

excessive noise, which hampers precise core substructure extraction.
To address these challenges, we present an integrated dynamic framework called Iterative Substructure Extraction (ISE). ISE employs the Expectation-Maximization (EM) algorithm for MRL tasks, where the core substructures of interacting molecules are treated as latent variables and model parameters, respectively. Through iterative refinement, ISE gradually narrows the interactions from the entire molecular structures to just the core substructures.
Moreover, to ensure the extracted substructures are concise and compact, we propose the Interactive Graph Information Bottleneck (IGIB) theory, which focuses on capturing the most influential yet minimal interactive substructures. In summary, our approach, guided by the IGIB theory, achieves precise substructure extraction within the ISE framework and is encapsulated in the IGIB-ISE}
Extensive experiments validate the superiority of our model over state-of-the-art baselines across various tasks in terms of accuracy, generalizability, and interpretability.
****
Avik Pal,Max van Spengler,Guido Maria D'Amely di Melendugno,Alessandro Flaborea,Fabio Galasso,Pascal Mettes
Compositional Entailment Learning for Hyperbolic Vision-Language Models
Image-text representation learning forms a cornerstone in vision-language models, where pairs of images and textual descriptions are contrastively aligned in a shared embedding space. Since visual and textual concepts are naturally hierarchical, recent work has shown that hyperbolic space can serve as a high-potential manifold to learn vision-language representation with strong downstream performance. In this work, for the first time we show how to fully leverage the innate hierarchical nature of hyperbolic embeddings by looking beyond individual image-text pairs. We propose Compositional Entailment Learning for hyperbolic vision-language models. The idea is that an image is not only described by a sentence but is itself a composition of multiple object boxes, each with their own textual description. Such information can be obtained freely by extracting nouns from sentences and using openly available localized grounding models. We show how to hierarchically organize images, image boxes, and their textual descriptions through contrastive and entailment-based objectives. Empirical evaluation on a hyperbolic vision-language model trained with millions of image-text pairs shows that the proposed compositional learning approach outperforms conventional Euclidean CLIP learning, as well as recent hyperbolic alternatives, with better zero-shot and retrieval generalization and clearly stronger hierarchical performance.
****
Hengrui Zhang,Liancheng Fang,Qitian Wu,Philip S. Yu
DiffPuter: Empowering Diffusion Models for Missing Data Imputation
Generative models play an important role in missing data imputation in that they aim to learn the joint distribution of full data. However, applying advanced deep generative models (such as Diffusion models) to missing data imputation is challenging due to 1) the inherent incompleteness of the training data and 2) the difficulty in performing conditional inference from unconditional generative models. To deal with these challenges, this paper introduces DiffPuter, a tailored diffusion model combined with the Expectation-Maximization (EM) algorithm for missing data imputation. DiffPuter iteratively trains a diffusion model to learn the joint distribution of missing and observed data and performs an accurate conditional sampling to update the missing values using a tailored reversed sampling strategy. Our theoretical analysis shows that DiffPuter's training step corresponds to the maximum likelihood estimation of data density (M-step), and its sampling step represents the Expected A Posteriori estimation of missing values (E-step). Extensive experiments across ten diverse datasets and comparisons with 17 different imputation methods demonstrate DiffPuter's superior performance. Notably, DiffPuter achieves an average improvement of 8.10\% in MAE and 5.64\% in RMSE compared to the most competitive existing method.
****
Nicolas BEREUX,Aurélien Decelle,Cyril Furtlehner,Lorenzo Rosset,Beatriz Seoane
Fast training and sampling of Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBMs) are powerful tools for modeling complex sys
tems and extracting insights from data, but their training is hindered by the sl
ow mixing of Markov Chain Monte Carlo (MCMC) processes, especially with highly s
tructured datasets. In this study, we build on recent theoretical advances in RB
M training and focus on the stepwise encoding of data patterns into singular vec
tors of the coupling matrix, significantly reducing the cost of generating new s
amples and evaluating the quality of the model, as well as the training cost in
highly clustered datasets. The learning process is analogous to the thermodynam
ic continuous phase transitions observed in ferromagnetic models, where new mode
s in the probability measure emerge in a continuous manner.
We leverage the continuous transitions in the training process to define a smoot
h annealing trajectory that enables reliable and computationally efficient log-l
ikelihood estimates. This approach enables online assessment during training and
 introduces a novel sampling strategy called Parallel Trajectory Tempering (PTT)
 that outperforms previously optimized MCMC methods.
To mitigate the critical slowdown effect in the early stages of training, we pro
pose a pre-training phase. In this phase, the principal components are encoded i
nto a low-rank RBM through a convex optimization process, facilitating efficient
 static Monte Carlo sampling and accurate computation of the partition function.
Our results demonstrate that this pre-training strategy allows RBMs to efficient
ly handle highly structured datasets where conventional methods fail. Additional
ly, our log-likelihood estimation outperforms computationally intensive approach
es in controlled scenarios, while the PTT algorithm significantly accelerates MC
MC processes compared to conventional methods.
****
Bharat Srikishan,Daniel O'Malley,Mohamed Mehana,Nicholas Lubbers,Nikhil Muralidh
ar
Model-Agnostic Knowledge Guided Correction for Improved Neural Surrogate Rollout
Modeling the evolution of physical systems is critical to many applications in s
cience and engineering. As the evolution of these systems is governed by partial
 differential equations (PDEs), there are a number of computational simulations
which resolve these systems with high accuracy. However, as these simulations in
cur high computational costs, they are infeasible to be employed for large-scale
 analysis. A popular alternative to simulators are neural network surrogates whi
ch are trained in a data-driven manner and are much more computationally efficie
nt. However, these surrogate models suffer from high rollout error when used aut
oregressively, especially when confronted with training data paucity. Existing w
ork proposes to improve surrogate rollout error by either including physical los
s terms directly in the optimization of the model or incorporating computational
 simulators as `differentiable layers' in the neural network. Both of these appr
oaches have their challenges, with physical loss functions suffering from slow c
onvergence for stiff PDEs and simulator layers requiring gradients which are not
 always available, especially in legacy simulators. We propose the Hybrid PDE Pr
edictor with Reinforcement Learning (HyPER) model: a model-agnostic, RL based, c
ost-aware model which combines a neural surrogate, RL decision model, and a phys
ics simulator (with or without gradients) to reduce surrogate rollout error sign
ificantly. In addition to reducing in-distribution rollout error by **47%-78%**,
 HyPER learns an intelligent policy that is adaptable to changing physical condi
tions and resistant to noise corruption. Code available at https://github.com/sc
ailab/HyPER.
****
Weronika Ormaniec,Felix Dangel,Sidak Pal Singh
What Does It Mean to Be a Transformer? Insights from a Theoretical Hessian Analy
sis
The Transformer architecture has inarguably revolutionized deep learning, overta
king classical architectures like multi-layer perceptions (MLPs) and convolution
al neural networks (CNNs). At its core, the attention block differs in form and
functionality from most other architectural components in deep learning—to the e
xtent that, in comparison to MLPs/CNNs, Transformers are more often accompanied
by adaptive optimizers, layer normalization, learning rate warmup, etc. The root

causes behind these outward manifestations and the precise mechanisms that gove
rn them remain poorly understood. In this work, we bridge this gap by providing
a fundamental understanding of what distinguishes the Transformer from the other
architectures—grounded in a theoretical comparison of the (loss) Hessian. Concr
etely, for a single self-attention layer, (a) we first entirely derive the Trans
former’s Hessian and express it in matrix derivatives; (b) we then characterize
it in terms of data, weight, and attention moment dependencies; and (c) while do
ing so further highlight the important structural differences to the Hessian of
classical networks. Our results suggest that various common architectural and op
timization choices in Transformers can be traced back to their highly non-linear
dependencies on the data and weight matrices, which vary heterogeneously across
parameters. Ultimately, our findings provide a deeper understanding of the Tran
sformer’s unique optimization landscape and the challenges it poses.
****
Jin Peng Zhou,Christian K Belardi,Ruihan Wu,Travis Zhang,Carla P Gomes,Wen Sun,K
ilian Q Weinberger
On Speeding Up Language Model Evaluation
Developing prompt-based methods with Large Language Models (LLMs) requires makin
g numerous decisions, which give rise to a combinatorial search problem over hyp
er-parameters. This exhaustive evaluation can be time-consuming and costly. In t
his paper, we propose an \textit{adaptive} approach to explore this space. We ar
e exploiting the fact that often only few samples are needed to identify clearly
superior or inferior settings, and that many evaluation tests are highly correl
ated. We lean on multi-armed bandits to sequentially identify the next (method,
validation sample)-pair to evaluate and utilize low-rank matrix factorization to
fill in missing evaluations. We carefully assess the efficacy of our approach o
n several competitive benchmark problems and show that it can identify the top-p
erforming method using only 5-15% of the typical resources---resulting in 85-95%
LLM cost savings. Our code is available at https://github.com/kilian-group/band
iteval.
****
Zhiliang Chen,Xinyuan Niu,Chuan-Sheng Foo,Bryan Kian Hsiang Low
Broaden your SCOPE! Efficient Multi-turn Conversation Planning for LLMs with Sem
antic Space
Large language models (LLMs) are used in chatbots or AI assistants to hold conve
rsations with a human user. In such applications, the quality (e.g., user engage
ment, safety) of a conversation is important and can only be exactly known at th
e end of the conversation. To maximize its expected quality, conversation planni
ng reasons about the stochastic transitions within a conversation to select the
optimal LLM response at each turn. Existing simulation-based conversation planni
ng algorithms typically select the optimal response by simulating future convers
ations with a large number of LLM queries at every turn. However, this process i
s extremely time-consuming and hence impractical for real-time conversations. Th
is paper presents a novel approach called Semantic space COnversation Planning w
ith improved Efficiency (SCOPE) that exploits the dense semantic representation
of conversations to perform conversation planning efficiently. In particular, SC
OPE models the stochastic transitions in conversation semantics and their associ
ated rewards to plan entirely within the semantic space. This allows us to selec
t the optimal LLM response at every conversation turn without needing additional
LLM queries for simulation. As a result, SCOPE can perform conversation plannin
g 70 times faster than conventional simulation-based planning algorithms when ap
plied to a wide variety of conversation starters and two reward functions seen i
n the real world, yet achieving a higher reward within a practical planning budg
et. Our code can be found at: https://github.com/chenzhiliang94/convo-plan-SCOPE
.
****
James Burgess,Xiaohan Wang,Yuhui Zhang,Anita Rau,Alejandro Lozano,Lisa Dunlap,Tr
evor Darrell,Serena Yeung-Levy
Video Action Differencing
How do two individuals differ when performing the same action? In this work, we

introduce Video Action Differencing (VidDiff), the novel task of identifying subtle differences between videos of the same action, which has numerous applications, such as coaching and skill learning. To enable development on this new task, we first create VidDiffBench, a benchmark dataset containing 549 video pairs, with human annotations of 4,469 fine-grained action differences and 2,075 timestamps indicating where these differences occur. Our experiments demonstrate that VidDiffBench poses a significant challenge for state-of-the-art large multimodal models (LMMs), such as GPT-4o and Qwen2-VL. By analyzing the failure cases of LMMs on VidDiffBench, we highlight two key challenges for this task: localizing relevant sub-actions over two videos and fine-grained frame comparison. To overcome these, we propose the VidDiff method, an agentic workflow that breaks the task into three stages: action difference proposal, keyframe localization, and frame differencing, each stage utilizing specialized foundation models. To encourage future research in this new task, we release the benchmark and code.
****
Xianwei Zhuang,Zhihong Zhu,Zhichang Wang,Xuxin Cheng,Yuexian Zou
UniCoTT: A Unified Framework for Structural Chain-of-Thought Distillation
Chains of thought (CoTs) have achieved success in enhancing the reasoning capabilities of large language models (LLMs), while their effectiveness is predominantly observed in LLMs. Existing solutions methods adopt distillation to inject chain-of-thought capabilities into small models (SLMs). However, they: (1) can not guarantee the rationality of the generated explanation due to hallucinations; (2) ignore diverse structures of CoT during knowledge transfer. In this paper, we propose a unified CoT distillation framework termed UniCoTT for considering diverse structural CoTs (\emph{i.e.}, chain, tree, and graph). UniCoTT contains two core strategies: iterative construction for structured CoTs and the structural constraint strategy. Specifically, UniCoTT prompts LLMs to iteratively produce accurate explanations with answers and unifies structured explanations as UniCoT which is seen as a bridge for knowledge transfer. Furthermore, UniCoTT utilizes the proposed unified supervised learning and structural consistency learning strategies to transfer knowledge of structured CoT to SLMs. Experimental results show that UniCoTT can significantly improve the performance of SLMs on multiple datasets across different NLP tasks. Our code is available at https://github.com/mengchuang123/UniCoTT.
****
Hantao Zhang,Yuhe Liu,Jiancheng Yang,Shouhong Wan,Xinyuan Wang,Wei Peng,Pascal Fua
LeFusion: Controllable Pathology Synthesis via Lesion-Focused Diffusion Models
Patient data from real-world clinical practice often suffers from data scarcity and long-tail imbalances, leading to biased outcomes or algorithmic unfairness. This study addresses these challenges by generating lesion-containing image-segmentation pairs from lesion-free images. Previous efforts in medical imaging synthesis have struggled with separating lesion information from background, resulting in low-quality backgrounds and limited control over the synthetic output. Inspired by diffusion-based image inpainting, we propose LeFusion, a lesion-focused diffusion model. By redesigning the diffusion learning objectives to focus on lesion areas, we simplify the learning process and improve control over the output while preserving high-fidelity backgrounds by integrating forward-diffused background contexts into the reverse diffusion process. Additionally, we tackle two major challenges in lesion texture synthesis: 1) multi-peak and 2) multi-class lesions. We introduce two effective strategies: histogram-based texture control and multi-channel decomposition, enabling the controlled generation of high-quality lesions in difficult scenarios. Furthermore, we incorporate lesion mask diffusion, allowing control over lesion size, location, and boundary, thus increasing lesion diversity. Validated on 3D cardiac lesion MRI and lung nodule CT datasets, LeFusion-generated data significantly improves the performance of state-of-the-art segmentation models, including nnUNet and SwinUNETR.
****
Yihao Hu,Congyu Qiao,Xin Geng,Ning Xu
Selective Label Enhancement Learning for Test-Time Adaptation

Test-time adaptation (TTA) aims to adapt a pre-trained model to the target domain using only unlabeled test samples. Most existing TTA approaches rely on definite pseudo-labels, inevitably introducing false labels and failing to capture uncertainty for each test sample. This prevents pseudo-labels from being flexibly refined as the model adapts during training, limiting their potential for performance improvement. To address this, we propose the Progressive Adaptation with Selective Label Enhancement (PASLE) framework. Instead of definite labels, PASLE assigns candidate pseudo-label sets to uncertain ones via selective label enhancement. Specifically, PASLE partitions data into confident/uncertain subsets, assigning one-hot labels to confident samples and candidate sets to uncertain ones. The model progressively trains on certain/uncertain pseudo-labeled data while dynamically refining uncertain pseudo-labels, leveraging increasing target adaptation monitored throughout training. Experiments on various benchmark datasets validate the effectiveness of the proposed approach.
****
Yifei Sun,Zemin Liu,Bryan Hooi,Yang Yang,Rizal Fathony,Jia Chen,Bingsheng He
Multi-Label Node Classification with Label Influence Propagation
Graphs are a complex and versatile data structure used across various domains, with possibly multi-label nodes playing a particularly crucial role.
Examples include proteins in PPI networks with multiple functions and users in social or e-commerce networks exhibiting diverse interests.
Tackling multi-label node classification (MLNC) on graphs has led to the development of various approaches. Some methods leverage graph neural networks (GNNs) to exploit label co-occurrence correlations, while others incorporate label embeddings to capture label proximity. However, these approaches fail to account for the intricate influences between labels in non-Euclidean graph data.
To address this issue, we decompose the message passing process in GNNs into two operations: propagation and transformation.
We then conduct a comprehensive analysis and quantification of the influence correlations between labels in each operation.
Building on these insights, we propose a novel model, Label Influence Propagation (LIP).
Specifically, we construct a label influence graph based on the integrated label correlations.
Then, we propagate high-order influences through this graph, dynamically adjusting the learning process by amplifying labels with positive contributions and mitigating those with negative influence.
Finally, our framework is evaluated on comprehensive benchmark datasets, consistently outperforming SOTA methods across various settings, demonstrating its effectiveness on MLNC tasks.
****
Shoubin Yu,Jaehong Yoon,Mohit Bansal
CREMA: Generalizable and Efficient Video-Language Reasoning via Multimodal Modular Fusion
Despite impressive advancements in recent multimodal reasoning approaches, they are still limited in flexibility and efficiency, as these models typically process only a few fixed modality inputs and require updates to numerous parameters.
This paper tackles these critical challenges and proposes CREMA, a generalizable, highly efficient, and modular modality-fusion framework that can incorporate many new modalities to enhance video reasoning. We first augment multiple informative modalities (such as optical flow, 3D point cloud, audio, thermal heatmap, and touch map) from given videos without extra human annotation by leveraging sensors or existing pre-trained models.
Next, we introduce a query transformer with multiple parameter-efficient modules associated with each accessible modality.
It projects diverse modality features to the LLM token embedding space, allowing the model to integrate different data types for response generation. Furthermore, we propose a novel progressive multimodal fusion design supported by a lightweight fusion module and modality-sequential training strategy. It helps compress information across various assisting modalities, maintaining computational effi

ciency in the LLM while improving performance. We validate our method on seven v
ideo-language reasoning tasks assisted by diverse modalities, including conventi
onal VideoQA and Video-Audio/3D/Touch/Thermal QA, and achieve better/equivalent
performance against strong multimodal LLMs, including OneLLM, BLIP-2, and SeViLA
 while reducing over 90% trainable parameters. We provide extensive analyses of
CREMA, including the impact of each modality on reasoning domains, the design of
 the fusion module, and example visualizations.
****
Hongjin SU,Ruoxi Sun,Jinsung Yoon,Pengcheng Yin,Tao Yu,Sercan O Arik
Learn-by-interact: A Data-Centric Framework For Self-Adaptive Agents in Realisti
c Environments
Autonomous agents powered by large language models (LLMs) have the potential to
enhance human capabilities, assisting with digital tasks from sending emails to
performing data analysis. The abilities of existing LLMs at such tasks are oft
en hindered by the lack of high-quality agent data from the corresponding enviro
nments they interact with. We propose LEARN-BY-INTERACT, a data-centric framewo
rk to adapt LLM agents to any given environments without human annotations. LE
ARN-BY-INTERACT synthesizes trajectories of agent-environment interactions based
 on documentations, and constructs instructions by summarizing or abstracting th
e interaction histories, a process called backward construction. We assess the q
uality of our synthetic data by using them in both training-based scenarios and
training-free in-context learning (ICL), where we craft innovative retrieval app
roaches optimized for agents. Extensive experiments on SWE-bench, WebArena, OSWo
rld, and Spider2-V spanning across realistic coding, web, and desktop environmen
ts show the effectiveness of LEARN-BY-INTERACT in various downstream agentic tas
ks — baseline results are improved up to 11.1% for ICL with Claude-3.5 and 23.1%
 for training with Codestral-22B. We further demonstrate the critical role of ba
ckward construction, which provides up to 10.6% improvement for training. Our a
blation studies demonstrate the efficiency provided by our synthesized data in I
CL and the superiority of our retrieval pipeline over alternative approaches lik
e conventional retrieval-augmented generation (RAG). We expect that LEARN-BY-INT
ERACT will serve as a foundation for agent data synthesis as LLMs are increasing
ly deployed at real-world environments.
****
Wanpeng Zhang,Zilong Xie,Yicheng Feng,Yijiang Li,Xingrun Xing,Sipeng Zheng,Zongq
ing Lu
From Pixels to Tokens: Byte-Pair Encoding on Quantized Visual Modalities
Multimodal Large Language Models have made significant strides in integrating vi
sual and textual information, yet they often struggle with effectively aligning
these modalities. We introduce a novel image tokenizer that bridges this gap by
applying the principle of Byte-Pair Encoding (BPE) to visual data. Unlike conven
tional approaches that rely on separate visual encoders, our method directly inc
orporates structural prior information into image tokens, mirroring the successf
ul tokenization strategies used in text-only Large Language Models. This innovat
ive approach enables Transformer models to more effectively learn and reason acr
oss modalities. Through theoretical analysis and extensive experiments, we demon
strate that our BPE Image Tokenizer significantly enhances MLLMs' multimodal und
erstanding capabilities, even with limited training data. Leveraging this method
, we develop Being-VL-0, a model that demonstrates superior performance across v
arious benchmarks and shows promising scalability, potentially paving the way fo
r more efficient and capable multimodal foundation models. For further details,
visit our website https://github.com/BeingBeyond/Being-VL-0.
****
Leonardo Barcellona,Andrii Zadaianchuk,Davide Allegro,Samuele Papa,Stefano Ghido
ni,Efstratios Gavves
Dream to Manipulate: Compositional World Models Empowering Robot Imitation Learn
ing with Imagination
A world model provides an agent with a representation of its environment, enabli
ng it to predict the causal consequences of its actions. Current world models ty
pically cannot directly and explicitly imitate the actual environment in front o

f a robot, often resulting in unrealistic behaviors and hallucinations that make them unsuitable for real-world robotics applications.
To overcome those challenges, we propose to rethink robot world models as learnable digital twins. We introduce DreMa, a new approach for constructing digital twins automatically using learned explicit representations of the real world and its dynamics, bridging the gap between traditional digital twins and world models.
DreMa replicates the observed world and its structure by integrating Gaussian Splatting and physics simulators, allowing robots to imagine novel configurations of objects and to predict the future consequences of robot actions thanks to its compositionality.
We leverage this capability to generate new data for imitation learning by applying equivariant transformations to a small set of demonstrations. Our evaluations across various settings demonstrate significant improvements in accuracy and robustness by incrementing actions and object distributions, reducing the data needed to learn a policy and improving the generalization of the agents.
As a highlight, we show that a real Franka Emika Panda robot, powered by DreMa’s imagination, can successfully learn novel physical tasks from just a single example per task variation (one-shot policy learning).
Our project page can be found in: https://dreamtomanipulate.github.io/.
****

Kaiwen Zhu,Jinjin Gu,Zhiyuan You,Yu Qiao,Chao Dong
An Intelligent Agentic System for Complex Image Restoration Problems
Real-world image restoration (IR) is inherently complex and often requires combining multiple specialized models to address diverse degradations. Inspired by human problem-solving, we propose AgenticIR, an agentic system that mimics the human approach to image processing by following five key stages: Perception, Scheduling, Execution, Reflection, and Rescheduling. AgenticIR leverages large language models (LLMs) and vision-language models (VLMs) that interact via text generation to dynamically operate a toolbox of IR models. We fine-tune VLMs for image quality analysis and employ LLMs for reasoning, guiding the system step by step.
To compensate for LLMs' lack of specific IR knowledge and experience, we introduce a self-exploration method, allowing the LLM to observe and summarize restoration results into referenceable documents. Experiments demonstrate AgenticIR's potential in handling complex IR tasks, representing a promising path toward achieving general intelligence in visual processing.
****

Bohan Zeng,Shanglin Li,Yutang Feng,Ling Yang,Juan Zhang,Hong Li,Jiaming Liu,Conghui He,Wentao Zhang,Jianzhuang Liu,Baochang Zhang,Shuicheng YAN
IPDreamer: Appearance-Controllable 3D Object Generation with Complex Image Prompts
Recent advances in 3D generation have been remarkable, with methods such as DreamFusion leveraging large-scale text-to-image diffusion-based models to guide 3D object generation. These methods enable the synthesis of detailed and photorealistic textured objects. However, the appearance of 3D objects produced by such text-to-3D models is often unpredictable, and it is hard for single-image-to-3D methods to deal with images lacking a clear subject, complicating the generation of appearance-controllable 3D objects from complex images. To address these challenges, we present IPDreamer, a novel method that captures intricate appearance features from complex **I**mage **P**rompts and aligns the synthesized 3D object with these extracted features, enabling high-fidelity, appearance-controllable 3D object generation. Our experiments demonstrate that IPDreamer consistently generates high-quality 3D objects that align with both the textual and complex image prompts, highlighting its promising capability in appearance-controlled, complex 3D object generation.
****

Sreyan Ghosh,Chandra Kiran Reddy Evuru,Sonal Kumar,Utkarsh Tyagi,Oriol Nieto,Zeyu Jin,Dinesh Manocha
Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs

Large Vision-Language Models (LVLMs) often produce responses that misalign with factual information, a phenomenon known as hallucinations. While hallucinations are well-studied, the exact causes behind them remain underexplored. In this paper, we first investigate the root causes of hallucinations in LVLMs. Our findings reveal that existing mitigation techniques primarily reduce hallucinations for visual recognition prompts—those that require simple descriptions of visual elements—but fail for cognitive prompts that demand deliberate reasoning. We identify the core issue as a lack of true visual perception in LVLMs: although they can accurately recognize visual elements, they struggle to fully interpret these elements in the context of the input prompt and effectively link this recognition to their internal knowledge, which is critical for reasoning. To address this gap, we introduce Visual Description Grounded Decoding (VDGD), a simple, robust, and training-free method designed to enhance visual perception and improve reasoning capabilities in LVLMs. VDGD works by first generating a detailed description of the image and appending it as a prefix to the instruction. During response generation, tokens are sampled based on their KL divergence to the description, favoring candidates with lower divergence. Experimental results on multiple visual reasoning benchmarks and LVLMs demonstrate that VDGD consistently outperforms existing baselines 2% - 33%. Finally, we introduce VaLLu, a benchmark designed for comprehensive evaluation of the cognitive capabilities of LVLMs.
****
Millicent Li,Tongfei Chen,Benjamin Van Durme,Patrick Xia
Multi-Field Adaptive Retrieval
Document retrieval for tasks such as search and retrieval-augmented generation typically involves datasets that are _unstructured_: free-form text without explicit internal structure in each document. However, documents can have some structure, containing fields such as an article title, a message body, or an HTML header. To address this gap, we introduce Multi-Field Adaptive Retrieval (mFAR), a flexible framework that accommodates any number and any type of document indices on _semi-structured_ data. Our framework consists of two main steps: (1) the decomposition of an existing document into fields, each indexed independently through dense and lexical methods, and (2) learning a model which adaptively predicts the importance of a field by conditioning on the document query, allowing on-the-fly weighting of the most likely field(s). We find that our approach allows for the optimized use of dense versus lexical representations across field types, significantly improves in document ranking over a number of existing retrievers, and achieves state-of-the-art performance for multi-field structured data.
****
Hritik Bansal,Arian Hosseini,Rishabh Agarwal,Vinh Q. Tran,Mehran Kazemi
Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling
Training on high-quality synthetic data from strong language models (LMs) is a common strategy to improve the reasoning performance of LMs. In this work, we revisit whether this strategy is compute-optimal under a fixed inference budget (e.g., FLOPs). To do so, we investigate the trade-offs between generating synthetic data using a stronger but more expensive (SE) model versus a weaker but cheaper (WC) model. We evaluate the generated data across three key metrics: coverage, diversity, and false positive rate, and show that the data from WC models may have higher coverage and diversity, but also exhibit higher false positive rates. We then finetune LMs on data from SE and WC models in different settings: knowledge distillation, self-improvement, and a novel weak-to-strong improvement setup where a weaker LM teaches reasoning to a stronger LM. Our findings reveal that models finetuned on WC-generated data consistently outperform those trained on SE-generated data across multiple benchmarks and multiple choices of WC and SE models. These results challenge the prevailing practice of relying on SE models for synthetic data generation, suggesting that WC may be the compute-optimal approach for training advanced LM reasoners.
****
Tianyuan Zhang,Zhengfei Kuang,Haian Jin,Zexiang Xu,Sai Bi,Hao Tan,He Zhang,Yiwei Hu,Milos Hasan,William T. Freeman,Kai Zhang,Fujun Luan
RelitLRM: Generative Relightable Radiance for Large Reconstruction Models

We propose RelitLRM, a Large Reconstruction Model (LRM) for generating high-quality Gaussian splatting representations of 3D objects under novel illuminations from sparse (4-8) posed images captured under unknown static lighting. Unlike prior inverse rendering methods requiring dense captures and slow optimization, often causing artifacts like incorrect highlights or shadow baking, RelitLRM adopts a feed-forward transformer-based model with a novel combination of a geometry reconstructor and a relightable appearance generator based on diffusion. The model is trained end-to-end on synthetic multi-view renderings of objects under varying known illuminations. This architecture design enables to effectively decompose geometry and appearance, resolve the ambiguity between material and lighting, and capture the multi-modal distribution of shadows and specularity in the relit appearance. We show our sparse-view feed-forward RelitLRM offers competitive relighting results to state-of-the-art dense-view optimization-based baselines while being significantly faster. Our project page is available at: https://relit-lrm.github.io/.
****

Alessandro Palma,Till Richter,Hanyi Zhang,Manuel Lubetzki,Alexander Tong,Andrea Dittadi,Fabian J Theis

Multi-Modal and Multi-Attribute Generation of Single Cells with CFGen

Generative modeling of single-cell RNA-seq data is crucial for tasks like trajectory inference, batch effect removal, and simulation of realistic cellular data. However, recent deep generative models simulating synthetic single cells from noise operate on pre-processed continuous gene expression approximations, overlooking the discrete nature of single-cell data, which limits their effectiveness and hinders the incorporation of robust noise models. Additionally, aspects like controllable multi-modal and multi-label generation of cellular data remain underexplored. This work introduces CellFlow for Generation (CFGen), a flow-based conditional generative model that preserves the inherent discreteness of single-cell data. CFGen generates whole-genome multi-modal single-cell data reliably, improving the recovery of crucial biological data characteristics while tackling relevant generative tasks such as rare cell type augmentation and batch correction. We also introduce a novel framework for compositional data generation using Flow Matching. By showcasing CFGen on a diverse set of biological datasets and settings, we provide evidence of its value to the fields of computational biology and deep generative models.
****

Saeed Vahidian,Mingyu Wang,Jianyang Gu,Vyacheslav Kungurtsev,Wei Jiang,Yiran Chen

Group Distributionally Robust Dataset Distillation with Risk Minimization

Dataset distillation (DD) has emerged as a widely adopted technique for crafting a synthetic dataset that captures the essential information of a training dataset, facilitating the training of accurate neural models. Its applications span various domains, including transfer learning, federated learning, and neural architecture search. The most popular methods for constructing the synthetic data rely on matching the convergence properties of training the model with the synthetic dataset and the training dataset. However, using the empirical loss as the criterion must be thought of as auxiliary in the same sense that the training set is an approximate substitute for the population distribution, and the latter is the data of interest. Yet despite its popularity, an aspect that remains unexplored is the relationship of DD to its generalization, particularly across uncommon subgroups. That is, how can we ensure that a model trained on the synthetic dataset performs well when faced with samples from regions with low population density? Here, the representativeness and coverage of the dataset become salient over the guaranteed training error at inference. Drawing inspiration from distributionally robust optimization, we introduce an algorithm that combines clustering with the minimization of a risk measure on the loss to conduct DD. We provide a theoretical rationale for our approach and demonstrate its effective generalization and robustness across subgroups through numerical experiments.
****

Po-Wei Huang,Pei-Chiun Peng,Hung Guei,Ti-Rong Wu

OptionZero: Planning with Learned Options
Planning with options -- a sequence of primitive actions -- has been shown effec
tive in reinforcement learning within complex environments. Previous studies hav
e focused on planning with predefined options or learned options through expert
demonstration data.
Inspired by MuZero, which learns superhuman heuristics without any human knowled
ge, we propose a novel approach, named *OptionZero*. OptionZero incorporates an
option network into MuZero, providing autonomous discovery of options through
self-play games. Furthermore, we modify the dynamics network to provide environm
ent transitions when using options, allowing searching deeper under the same sim
ulation constraints. Empirical experiments conducted in 26 Atari games demonstra
te that OptionZero outperforms MuZero, achieving a 131.58% improvement in mean h
uman-normalized score. Our behavior analysis shows that OptionZero not only lear
ns options but also acquires strategic skills tailored to different game charact
eristics. Our findings show promising directions for discovering and using optio
ns in planning. Our code is available at https://rlg.iis.sinica.edu.tw/papers/op
tionzero.
****
Shengda Fan,Xin Cong,Yuepeng Fu,Zhong Zhang,Shuyan Zhang,Yuanwei Liu,Yesai Wu,Ya
nkai Lin,Zhiyuan Liu,Maosong Sun
WorkflowLLM: Enhancing Workflow Orchestration Capability of Large Language Model
s
Recent advancements in large language models (LLMs) have driven a revolutionary
paradigm shift in process automation from Robotic Process Automation to Agentic
Process Automation by automating the workflow orchestration procedure based on L
LMs. However, existing LLMs (even the advanced OpenAI GPT-4o) are confined to ac
hieving satisfactory capability in workflow orchestration. To address this limit
ation, we present WorkflowLLM, a data-centric framework elaborately designed to
enhance the capability of LLMs in workflow orchestration. It first constructs a
large-scale fine-tuning dataset WorkflowBench with 106, 763 samples, covering 1,
 503 APIs from 83 applications across 28 categories. Specifically, the construct
ion process can be divided into three phases: (1) Data Collection: we collect re
al-world workflow data from Apple Shortcuts and RoutineHub, transcribing them in
to Python-style code. We further equip them with generated hierarchical thought
via GPT-4o-mini. (2) Query Expansion: we prompt GPT-4o-mini to generate more tas
k queries to enrich the diversity and complexity of workflows. (3) Workflow Gene
ration: we leverage an annotator model trained on collected data to generate wor
kflows for synthesized queries. Finally, we merge the synthetic samples that pas
s quality confirmation with the collected samples to obtain the WorkflowBench. B
ased on WorkflowBench, we fine-tune Llama-3.1-8B to obtain WorkflowLlama. Our ex
periments show that WorkflowLlama demonstrates a strong capacity to orchestrate
complex workflows, while also achieving notable generalization performance on pr
eviously unseen APIs. Additionally, WorkflowBench exhibits robust zero-shot gene
ralization capabilities on an out-of-distribution task planning dataset, T-Eval.
 Our data and code are available at https://github.com/OpenBMB/WorkflowLLM.
****
Dongjun Lee,Juyong Lee,Kyuyoung Kim,Jihoon Tack,Jinwoo Shin,Yee Whye Teh,Kimin L
ee
Learning to Contextualize Web Pages for Enhanced Decision Making by LLM Agents
Recent advances in large language models (LLMs) have led to a growing interest i
n developing LLM-based agents for automating web tasks. However, these agents of
ten struggle with even simple tasks on real-world websites due to their limited
capability to understand and process complex web page structures. In this work,
we introduce LCoW, a framework for Learning language models to Contextualize com
plex Web pages into a more comprehensible form, thereby enhancing decision makin
g by LLM agents. LCoW decouples web page understanding from decision making by t
raining a separate contextualization module to transform complex web pages into
comprehensible format, which are then utilized by the decision-making agent. We
demonstrate that our contextualization module effectively integrates with LLM ag
ents of various scales to significantly enhance their decision-making capabiliti

es in web automation tasks. Notably, LCoW improves the success rates of closed-s
ource LLMs (e.g., Gemini-1.5-flash, GPT-4o, Claude-3.5-Sonnet) by an average of
15.6%, and demonstrates a 23.7% average improvement in success rates for open-so
urce LMs (e.g., Llama-3.1-8B, Llama-3.1-70B) on the WorkArena benchmark.
Moreover, the Gemini-1.5-flash agent with LCoW achieves state-of-the-art results
on the WebShop benchmark, outperforming human experts. The relevant code materi
als are available at our project page: https://lcowiclr2025.github.io.
****
Yifeng Xu,Zhenliang He,Shiguang Shan,Xilin Chen
CtrLoRA: An Extensible and Efficient Framework for Controllable Image Generation
Recently, large-scale diffusion models have made impressive progress in text-to-
image (T2I) generation. To further equip these T2I models with fine-grained spat
ial control, approaches like ControlNet introduce an extra network that learns t
o follow a condition image. However, for every single condition type, ControlNet
requires independent training on millions of data pairs with hundreds of GPU ho
urs, which is quite expensive and makes it challenging for ordinary users to exp
lore and develop new types of conditions. To address this problem, we propose th
e CtrLoRA framework, which trains a Base ControlNet to learn the common knowledg
e of image-to-image generation from multiple base conditions, along with conditi
on-specific LoRAs to capture distinct characteristics of each condition. Utilizi
ng our pretrained Base ControlNet, users can easily adapt it to new conditions,
requiring as few as 1,000 data pairs and less than one hour of single-GPU traini
ng to obtain satisfactory results in most scenarios. Moreover, our CtrLoRA reduc
es the learnable parameters by 90% compared to ControlNet, significantly lowerin
g the threshold to distribute and deploy the model weights. Extensive experiment
s on various types of conditions demonstrate the efficiency and effectiveness of
our method. Codes and model weights will be released at
https://github.com/xyfJASON/ctrlora.
****
Jiayi Ye,Yanbo Wang,Yue Huang,Dongping Chen,Qihui Zhang,Nuno Moniz,Tian Gao,Wern
er Geyer,Chao Huang,Pin-Yu Chen,Nitesh V Chawla,Xiangliang Zhang
Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge
LLM-as-a-Judge has been widely utilized as an evaluation method in various bench
marks and served as supervised rewards in model training. However, despite their
excellence in many domains, potential issues are under-explored, undermining th
eir reliability and the scope of their utility.
Therefore, we identify 12 key potential biases and propose a new automated bias
quantification framework—CALM—which systematically quantifies and analyzes each
type of bias in LLM-as-a-Judge by using automated and principle-guided modificat
ion. Our experiments cover multiple popular language models, and the results ind
icate that while advanced models have achieved commendable overall performance,
significant biases persist in certain specific tasks. Empirical results suggest
that there remains room for improvement in the reliability of LLM-as-a-Judge. Mo
reover, we also discuss the explicit and implicit influence of these biases and
give some suggestions for the reliable application of LLM-as-a-Judge. Our work h
ighlights the need for stakeholders to address these issues and remind users to
exercise caution in LLM-as-a-Judge applications.
****
Jesus de la Fuente Cedeño,Robert Lehmann,Carlos Ruiz-Arenas,Jan Voges,Irene Marí
n-Goñi,Xabier Martinez de Morentin,David Gomez-Cabrero,Idoia Ochoa,Jesper Tegnér
,Vincenzo Lagani,Mikel Hernaez
Interpretable Causal Representation Learning for Biological Data in the Pathway
Space
Predicting the impact of genomic and drug perturbations in cellular function is
crucial for understanding gene functions and drug effects, ultimately leading to
improved therapies. To this end, Causal Representation Learning (CRL) constitut
es one of the most promising approaches, as it aims to identify the latent facto
rs that causally govern biological systems, thus facilitating the prediction of
the effect of unseen perturbations. Yet, current CRL methods fail in reconciling
their principled latent representations with known biological processes, leadin

g to models that are not interpretable. To address this major issue, in this wor
k we present SENA-discrepancy-VAE, a model based on the recently proposed CRL me
thod discrepancy-VAE, that produces representations where each latent factor can
 be interpreted as the (linear) combination of the activity of a (learned) set o
f biological processes. To this extent, we present an encoder, SENA-δ, th
at efficiently compute and map biological processes' activity levels to the late
nt causal factors. We show that SENA-discrepancy-VAE achieves predictive perform
ances on unseen combinations of interventions that are comparable with its origi
nal, non-interpretable counterpart, while inferring causal latent factors that a
re biologically meaningful.
****
USVSN Sai Prashanth,Alvin Deng,Kyle O'Brien,Jyothir S V,Mohammad Aflah Khan,Jayd
eep Borkar,Christopher A. Choquette-Choo,Jacob Ray Fuehne,Stella Biderman,Tracy
Ke,Katherine Lee,Naomi Saphra
Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon
Memorization in language models is typically treated as a homogenous phenomenon,
 neglecting the specifics of the memorized data. We instead model memorization a
s the effect of a set of complex factors that describe each sample and relate it
 to the model and corpus. To build intuition around these factors, we break memo
rization down into a taxonomy: recitation of highly duplicated sequences, recons
truction of inherently predictable sequences, and recollection of sequences that
 are neither. We demonstrate the usefulness of our taxonomy by using it to const
ruct a predictive model for memorization. By analyzing dependencies and inspecti
ng the weights of the predictive model, we find that different factors have diff
erent influences on the likelihood of memorization depending on the taxonomic ca
tegory.
****
Alireza Ganjdanesh,Reza Shirkavand,Shangqian Gao,Heng Huang
Not All Prompts Are Made Equal: Prompt-based Pruning of Text-to-Image Diffusion
Models
Text-to-image (T2I) diffusion models have demonstrated impressive image generati
on capabilities. Still, their computational intensity prohibits resource-constra
ined organizations from deploying T2I models after fine-tuning them on their int
ernal *target* data. While pruning techniques offer a potential solution to redu
ce the computational burden of T2I models, static pruning methods use the same p
runed model for all input prompts, overlooking the varying capacity requirements
 of different prompts. Dynamic pruning addresses this issue by utilizing a separ
ate sub-network for each prompt, but it prevents batch parallelism on GPUs. To o
vercome these limitations, we introduce Adaptive Prompt-Tailored Pruning (APTP),
 a novel prompt-based pruning method designed for T2I diffusion models. Central
to our approach is a *prompt router* model, which learns to determine the requir
ed capacity for an input text prompt and routes it to an architecture code, give
n a total desired compute budget for prompts. Each architecture code represents
a specialized model tailored to the prompts assigned to it, and the number of co
des is a hyperparameter. We train the prompt router and architecture codes using
 contrastive learning, ensuring that similar prompts are mapped to nearby codes.
 Further, we employ optimal transport to prevent the codes from collapsing into
a single one. We demonstrate APTP's effectiveness by pruning Stable Diffusion (S
D) V2.1 using CC3M and COCO as *target* datasets. APTP outperforms the single-mo
del pruning baselines in terms of FID, CLIP, and CMMD scores. Our analysis of th
e clusters learned by APTP reveals they are semantically meaningful. We also sho
w that APTP can automatically discover previously empirically found challenging
prompts for SD, *e.g.,* prompts for generating text images, assigning them to hi
gher capacity codes.
****
Yuhao Wu,Ming Shan Hee,Zhiqiang Hu,Roy Ka-Wei Lee
LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs
Current benchmarks like ``$\textit{Needle-in-a-Haystack}$'' ($\textit{NIAH}$), $
\textit{Ruler}$, and $\textit{Needlebench}$ focus on models' ability to understa
nd long-context input sequences but fail to capture a critical dimension: the ge

neration of high-quality long-form text. Applications such as design proposals, technical documentation, and creative writing rely on coherent, instruction-following outputs over extended sequences—a challenge that existing benchmarks do not adequately address. To fill this gap, we introduce $\textit{LongGenBench}$, a novel benchmark designed to rigorously evaluate large language models' (LLMs) ability to generate long text while adhering to complex instructions. Through tasks requiring specific events or constraints within generated text, $\textit{LongGenBench}$ evaluates model performance across four distinct scenarios, three instruction types, and two generation-lengths (16K and 32K tokens). Our evaluation of ten state-of-the-art LLMs reveals that, despite strong results on $\textit{Ruler}$, all models struggled with long text generation on $\textit{LongGenBench}$, particularly as text length increased. This suggests that current LLMs are not yet equipped to meet the demands of real-world, long-form text generation. We open-source $\textit{LongGenBench}$ to promote comprehensive evaluation and improvement in this critical area, with code and data available at ${anonymousurl}$.
****
Congpei Qiu,Yanhao Wu,Wei Ke,Xiuxiu Bai,Tong Zhang
Refining CLIP's Spatial Awareness: A Visual-Centric Perspective
Contrastive Language-Image Pre-training (CLIP) excels in global alignment with language but exhibits limited sensitivity to spatial information, leading to strong performance in zero-shot classification tasks but underperformance in tasks requiring precise spatial understanding. Recent approaches have introduced Region-Language Alignment (RLA) to enhance CLIP's performance in dense multimodal tasks by aligning regional visual representations with corresponding text inputs. However, we find that CLIP ViTs fine-tuned with RLA suffer from notable loss in spatial awareness, which is crucial for dense prediction tasks. To address this, we propose the Spatial Correlation Distillation (SCD) framework, which preserves CLIP's inherent spatial structure and mitigates above degradation. To further enhance spatial correlations, we introduce a lightweight Refiner that extracts refined correlations directly from CLIP before feeding them into SCD, based on an intriguring finding that CLIP naturally capture high-quality dense features. Together, these components form a robust distillation framework that enables CLIP ViTs to integrate both visual-language and visual-centric improvements, achieving state-of-the-art results across various open-vocabulary dense prediction benchmarks.
****
Xiaochi Qian,Shangtong Zhang
Revisiting a Design Choice in Gradient Temporal Difference Learning
Off-policy learning enables a reinforcement learning (RL) agent to reason counterfactually about policies that are not executed and is one of the most important ideas in RL. It, however, can lead to instability when combined with function approximation and bootstrapping, two arguably indispensable ingredients for large-scale reinforcement learning. This is the notorious deadly triad. The seminal work Sutton et al. (2008) pioneers Gradient Temporal Difference learning (GTD) as the first solution to the deadly triad, which has enjoyed massive success thereafter. During the derivation of GTD, some intermediate algorithm, called $A^\top$TD, was invented but soon deemed inferior. In this paper, we revisit this $A^\top$TD and prove that a variant of $A^\top$TD, called $A_t^\top$TD, is also an effective solution to the deadly triad. Furthermore, this $A_t^\top$TD only needs one set of parameters and one learning rate. By contrast, GTD has two sets of parameters and two learning rates, making it hard to tune in practice. We provide asymptotic analysis for $A^\top_t$TD and finite sample analysis for a variant of $A^\top_t$TD that additionally involves a projection operator. The convergence rate of this variant is on par with the canonical on-policy temporal difference learning.
****
Valerii Iakovlev,Harri Lähdesmäki
Learning Spatiotemporal Dynamical Systems from Point Process Observations
Spatiotemporal dynamics models are fundamental for various domains, from heat propagation in materials to oceanic and atmospheric flows. However, currently avai

lable neural network-based spatiotemporal modeling approaches fall short when fa
ced with data that is collected randomly over time and space, as is often the ca
se with sensor networks in real-world applications like crowdsourced earthquake
detection or pollution monitoring. In response, we developed a new method that c
an effectively learn spatiotemporal dynamics from such point process observation
s. Our model integrates techniques from neural differential equations, neural po
int processes, implicit neural representations and amortized variational inferen
ce to model both the dynamics of the system and the probabilistic locations and
timings of observations. It outperforms existing methods on challenging spatiote
mporal datasets by offering substantial improvements in predictive accuracy and
computational efficiency, making it a useful tool for modeling and understanding
 complex dynamical systems observed under realistic, unconstrained conditions.
****

Guy Kaplan,Matanel Oren,Yuval Reif,Roy Schwartz

From Tokens to Words: On the Inner Lexicon of LLMs

Natural language is composed of words, but modern large language models (LLMs) p
rocess sub-words as input. A natural question raised by this discrepancy is whet
her LLMs encode words internally, and if so how. We present evidence that LLMs e
ngage in an intrinsic detokenization process, where subword sequences are combin
ed into coherent whole-word representations at their last token. Our experiments
 show that this process primarily takes place within the early and middle layers
 of the model. We further demonstrate its robustness to arbitrary splits (e.g.,
“cats” to “ca” and “ts”), typos, and importantly—to out-of-vocabulary words: whe
n feeding the last token internal representations of such words to the model as
input, it can “understand” them as the complete word despite never seeing such r
epresentations as input during training. Our findings suggest that LLMs maintain
 a latent vocabulary beyond the tokenizer’s scope. These insights provide a prac
tical, finetuning-free application for expanding the vocabulary of pre-trained m
odels. By enabling the addition of new vocabulary words, we reduce input length
and inference iterations, which reduces both space and model latency, with littl
e to no loss in model accuracy.
****

Md Rifat Arefin,Gopeshh Subbaraj,Nicolas Gontier,Yann LeCun,Irina Rish,Ravid Shw
artz-Ziv,Christopher Pal

Seq-VCR: Preventing Collapse in Intermediate Transformer Representations for En
hanced Reasoning

Decoder-only Transformers often struggle with complex reasoning tasks, particula
rly arithmetic reasoning requiring multiple sequential operations. In this work,
 we identify representation collapse in the model’s intermediate layers as a key
 factor limiting their reasoning capabilities. To address this, we propose Seque
ntial Variance-Covariance Regularization (Seq-VCR), which enhances the entropy o
f intermediate representations and prevents collapse. Combined with dummy pause
tokens as substitutes for chain-of-thought (CoT) tokens, our method significantl
y improves performance in arithmetic reasoning problems. In the challenging 5 ×
5 integer multiplication task, our approach achieves 99.5% exact match accuracy,
 outperforming models of the same size (which yield 0% accuracy) and GPT-4 with
five-shot CoT prompting (44%). We also demonstrate superior results on arithmeti
c expression and longest increasing subsequence (LIS) datasets. Our findings hig
hlight the importance of preventing intermediate layer representation collapse t
o enhance the reasoning capabilities of Transformers and show that Seq-VCR offer
s an effective solution without requiring explicit CoT supervision.
****

Mor Ventura,Michael Toker,Nitay Calderon,Zorik Gekhman,Yonatan Bitton,Roi Reicha
rt

NL-Eye: Abductive NLI For Images

Will a Visual Language Model (VLM)-based bot warn us about slipping if it detect
s a wet floor? Recent VLMs have demonstrated impressive capabilities, yet their
ability to infer outcomes and causes remains underexplored. To address this, we
introduce NL-Eye, a benchmark designed to assess VLMs' visual abductive reasonin
g skills. NL-Eye adapts the abductive Natural Language Inference (NLI) task to t

he visual domain, requiring models to evaluate the plausibility of hypothesis im
ages based on a premise image and explain their decisions. NL-Eye consists of 35
0 carefully curated triplet examples (1,050 images) spanning diverse reasoning c
ategories: physical, functional, logical, emotional, cultural, and social. The d
ata curation process involved two steps—writing textual descriptions and generat
ing images using text-to-image models, both requiring substantial human involvem
ent to ensure high-quality and challenging scenes. Our experiments show that VLM
s struggle significantly on NL-Eye, often performing at random baseline levels,
while humans excel in both plausibility prediction and explanation quality. This
 demonstrates a deficiency in the abductive reasoning capabilities of modern VLM
s. NL-Eye represents a crucial step toward developing VLMs capable of robust mul
timodal reasoning for real-world applications, including accident-prevention bot
s and generated video verification.
****
Théophane Vallaeys,Matthew J. Muckley,Jakob Verbeek,Matthijs Douze
Qinco2: Vector Compression and Search with Improved Implicit Neural Codebooks
Vector quantization is a fundamental technique for compression and large-scale n
earest neighbor search. For high-accuracy operating points, multi-codebook quant
ization associates data vectors with one element from each of multiple codebook
s. An example is residual quantization (RQ), which iteratively quantizes the res
idual error of previous steps. Dependencies between the different parts of the c
ode are, however, ignored in RQ, which leads to suboptimal rate-distortion perfo
rmance. Qinco recently addressed this inefficiency by using a neural network to
determine the quantization codebook in RQ based on the vector reconstruction fro
m previous steps. In this paper we introduce Qinco2 which extends and improves
Qinco with (i) improved vector encoding using codeword pre-selection and beam-
search, (ii) a fast approximate decoder leveraging codeword pairs to establish
 accurate short-lists for search, and (iii) an optimized training procedure and
network architecture. We conduct experiments on four datasets to evaluate Qinco2
 for vector compression and billion-scale nearest neighbor search. We obtain ou
tstanding results in both settings, improving the state-of-the-art reconstructi
on MSE by 44% for 16-byte vector compression on BigANN, and search accuracy by 2
4% with 8-byte encodings on Deep1M.
****
Zeqi Xiao,Wenqi Ouyang,Yifan Zhou,Shuai Yang,Lei Yang,Jianlou Si,Xingang Pan
Trajectory attention for fine-grained video motion control
Recent advancements in video generation have been greatly driven by video diffus
ion models, with camera motion control emerging as a crucial challenge in creati
ng view-customized visual content. This paper introduces trajectory attention, a
 novel approach that performs attention along available pixel trajectories for f
ine-grained camera motion control. Unlike existing methods that often yield impr
ecise outputs or neglect temporal correlations, our approach possesses a stronge
r inductive bias that seamlessly injects trajectory information into the video g
eneration process. Importantly, our approach models trajectory attention as an a
uxiliary branch alongside traditional temporal attention. This design enables th
e original temporal attention and the trajectory attention to work in synergy, e
nsuring both
precise motion control and new content generation capability, which is critical
when the trajectory is only partially available. Experiments on camera motion co
ntrol for images and videos demonstrate significant improvements in precision an
d long-range consistency while maintaining high-quality generation. Furthermore,
 we show that our approach can be extended to other video motion control tasks,
such as first-frame-guided video editing, where it excels in maintaining content
 consistency over large spatial and temporal ranges.
****
Shengyuan Zhang,Ling Yang,Zejian Li,An Zhao,Chenye Meng,Changyuan Yang,Guang Yan
g,Zhiyuan Yang,Lingyun Sun
Distribution Backtracking Builds A Faster Convergence Trajectory for Diffusion D
istillation
Accelerating the sampling speed of diffusion models remains a significant challe

nge. Recent score distillation methods distill a heavy teacher model into a stud
ent generator to achieve one-step generation, which is optimized by calculating
the difference between two score functions on the samples generated by the stude
nt model.
However, there is a score mismatch issue in the early stage of the score distill
ation process, since existing methods mainly focus on using the endpoint of pre-
trained diffusion models as teacher models, overlooking the importance of the co
nvergence trajectory between the student generator and the teacher model.
To address this issue, we extend the score distillation process by introducing t
he entire convergence trajectory of the teacher model and propose $\textbf{Dis}$
tribution $\textbf{Back}$tracking Distillation ($\textbf{DisBack}$). DisBask is
composed of two stages: $\textit{Degradation Recording}$ and $\textit{Distributi
on Backtracking}$.
$\textit{Degradation Recording}$ is designed to obtain the convergence trajector
y by recording the degradation path from the pre-trained teacher model to the un
trained student generator.
The degradation path implicitly represents the intermediate distributions betwee
n the teacher and the student, and its reverse can be viewed as the convergence
trajectory from the student generator to the teacher model.
Then $\textit{Distribution Backtracking}$ trains the student generator to backtr
ack the intermediate distributions along the path to approximate the convergence
 trajectory of the teacher model.
Extensive experiments show that DisBack achieves faster and better convergence t
han the existing distillation method and achieves comparable or better generatio
n performance, with an FID score of 1.38 on the ImageNet 64$\times$64 dataset.
DisBack is easy to implement and can be generalized to existing distillation met
hods to boost performance.
****

Quan Xiao,Tianyi Chen

Unlocking Global Optimality in Bilevel Optimization: A Pilot Study

Bilevel optimization has witnessed a resurgence of interest, driven by its criti
cal role in trustworthy and efficient AI applications. Recent focus has been on
finding efficient methods with provable convergence guarantees. However, while
many prior works have established convergence to stationary points or local mini
ma, obtaining the global optimum of bilevel optimization remains an important ye
t open problem. The difficulty lies in the fact that unlike many prior non-conve
x single-level problems, bilevel problems often do not admit a ``benign" landsca
pe, and may indeed have multiple spurious local solutions. Nevertheless, attaini
ng the global optimality is indispensable for ensuring reliability, safety, and
cost-effectiveness, particularly in high-stakes engineering applications that re
ly on bilevel optimization. In this paper, we first explore the challenges of es
tablishing a global convergence theory for bilevel optimization, and present two
 sufficient conditions for global convergence. We provide {\em algorithm-depende
nt} proofs to rigorously substantiate these sufficient conditions on two specifi
c bilevel learning scenarios: representation learning and data hypercleaning (a.
k.a. reweighting). Experiments corroborate the theoretical findings, demonstrati
ng convergence to global minimum in both cases.
****

Yixian Zhang,Huaze Tang,Huijing Lin,Wenbo Ding

Residual Kernel Policy Network: Enhancing Stability and Robustness in RKHS-Based
 Reinforcement Learning

Achieving optimal performance in reinforcement learning requires robust policies
 supported by training processes that ensure both sample efficiency and stabilit
y. Modeling the policy in reproducing kernel Hilbert space (RKHS) enables effici
ent exploration of local optimal solutions. However, the stability of existing R
KHS-based methods is hindered by significant variance in gradients, while the ro
bustness of the learned policies is often compromised due to the sensitivity of
hyperparameters. In this work, we conduct a comprehensive analysis of the signif
icant instability in RKHS policies and reveal that the variance of the policy gr
adient increases substantially when a wide-bandwidth kernel is employed. To addr

ess these challenges, we propose a novel RKHS policy learning method integrated with representation learning to dynamically process observations in complex environments, enhancing the robustness of RKHS policies. Furthermore, inspired by the advantage functions, we introduce a residual layer that further stabilizes the training process by significantly reducing gradient variance in RKHS. Our novel algorithm, the Residual Kernel Policy Network (ResKPN), demonstrates state-of-the-art performance, achieving a 30% improvement in episodic rewards across complex environments.
****
Qiming Huang,Han Hu,Jianbo Jiao
Revisit the Open Nature of Open Vocabulary Semantic Segmentation
In Open Vocabulary Semantic Segmentation (OVS), we observe a consistent drop in model performance as the query vocabulary set expands, especially when it includes semantically similar and ambiguous vocabularies, such as ‘sofa’ and ‘couch’. The previous OVS evaluation protocol, however, does not account for such ambiguity, as any mismatch between model-predicted and human-annotated pairs is simply treated as incorrect on a pixel-wise basis. This contradicts the open nature of OVS, where ambiguous categories may both be correct from an open-world perspective. To address this, in this work, we study the open nature of OVS and propose a mask-wise evaluation protocol that is based on matched and mismatched mask pairs between prediction and annotation respectively. Extensive experimental evaluations show that the proposed mask-wise protocol provides a more effective and reliable evaluation framework for OVS models compared to the previous pixel-wise approach on the perspective of open-world. Moreover, analysis of mismatched mask pairs reveals that a large amount of ambiguous categories exist in commonly used OVS datasets. Interestingly, we find that reducing these ambiguities during both training and inference enhances capabilities of OVS models. These findings and the new evaluation protocol encourage further exploration of the open nature of OVS, as well as broader open-world challenges. Project page: https://qiming-huang.github.io/RevisitOVS/.
****
Shangding Gu,Laixi Shi,Muning Wen,Ming Jin,Eric Mazumdar,Yuejie Chi,Adam Wierman,Costas Spanos
Robust Gymnasium: A Unified Modular Benchmark for Robust Reinforcement Learning
Driven by inherent uncertainty and the sim-to-real gap, robust reinforcement learning (RL) seeks to improve resilience against the complexity and variability in agent-environment sequential interactions. Despite the existence of a large number of RL benchmarks, there is a lack of standardized benchmarks for robust RL. Current robust RL policies often focus on a specific type of uncertainty and are evaluated in distinct, one-off environments. In this work, we introduce Robust-Gymnasium, a unified modular benchmark designed for robust RL that supports a wide variety of disruptions across all key RL components—agents' observed state and reward, agents' actions, and the environment. Offering over sixty diverse task environments spanning control and robotics, safe RL, and multi-agent RL, it provides an open-source and user-friendly tool for the community to assess current methods and foster the development of robust RL algorithms.
In addition, we benchmark existing standard and robust RL algorithms within this framework, uncovering significant deficiencies in each and offering new insights.
****
Caleb Chuck,Fan Feng,Carl Qi,Chang Shi,Siddhant Agarwal,Amy Zhang,Scott Niekum
Null Counterfactual Factor Interactions for Goal-Conditioned Reinforcement Learning
Hindsight relabeling is a powerful tool for overcoming sparsity in goal-conditioned reinforcement learning (GCRL), especially in certain domains such as navigation and locomotion. However, hindsight relabeling can struggle in object-centric domains. For example, suppose that the goal space consists of a robotic arm pus

hing a particular target block to a goal location. In this case, hindsight relab
eling will give high rewards to any trajectory that does not interact with the b
lock. However, these behaviors are only useful when the object is already at the
 goal---an extremely rare case in practice. A dataset dominated by these kinds o
f trajectories can complicate learning and lead to failures. In object-centric d
omains, one key intuition is that meaningful trajectories are often characterize
d by object-object interactions such as pushing the block with the gripper. To l
everage this intuition, we introduce Hindsight Relabeling using Interactions (HI
nt), which combines interactions with hindsight relabeling to improve the sample
 efficiency of downstream RL. However, interactions do not have a consensus stat
istical definition that is tractable for downstream GCRL. Therefore, we propose
a definition of interactions based on the concept of _null counterfactual_: a ca
use object is interacting with a target object if, in a world where the cause ob
ject did not exist, the target object would have different transition dynamics.
We leverage this definition to infer interactions in Null Counterfactual Interac
tion Inference (NCII), which uses a ``nulling'' operation with a learned model t
o simulate absences and infer interactions. We demonstrate that NCII is able to
achieve significantly improved interaction inference accuracy in both simple lin
ear dynamics domains and dynamic robotic domains in Robosuite, Robot Air Hockey,
 and Franka Kitchen. Furthermore, we demonstrate that HInt improves sample effic
iency by up to $4\times$ in these domains as goal-conditioned tasks.
****
Can Demircan,Tankred Saanum,Akshay Kumar Jagadish,Marcel Binz,Eric Schulz
Sparse Autoencoders Reveal Temporal Difference Learning in Large Language Models
In-context learning, the ability to adapt based on a few examples in the input p
rompt, is a ubiquitous feature of large language models (LLMs). However, as LLMs
' in-context learning abilities continue to improve, understanding this phenomen
on mechanistically becomes increasingly important. In particular, it is not well
-understood how LLMs learn to solve specific classes of problems, such as reinfo
rcement learning (RL) problems, in-context. Through three different tasks, we fi
rst show that Llama 3 70B can solve simple RL problems in-context. We then a
nalyze the residual stream of Llama using Sparse Autoencoders (SAEs) and find re
presentations that closely match temporal difference (TD) errors. Notably, these
 representations emerge despite the model only being trained to predict the next
 token. We verify that these representations are indeed causally involved in the
 computation of TD errors and Q-values by performing carefully designed interv
entions on them. Taken together, our work establishes a methodology for studying
 and manipulating in-context learning with SAEs, paving the way for a more mecha
nistic understanding.
****
Xiao Liu,Tianjie Zhang,Yu Gu,Iat Long Iong,Song XiXuan,Yifan Xu,Shudan Zhang,Han
yu Lai,Jiadai Sun,Xinyue Yang,Yu Yang,Zehan Qi,Shuntian Yao,Xueqiao Sun,Siyi Che
ng,Qinkai Zheng,Hao Yu,Hanchen Zhang,Wenyi Hong,Ming Ding,Lihang Pan,Xiaotao Gu,
Aohan Zeng,Zhengxiao Du,Chan Hee Song,Yu Su,Yuxiao Dong,Jie Tang
VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents
Large Multimodal Models (LMMs) have ushered in a new era in artificial intellige
nce, merging capabilities in both language and vision to form highly capable \te
xtbf{Visual Foundation Agents} that are postulated to excel across a myriad of t
asks. However, existing benchmarks fail to sufficiently challenge or showcase th
e full potential of LMMs as visual foundation agents in complex, real-world envi
ronments. To address this gap, we introduce VisualAgentBench (VAB), a comprehens
ive and unified benchmark specifically designed to train and evaluate LMMs as vi
sual foundation agents across diverse scenarios in one standard setting, includi
ng Embodied, Graphical User Interface, and Visual Design, with tasks formulated
to probe the depth of LMMs' understanding and interaction capabilities. Through
rigorous testing across 9 proprietary LMM APIs and 9 open models (18 in total),
we demonstrate the considerable yet still developing visual agent capabilities o
f these models. Additionally, VAB explores the synthesizing of visual agent traj
ectory data through hybrid methods including Program-based Solvers, LMM Agent Bo
otstrapping, and Human Demonstrations, offering insights into obstacles, solutio

ns, and trade-offs one may meet in developing open LMM agents. Our work not only aims to benchmark existing models but also provides an instrumental playground for future development into visual foundation agents. Code, train, and test data are available at \url{https://github.com/THUDM/VisualAgentBench}.
****
Minjun Kim,Jongjin Kim,U Kang
SynQ: Accurate Zero-shot Quantization by Synthesis-aware Fine-tuning
How can we accurately quantize a pre-trained model without any data?
Quantization algorithms are widely used for deploying neural networks on resource-constrained edge devices.
Zero-shot Quantization (ZSQ) addresses the crucial and practical scenario where training data are inaccessible for privacy or security reasons.
However, three significant challenges hinder the performance of existing ZSQ methods: 1) noise in the synthetic dataset, 2) predictions based on off-target patterns, and the 3) misguidance by erroneous hard labels.
In this paper, we propose SynQ (Synthesis-aware Fine-tuning for Zero-shot Quantization),
a carefully designed ZSQ framework to overcome the limitations of existing methods.
SynQ minimizes the noise from the generated samples by exploiting a low-pass filter.
Then, SynQ trains the quantized model to improve accuracy by aligning its class activation map with the pre-trained model.
Furthermore, SynQ mitigates misguidance from the pre-trained model's error by leveraging only soft labels for difficult samples.
Extensive experiments show that SynQ provides the state-of-the-art accuracy, over existing ZSQ methods.
****
Jiacheng Chen,Tianhao Liang,Sherman Siu,Zhengqing Wang,Kai Wang,Yubo Wang,Yuansheng Ni,Ziyan Jiang,Wang Zhu,Bohan Lyu,Dongfu Jiang,Xuan He,Yuan Liu,Hexiang Hu,Xiang Yue,Wenhu Chen
MEGA-Bench: Scaling Multimodal Evaluation to over 500 Real-World Tasks
We present MEGA-Bench, an evaluation suite that scales multimodal evaluation to over 500 real-world tasks, to address the highly heterogeneous daily use cases of end users.
Our objective is to optimize for a set of high-quality data samples that cover a highly diverse and rich set of multimodal tasks, while enabling cost-effective and accurate model evaluation.
In particular, we collected 505 realistic tasks encompassing over 8,000 samples from 16 expert annotators to extensively cover the multimodal task space. Instead of unifying these problems into standard multi-choice questions (like MMMU, MM-Bench, and MMT-Bench), we embrace a wide range of output formats like numbers, phrases, code, \LaTeX, coordinates, JSON, free-form, etc. To accommodate these formats, we developed over 40 metrics to evaluate these tasks.
Unlike existing benchmarks, MEGA-Bench offers a fine-grained capability report across multiple dimensions (e.g., application, input type, output format, skill), allowing users to interact with and visualize model capabilities in depth. We evaluate a wide variety of frontier vision-language models on MEGA-Bench to understand their capabilities across these dimensions.
****
Ola Rønning,Eric Nalisnick,Christophe Ley,Padhraic Smyth,Thomas Hamelryck
ELBOing Stein: Variational Bayes with Stein Mixture Inference
Stein variational gradient descent (SVGD) (Liu & Wang, 2016) performs approximate Bayesian inference by representing the posterior with a set of particles.
However, SVGD suffers from variance collapse, i.e. poor predictions due to underestimating uncertainty (Ba et al., 2021), even for moderately-dimensional models such as small Bayesian neural networks (BNNs). To address this issue, we generalize SVGD by letting each particle parameterize a component distribution in a mixture model. Our method, Stein Mixture Inference (SMI), optimizes a lower bound to the evidence (ELBO) and introduces user-specified guides parameterized

by particles. SMI extends the Nonlinear SVGD framework (Wang & Liu, 2019) to the case of variational Bayes. SMI effectively avoids variance collapse, judging by a previously described test developed for this purpose, and performs well on standard data sets. In addition, SMI requires considerably fewer particles than SVGD to accurately estimate uncertainty for small BNNs. The synergistic combination of NSVGD, ELBO optimization and user-specified guides establishes a promising approach towards variational Bayesian inference in the case of tall and wide data.
****
Qingming LIU,Yuan Liu,Jiepeng Wang,Xianqiang Lyu,Peng Wang,Wenping Wang,Junhui Hou
MoDGS: Dynamic Gaussian Splatting from Casually-captured Monocular Videos with Depth Priors
In this paper, we propose MoDGS, a new pipeline to render novel-view images in dynamic scenes using only casually captured monocular videos. Previous monocular dynamic NeRF or Gaussian Splatting methods strongly rely on the rapid movement of input cameras to construct multiview consistency but fail to reconstruct dynamic scenes on casually captured input videos whose cameras are static or move slowly. To address this challenging task, MoDGS adopts recent single-view depth estimation methods to guide the learning of the dynamic scene. Then, a novel 3D-aware initialization method is proposed to learn a reasonable deformation field and a new robust depth loss is proposed to guide the learning of dynamic scene geometry. Comprehensive experiments demonstrate that MoDGS is able to render high-quality novel view images of dynamic scenes from just a casually captured monocular video, which outperforms baseline methods by a significant margin. Project page: https://MoDGS.github.io
****
Yu Chen,Jiatai Huang,Yan Dai,Longbo Huang
uniINF: Best-of-Both-Worlds Algorithm for Parameter-Free Heavy-Tailed MABs
In this paper, we present a novel algorithm, `uniINF`, for the Heavy-Tailed Multi-Armed Bandits (HTMAB) problem, demonstrating robustness and adaptability in both stochastic and adversarial environments. Unlike the stochastic MAB setting where loss distributions are stationary with time, our study extends to the adversarial setup, where losses are generated from heavy-tailed distributions that depend on both arms and time. Our novel algorithm `uniINF` enjoys the so-called Best-of-Both-Worlds (BoBW) property, performing optimally in both stochastic and adversarial environments *without* knowing the exact environment type. Moreover, our algorithm also possesses a Parameter-Free feature, *i.e.*, it operates *without* the need of knowing the heavy-tail parameters (σ, α) a-priori. To be precise, `uniINF` ensures nearly-optimal regret in both stochastic and adversarial environments, matching the corresponding lower bounds when (σ, α) is known (up to logarithmic factors). To our knowledge, `uniINF` is the first parameter-free algorithm to achieve the BoBW property for the heavy-tailed MAB problem. Technically, we develop innovative techniques to achieve BoBW guarantees for Parameter-Free HTMABs, including a refined analysis for the dynamics of log-barrier, an auto-balancing learning rate scheduling scheme, an adaptive skipping-clipping loss tuning technique, and a stopping-time analysis for logarithmic regret.
****
Alizée Pace,Bernhard Schölkopf,Gunnar Ratsch,Giorgia Ramponi
Preference Elicitation for Offline Reinforcement Learning
Applying reinforcement learning (RL) to real-world problems is often made challenging by the inability to interact with the environment and the difficulty of designing reward functions. Offline RL addresses the first challenge by considering access to an offline dataset of environment interactions labeled by the reward function. In contrast, Preference-based RL does not assume access to the reward function and learns it from preferences, but typically requires an online inter

action with the environment. We bridge the gap between these frameworks by explo
ring efficient methods for acquiring preference feedback in a fully offline setu
p. We propose Sim-OPRL, an offline preference-based reinforcement learning algor
ithm, which leverages a learned environment model to elicit preference feedback
on simulated rollouts. Drawing on insights from both the offline RL and the pref
erence-based RL literature, our algorithm employs a pessimistic approach for out
-of-distribution data, and an optimistic approach for acquiring informative pref
erences about the optimal policy. We provide theoretical guarantees regarding th
e sample complexity of our approach, dependent on how well the offline data cove
rs the optimal policy. Finally, we demonstrate the empirical performance of Sim-
OPRL in various environments.
****

EungGu Yun,Heonjin Ha,Yeongwoo Nam,Bryan Dongik Lee
Language-Assisted Feature Transformation for Anomaly Detection
This paper introduces LAFT, a novel feature transformation method designed to in
corporate user knowledge and preferences into anomaly detection using natural la
nguage. Accurately modeling the boundary of normality is crucial for distinguish
ing abnormal data, but this is often challenging due to limited data or the pres
ence of nuisance attributes. While unsupervised methods that rely solely on data
 without user guidance are common, they may fail to detect anomalies of specific
 interest. To address this limitation, we propose Language-Assisted Feature Tran
sformation (LAFT), which leverages the shared image-text embedding space of visi
on-language models to transform visual features according to user-defined requir
ements. Combined with anomaly detection methods, LAFT effectively aligns visual
features with user preferences, allowing anomalies of interest to be detected. E
xtensive experiments on both toy and real-world datasets validate the effectiven
ess of our method.
****

Haodong Hong,Yanyuan Qiao,Sen Wang,Jiajun Liu,Qi Wu
General Scene Adaptation for Vision-and-Language Navigation
Vision-and-Language Navigation (VLN) tasks mainly evaluate agents based on one-t
ime execution of individual instructions across multiple environments, aiming to
 develop agents capable of functioning in any environment in a zero-shot manner.
 However, real-world navigation robots often operate in persistent environments
with relatively consistent physical layouts, visual observations, and language s
tyles from instructors. Such a gap in the task setting presents an opportunity t
o improve VLN agents by incorporating continuous adaptation to specific environm
ents. To better reflect these real-world conditions, we introduce GSA-VLN (Gener
al Scene Adaptation for VLN), a novel task requiring agents to execute navigatio
n instructions within a specific scene and simultaneously adapt to it for improv
ed performance over time. To evaluate the proposed task, one has to address two
challenges in existing VLN datasets: the lack of out-of-distribution (OOD) data,
 and the limited number and style diversity of instructions for each scene. Ther
efore, we propose a new dataset, GSA-R2R, which significantly expands the divers
ity and quantity of environments and instructions for the Room-to-Room (R2R) dat
aset to evaluate agent adaptability in both ID and OOD contexts. Furthermore, we
 design a three-stage instruction orchestration pipeline that leverages large la
nguage models (LLMs) to refine speaker-generated instructions and apply role-pla
ying techniques to rephrase instructions into different speaking styles. This is
 motivated by the observation that each individual user often has consistent sig
natures or preferences in their instructions, taking the use case of home robot
ic assistants as an example. We conducted extensive experiments on GSA-R2R to th
oroughly evaluate our dataset and benchmark various methods, revealing key facto
rs enabling agents to adapt to specific environments. Based on our findings, we
propose a novel method, Graph-Retained DUET (GR-DUET), which incorporates memory
-based navigation graphs with an environment-specific training strategy, achievi
ng state-of-the-art results on all GSA-R2R splits.
****

Marta Skreta,Lazar Atanackovic,Joey Bose,Alexander Tong,Kirill Neklyudov
The Superposition of Diffusion Models Using the Itô Density Estimator

The Cambrian explosion of easily accessible pre-trained diffusion models suggest s a demand for methods that combine multiple different pre-trained diffusion mod els without incurring the significant computational burden of re-training a larg er combined model. In this paper, we cast the problem of combining multiple pre-trained diffusion models at the generation stage under a novel proposed framewor k termed superposition. Theoretically, we derive superposition from rigorous fir st principles stemming from the celebrated continuity equation and design two no vel algorithms tailor-made for combining diffusion models in SuperDiff. SuperDif f leverages a new scalable Itô density estimator for the log likelihood of the d iffusion SDE which incurs *no additional overhead* compared to the well-known Hu tchinson's estimator needed for divergence calculations. We demonstrate that Sup erDiff is scalable to large pre-trained diffusion models as superposition is per formed *solely through composition during inference*, and also enjoys painless i mplementation as it combines different pre-trained vector fields through an auto mated re-weighting scheme. Notably, we show that SuperDiff is efficient during i nference time, and mimics traditional composition operators such as the logical OR and the logical AND. We empirically demonstrate the utility of using SuperDi ff for generating more diverse images on CIFAR-10, more faithful prompt conditio ned image editing using Stable Diffusion, as well as improved conditional molecu le generation and unconditional *de novo* structure design of proteins. https://github.com/necludov/super-diffusion

****

Zizheng Pan,Bohan Zhuang,De-An Huang,Weili Nie,Zhiding Yu,Chaowei Xiao,Jianfei Cai,Anima Anandkumar

T-Stitch: Accelerating Sampling in Pre-Trained Diffusion Models with Trajectory Stitching

Sampling from diffusion probabilistic models (DPMs) is often expensive for high-quality image generation and typically requires many steps with a large model. In this paper, we introduce sampling Trajectory Stitching (T-Stitch), a simple yet efficient technique to improve the sampling efficiency with little or no generation degradation. Instead of solely using a large DPM for the entire sampling trajectory, T-Stitch first leverages a smaller DPM in the initial steps as a cheap drop-in replacement of the larger DPM and switches to the larger DPM at a later stage. Our key insight is that different diffusion models learn similar encodings under the same training data distribution and smaller models are capable of generating good global structures in the early steps. Extensive experiments demonstrate that T-Stitch is training-free, generally applicable for different architectures, and complements most existing fast sampling techniques with flexible speed and quality trade-offs. On DiT-XL, for example, 40% of the early timesteps can be safely replaced with a 10x faster DiT-S without performance drop on class-conditional ImageNet generation. We further show that our method can also be used as a drop-in technique to not only accelerate the popular pretrained stable diffusion (SD) models but also improve the prompt alignment of stylized SD models from the public model zoo. Finally, the explicit model allocation strategy of T-Stitch significantly reduces the need of training or searching, delivering high deployment efficiency.

****

Ruixuan Liu,Zhiqi Bu

Towards hyperparameter-free optimization with differential privacy

Differential privacy (DP) is a privacy-preserving paradigm that protects the training data when training deep learning models. Critically, the performance of models is determined by the training hyperparameters, especially those of the learning rate schedule, thus requiring fine-grained hyperparameter tuning on the data. In practice, it is common to tune the learning rate hyperparameters through the grid search that (1) is computationally expensive as multiple runs are needed, and (2) increases the risk of data leakage as the selection of hyperparameters is data-dependent. In this work, we adapt the automatic learning rate schedule to DP optimization for any models and optimizers, so as to significantly mitigate or even eliminate the cost of hyperparameter tuning when applied together with automatic per-sample gradient clipping. Our hyperparameter-free DP optimization

is almost as computationally efficient as the standard non-DP optimization, and achieves state-of-the-art DP performance on various language and vision tasks.
****

Dongzhuoran Zhou,Evgeny Kharlamov,Egor V. Kostylev

GLoRa: A Benchmark to Evaluate the Ability to Learn Long-Range Dependencies in Graphs

Learning on graphs is one of the most active research topics in machine learning (ML). Among the key challenges in this field, effectively learning long-range dependencies in graphs has been particularly difficult. It has been observed that, in practice, the performance of many ML approaches, including various types of graph neural networks (GNNs), degrades significantly when the learning task involves long-range dependencies—that is, when the answer is determined by the presence of a certain path of significant length in the graph. This issue has been attributed to several phenomena, including over-smoothing, over-squashing, and vanishing gradient. A number of solutions have been proposed to mitigate these causes. However, evaluation of these solutions is currently challenging because existing benchmarks do not effectively test systems for their ability to learn tasks based on long-range dependencies in a transparent manner. In this paper, we introduce GLoRa, a synthetic benchmark that allows testing of systems for this ability in a systematic way. We then evaluate state-of-the-art systems using GLoRa and conclude that none of them can confidently claim to learn long-range dependencies well. We also observe that this weak performance cannot be attributed to any of the three causes, highlighting the need for further investigation.
****

GUOJUN XIONG,Ujwal Dinesha,Debajoy Mukherjee,Jian Li,Srinivas Shakkottai

DOPL: Direct Online Preference Learning for Restless Bandits with Preference Feedback

Restless multi-armed bandits (RMAB) has been widely used to model constrained sequential decision making problems, where the state of each restless arm evolves according to a Markov chain and each state transition generates a scalar reward. However, the success of RMAB crucially relies on the availability and quality of reward signals. Unfortunately, specifying an exact reward function in practice can be challenging and even infeasible. In this paper, we introduce Pref-RMAB, a new RMAB model in the presence of preference signals, where the decision maker only observes pairwise preference feedback rather than scalar reward from the activated arms at each decision epoch. Preference feedback, however, arguably contains less information than the scalar reward, which makes Pref-RMAB seemingly more difficult. To address this challenge, we present a direct online preference learning (DOPL) algorithm for Pref-RMAB to efficiently explore the unknown environments, adaptively collect preference data in an online manner, and directly leverage the preference feedback for decision-makings. We prove that DOPL yields a sublinear regret. To our best knowledge, this is the first algorithm to ensure $\tilde{\mathcal{O}}(\sqrt{T\ln T})$ regret for RMAB with preference feedback. Experimental results further demonstrate the effectiveness of DOPL.
****

Changmin Yu,Maneesh Sahani,Máté Lengyel

Discovering Temporally Compositional Neural Manifolds with Switching Infinite GPFA

Gaussian Process Factor Analysis (GPFA) is a powerful latent variable model for extracting low-dimensional manifolds underlying population neural activities. However, one limitation of standard GPFA models is that the number of latent factors needs to be pre-specified or selected through heuristic-based processes, and that all factors contribute at all times. We propose the infinite GPFA model, a fully Bayesian non-parametric extension of the classical GPFA by incorporating an Indian Buffet Process (IBP) prior over the factor loading process, such that it is possible to infer a potentially infinite set of latent factors, and the identity of those factors that contribute to neural firings in a compositional manner at \textit{each} time point. Learning and inference in the infinite GPFA model is performed through variational expectation-maximisation, and we additionally propose scalable extensions based on sparse variational Gaussian Process method

s. We empirically demonstrate that the infinite GPFA model correctly infers dyna
mically changing activations of latent factors on a synthetic dataset. By fittin
g the infinite GPFA model to population activities of hippocampal place cells du
ring spatial tasks with alternating random foraging and spatial memory phases, w
e identify novel non-trivial and behaviourally meaningful dynamics in the neural
 encoding process.
****
Yi Liu,Changran XU,Yunhao Zhou,Zeju Li,Qiang Xu
DeepRTL: Bridging Verilog Understanding and Generation with a Unified Representa
tion Model
Recent advancements in large language models (LLMs) have shown significant poten
tial for automating hardware description language (HDL) code generation from hig
h-level natural language instructions. While fine-tuning has improved LLMs' perf
ormance in hardware design tasks, prior efforts have largely focused on Verilog
generation, overlooking the equally critical task of Verilog understanding. Furt
hermore, existing models suffer from weak alignment between natural language des
criptions and Verilog code, hindering the generation of high-quality, synthesiza
ble designs. To address these issues, we present DeepRTL, a unified representati
on model that excels in both Verilog understanding and generation. Based on Code
T5+, DeepRTL is fine-tuned on a comprehensive dataset that aligns Verilog code w
ith rich, multi-level natural language descriptions.
We also introduce the first benchmark for Verilog understanding and take the ini
tiative to apply embedding similarity and GPT Score to evaluate the models' unde
rstanding capabilities. These metrics capture semantic similarity more accuratel
y than traditional methods like BLEU and ROUGE, which are limited to surface-lev
el n-gram overlaps. By adapting curriculum learning to train DeepRTL, we enable
it to significantly outperform GPT-4 in Verilog understanding tasks, while achie
ving performance on par with OpenAI's o1-preview model in Verilog generation tas
ks.
****
Derek Qiang Xu,F Olcay Cirit,Reza Asadi,Yizhou Sun,Wei Wang
Mixture of In-Context Prompters for Tabular PFNs
Recent benchmarks find In-Context Learning (ICL) outperforms both deep learning
and tree-based algorithms on small tabular datasets. However, on larger datasets
, ICL for tabular learning suffers in both efficiency and effectiveness. In term
s of efficiency, transformers incur linear space and quadratic time complexity w
.r.t. context size. In terms of effectiveness, contexts at inference encounter d
istribution shift compared to contexts from pretraining. We propose MixturePFN,
which extends Sparse Mixture of Experts to the state-of-the-art ICL for tabular
learning model. Specifically, MixturePFN finetunes a specialized ICL expert on e
ach cluster of tabular data and routes new test samples to appropriate experts a
t inference. MixturePFN supports constant-size contexts by splitting large train
ing datasets into more manageable clusters. MixturePFN addresses distribution sh
ift by finetuning an expert on each training dataset cluster via bootstrapping.
Extensive experimental results shows MixturePFN outperforms 19 baselines both in
 mean rank and as the Condorcet winner across 36 diverse tabular datasets under
both accuracy and F1 score with statistical significance.
****
Zijian Li,Yifan Shen,Kaitao Zheng,Ruichu Cai,Xiangchen Song,Mingming Gong,Guangy
i Chen,Kun Zhang
On the Identification of Temporal Causal Representation with Instantaneous Depen
dence
Temporally causal representation learning aims to identify the latent causal pro
cess from time series observations, but most methods require the assumption that
 the latent causal processes do not have instantaneous relations. Although some
recent methods achieve identifiability in the instantaneous causality case, they
 require either interventions on the latent variables or grouping of the observa
tions, which are in general difficult to obtain in real-world scenarios. To fill
 this gap, we propose an \textbf{ID}entification framework for instantane\textbf
{O}us \textbf{L}atent dynamics (\textbf{IDOL}) by imposing a sparse influence co

nstraint that the latent causal processes have sparse time-delayed and instantan
eous relations. Specifically, we establish identifiability results of the latent
 causal process based on sufficient variability and the sparse influence constra
int by employing contextual information of time series data. Based on these theo
ries, we incorporate a temporally variational inference architecture to estimate
 the latent variables and a gradient-based sparsity regularization to identify t
he latent causal process. Experimental results on simulation datasets illustrate
 that our method can identify the latent causal process. Furthermore, evaluation
s on multiple human motion forecasting benchmarks with instantaneous dependencie
s indicate the effectiveness of our method in real-world settings.
****

Ying Fan,Yilun Du,Kannan Ramchandran,Kangwook Lee

Looped Transformers for Length Generalization

Recent work has shown that Transformers trained from scratch can successfully so
lve various arithmetic and algorithmic tasks, such as adding numbers and computi
ng parity. While these Transformers generalize well on unseen inputs of the same
 length, they struggle with length generalization, i.e., handling inputs of unse
en lengths. In this work, we demonstrate that looped Transformers with an adapti
ve number of steps significantly improve length generalization. We focus on task
s with a known iterative solution, involving multiple iterations of a RASP-L ope
ration—a length-generalizable operation that can be expressed by a finite-sized
Transformer. We train looped Transformers using our proposed learning algorithm
and observe that they learn highly length-generalizable solutions for various ta
sks.
****

Lifan Yuan,Ganqu Cui,Hanbin Wang,Ning Ding,Xingyao Wang,Boji Shan,Zeyuan Liu,Jia
 Deng,Huimin Chen,Ruobing Xie,Yankai Lin,Zhenghao Liu,Bowen Zhou,Hao Peng,Zhiyua
n Liu,Maosong Sun

Advancing LLM Reasoning Generalists with Preference Trees

We introduce EURUS, a suite of large language models (LLMs) optimized for reason
ing. Finetuned from Mistral-7B, Llama-3-8B, and Mixtral-8x22B, EURUS models achi
eve state-of-the-art results among open-source models on a diverse set of benchm
arks covering mathematics, code generation, and logical reasoning problems. Nota
bly, EURUX-8X22B outperforms GPT-3.5 Turbo in reasoning through a comprehensive
benchmarking across 12 test sets covering five tasks. The strong performance of
EURUS can be primarily attributed to ULTRAINTERACT, our newly-curated large-scal
e, high-quality training data dataset specifically designed for complex reasonin
g tasks. ULTRAINTERACT can be used in both supervised fine-tuning, preference le
arning, and reward modeling. It pairs each instruction with a preference tree co
nsisting of (1) reasoning chains with diverse planning strategies in a unified f
ormat, (2) multi-turn interaction trajectories with the environment and the crit
ique, and (3) pairwise positive and negative responses to facilitate preference
learning. ULTRAINTERACT allows us to conduct an in-depth exploration of preferen
ce learning for reasoning tasks. Our investigation reveals that some well-establ
ished preference learning algorithms may be less suitable for reasoning tasks co
mpared to their effectiveness in general conversations. The hypothesis is that i
n reasoning tasks, the space of correct answers is much smaller than that of inc
orrect ones, so it is necessary to explicitly increase the reward of chosen data
. Therefore, in addition to increasing the reward margin as many preference lear
ning algorithms do, the absolute values of positive responses' rewards should be
 positive and may serve as a proxy for performance. Inspired by this, we derive
a novel reward modeling objective and empirically that it leads to a stable rewa
rd modeling curve and better performance. Together with ULTRAINTERACT, we obtain
 a strong reward model.
****

Hongjun Wang,Sagar Vaze,Kai Han

HiLo: A Learning Framework for Generalized Category Discovery Robust to Domain S
hifts

Generalized Category Discovery (GCD) is a challenging task in which, given a par
tially labelled dataset, models must categorize all unlabelled instances, regard

less of whether they come from labelled categories or from new ones. In this pap
er, we challenge a remaining assumption in this task: that all images share the
same domain. Specifically, we introduce a new task and method to handle GCD when
 the unlabelled data also contains images from different domains to the labelled
 set. Our proposed `HiLo' networks extract High-level semantic and Low-level dom
ain features, before minimizing the mutual information between the representatio
ns. Our intuition is that the clusterings based on domain information and semant
ic information should be independent. We further extend our method with a specia
lized domain augmentation tailored for the GCD task, as well as a curriculum lea
rning approach. Finally, we construct a benchmark from corrupted fine-grained da
tasets as well as a large-scale evaluation on DomainNet with real-world domain s
hifts, reimplementing a number of GCD baselines in this setting. We demonstrate
that HiLo outperforms SoTA category discovery models by a large margin on all ev
aluations.
****
Siddhant Arora,Zhiyun Lu,Chung-Cheng Chiu,Ruoming Pang,Shinji Watanabe
Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics
The recent wave of audio foundation models (FMs) could provide new capabilities
for conversational modeling. However, there have been limited efforts to evaluat
e these audio FMs comprehensively on their ability to have natural and interacti
ve conversations. To engage in meaningful conversation with the end user, we wou
ld want the FMs to additionally perform a fluent succession of turns without too
 much overlapping speech or long stretches of silence. Inspired by this, we ask
whether the recently proposed audio FMs can understand, predict, and perform tur
n-taking events? To answer this, we propose a novel evaluation protocol that can
 assess spoken dialog system's turn-taking capabilities using a supervised model
 as a judge that has been trained to predict turn-taking events in human-human c
onversations. Using this protocol, we present the first comprehensive user study
 that evaluates existing spoken dialogue systems on their ability to perform tur
n-taking events and reveal many interesting insights, such as they sometimes do
not understand when to speak up, can interrupt too aggressively and rarely backc
hannel. We further evaluate multiple open-source and proprietary audio FMs acces
sible through APIs on carefully curated test benchmarks from Switchboard to meas
ure their ability to understand and predict turn-taking events and identify sign
ificant room for improvement. We will open source our evaluation platform to pro
mote the development of advanced conversational AI systems.
****
Xu Ouyang,Felix Xiaozhu Lin,Yangfeng Ji
SelectFormer in Data Markets: Privacy-Preserving and Efficient Data Selection fo
r Transformers with Multi-Party Computation
Critical to a free data market is $ \textit{private data selection}$, i.e. the m
odel owner selects and then appraises training data from the data owner before b
oth parties commit to a transaction. To keep the data and model private, this pr
ocess shall evaluate the target model to be trained over Multi-Party Computation
 (MPC). While prior work suggests that evaluating Transformer-based models over
MPC is prohibitively expensive, this paper makes it practical for the purpose of
 data selection. Our contributions are three: (1) a new pipeline for private dat
a selection over MPC; (2) emulating high-dimensional nonlinear operators with lo
w-dimension MLPs, which are trained on a small sample of the data of interest; (
3) scheduling MPC in a parallel, multiphase fashion. We evaluate our method on d
iverse Transformer models and NLP/CV benchmarks. Compared to directly evaluating
 the target model over MPC, our method reduces the delay from thousands of hours
 to tens of hours, while only seeing around 0.20% accuracy degradation from trai
ning with the selected data.
****
Jinwei Yao,Kaiqi Chen,Kexun Zhang,Jiaxuan You,Binhang Yuan,Zeke Wang,Tao Lin
DeFT: Decoding with Flash Tree-attention for Efficient Tree-structured LLM Infer
ence
Large language models (LLMs) are increasingly employed for complex tasks that pr
ocess multiple generation calls in a tree structure with shared prefixes of toke

ns, including few-shot prompting, multi-step reasoning, speculative decoding, et
c. However, existing inference systems for tree-based applications are inefficie
nt due to improper partitioning of queries and KV cache during attention calcula
tion.This leads to two main issues: (1) a lack of memory access (IO) reuse for K
V cache of shared prefixes, and (2) poor load balancing.As a result, there is re
dundant KV cache IO between GPU global memory and shared memory, along with low
GPU utilization. To address these challenges, we propose DeFT(Decoding with Flas
h Tree-Attention), a hardware-efficient attention algorithm with prefix-aware an
d load-balanced KV cache partitions. DeFT reduces the number of read/write opera
tions of KV cache during attention calculation through **KV-Guided Grouping**, a
 method that avoids repeatedly loading KV cache of shared prefixes in attention
computation. Additionally, we propose **Flattened Tree KV Splitting**, a mechani
sm that ensures even distribution of the KV cache across partitions with little
computation redundancy, enhancing GPU utilization during attention computations.
 By reducing 73-99% KV cache IO and nearly 100% IO for partial results during at
tention calculation, DeFT achieves up to 2.23/3.59$\times$ speedup in the end-to
-end/attention latency across three practical tree-based workloads compared to s
tate-of-the-art attention algorithms. Our code is available at https://github.co
m/LINs-lab/DeFT.
****
Luping Liu,Chao Du,Tianyu Pang,Zehan Wang,Chongxuan Li,Dong Xu
Improving Long-Text Alignment for Text-to-Image Diffusion Models
The rapid advancement of text-to-image (T2I) diffusion models has enabled them t
o generate unprecedented results from given texts. However, as text inputs becom
e longer, existing encoding methods like CLIP face limitations, and aligning the
 generated images with long texts becomes challenging. To tackle these issues, w
e propose LongAlign, which includes a segment-level encoding method for processi
ng long texts and a decomposed preference optimization method for effective alig
nment training. For segment-level encoding, long texts are divided into multiple
 segments and processed separately. This method overcomes the maximum input leng
th limits of pretrained encoding models. For preference optimization, we provide
 decomposed CLIP-based preference models to fine-tune diffusion models. Specific
ally, to utilize CLIP-based preference models for T2I alignment, we delve into t
heir scoring mechanisms and find that the preference scores can be decomposed in
to two components: a text-relevant part that measures T2I alignment and a text-i
rrelevant part that assesses other visual aspects of human preference. Additiona
lly, we find that the text-irrelevant part contributes to a common overfitting p
roblem during fine-tuning. To address this, we propose a reweighting strategy th
at assigns different weights to these two components, thereby reducing overfitti
ng and enhancing alignment. After fine-tuning $512 \\times 512$ Stable Diffusion
 (SD) v1.5 for about 20 hours using our method, the fine-tuned SD outperforms st
ronger foundation models in T2I alignment, such as PixArt-$\\alpha$ and Kandinsk
y v2.2. The code is available at https://github.com/luping-liu/LongAlign.
****
Harshit Kumar,Beomseok Kang,Biswadeep Chakraborty,Saibal Mukhopadhyay
Has the Deep Neural Network learned the Stochastic Process? An Evaluation Viewpo
int
This paper presents the first systematic study of evaluating Deep Neural Network
s (DNNs) designed to forecast the evolution of stochastic complex systems. We sh
ow that traditional evaluation methods like threshold-based classification metri
cs and error-based scoring rules assess a DNN's ability to replicate the observe
d ground truth but fail to measure the DNN's learning of the underlying stochast
ic process. To address this gap, we propose a new evaluation criteria called _Fi
delity to Stochastic Process (F2SP)_, representing the DNN's ability to predict
the system property _Statistic-GT_—the ground truth of the stochastic process—an
d introduce an evaluation metric that exclusively assesses F2SP. We formalize F2
SP within a stochastic framework and establish criteria for validly measuring it
. We formally show that Expected Calibration Error (ECE) satisfies the necessary
 condition for testing F2SP, unlike traditional evaluation methods. Empirical ex
periments on synthetic datasets, including wildfire, host-pathogen, and stock ma

rket models, demonstrate that ECE uniquely captures F2SP. We further extend our study to real-world wildfire data, highlighting the limitations of conventional evaluation and discuss the practical utility of incorporating F2SP into model as sessment. This work offers a new perspective on evaluating DNNs modeling complex systems by emphasizing the importance of capturing underlying the stochastic pr ocess.
****

O■uz Kaan Yüksel,Mathieu Even,Nicolas Flammarion
Long-Context Linear System Identification
This paper addresses the problem of long-context linear system identification, w here the state x_t of the system at time t depends linearly on previous stat es x_s over a fixed context window of length p. We establish a sample comple xity bound that matches the _i.i.d._ parametric rate, up to logarithmic factors for a broad class of systems, extending previous work that considered only first -order dependencies. Our findings reveal a ``learning-without-mixing'' phenomeno n, indicating that learning long-context linear autoregressive models is not hin dered by slow mixing properties potentially associated with extended context win dows. Additionally, we extend these results to _(i)_ shared low-rank feature rep resentations, where rank-regularized estimators improve rates with respect to di mensionality, and _(ii)_ misspecified context lengths in strictly stable systems , where shorter contexts offer statistical advantages.
****

Xiaoxuan Hou,Jiayi Yuan,Joel Z Leibo,Natasha Jaques
InvestESG: A multi-agent reinforcement learning benchmark for studying climate i nvestment as a social dilemma
InvestESG is a novel multi-agent reinforcement learning (MARL) benchmark des igned to study the impact of Environmental, Social, and Governance (ESG) disclos ure mandates on corporate climate investments. The benchmark models an intertemp oral social dilemma where companies balance short-term profit losses from climat e mitigation efforts and long-term benefits from reducing climate risk, while ES G-conscious investors attempt to influence corporate behavior through their inve stment decisions. Companies allocate capital across mitigation, greenwashing, an d resilience, with varying strategies influencing climate outcomes and investor preferences. We are releasing open-source versions of InvestESG in both PyTorch and JAX, which enable scalable and hardware-accelerated simulations for investig ating competing incentives in mitigate climate change. Our experiments show that without ESG-conscious investors with sufficient capital, corporate mitigation e fforts remain limited under the disclosure mandate. However, when a critical mas s of investors prioritizes ESG, corporate cooperation increases, which in turn r educes climate risks and enhances long-term financial stability. Additionally, p roviding more information about global climate risks encourages companies to inv est more in mitigation, even without investor involvement. Our findings align wi th empirical research using real-world data, highlighting MARL's potential to in form policy by providing insights into large-scale socio-economic challenges thr ough efficient testing of alternative policy and market designs.
****

Tongtian Zhu,Wenhao Li,Can Wang,Fengxiang He
DICE: Data Influence Cascade in Decentralized Learning
Decentralized learning offers a promising approach to crowdsource data consumpti ons and computational workloads across geographically distributed compute interc onnected through peer-to-peer networks, accommodating the exponentially increasi ng demands. However, proper incentives are still in absence, considerably discou raging participation. Our vision is that a fair incentive mechanism relies on fa ir attribution of contributions to participating nodes, which faces non-trivial challenges arising from the localized connections making influence ``cascade'' i n a decentralized network. To overcome this, we design the first method to esti mate Data Influence CascadE (DICE) in a decentralized environment. Theoretically , the framework derives tractable approximations of influence cascade over arbit rary neighbor hops, suggesting the influence cascade is determined by an interpl ay of data, communication topology, and the curvature of loss landscape.DICE als

o lays the foundations for applications including selecting suitable collaborato
rs and identifying malicious behaviors.
Project page is available at https://raiden-zhu.github.io/blog/2025/DICE.
****

Sheng-Feng Yu,Jia-Jiun Yao,Wei-Chen Chiu
Boost Self-Supervised Dataset Distillation via Parameterization, Predefined Augm
entation, and Approximation
Although larger datasets are crucial for training large deep models, the rapid g
rowth of dataset size has brought a significant challenge in terms of considerab
le training costs, which even results in prohibitive computational expenses. Dat
aset Distillation becomes a popular technique recently to reduce the dataset siz
e via learning a highly compact set of representative exemplars, where the model
 trained with these exemplars ideally should have comparable performance with re
spect to the one trained with the full dataset. While most of existing works upo
n dataset distillation focus on supervised datasets, \todo{we instead aim to dis
till images and their self-supervisedly trained representations into a distilled
 set. This procedure, named as Self-Supervised Dataset Distillation, effectively
 extracts rich information from real datasets, yielding the distilled sets with
enhanced cross-architecture generalizability.} Particularly, in order to preserv
e the key characteristics of original dataset more faithfully and compactly, sev
eral novel techniques are proposed: 1) we introduce an innovative parameterizati
on upon images and representations via distinct low-dimensional bases, where the
 base selection for parameterization is experimentally shown to play a crucial r
ole; 2) we tackle the instability induced by the randomness of data augmentation
 -- a key component in self-supervised learning but being underestimated in the
prior work of self-supervised dataset distillation -- by utilizing predetermined
 augmentations; 3) we further leverage a lightweight network to model the connec
tions among the representations of augmented views from the same image, leading
to more compact pairs of distillation. Extensive experiments conducted on variou
s datasets validate the superiority of our approach in terms of distillation eff
iciency, cross-architecture generalization, and transfer learning performance.
****

Thomas P Zollo,Andrew Wei Tung Siah,Naimeng Ye,Ang Li,Hongseok Namkoong
PersonalLLM: Tailoring LLMs to Individual Preferences
As LLMs become capable of complex tasks, there is growing potential for personal
ized interactions tailored to the subtle and idiosyncratic preferences of the us
er. We present a public benchmark, PersonalLLM, focusing on adapting LLMs to pro
vide maximal benefits for a particular user. Departing from existing alignment b
enchmarks that implicitly assume uniform preferences, we curate open-ended promp
ts paired with many high-quality answers over which users would be expected to d
isplay heterogeneous latent preferences. Instead of persona prompting LLMs based
 on high-level attributes (e.g., user race or response length), which yields hom
ogeneous preferences relative to humans, we develop a method that can simulate a
 large user base with diverse preferences from a set of pre-trained reward model
s. Our dataset and generated personalities offer an innovative testbed for devel
oping personalization algorithms that grapple with continual data sparsity---few
 relevant feedback from the particular user---by leveraging historical data from
 other (similar) users. We explore basic in-context learning and meta-learning b
aselines to illustrate the utility of PersonalLLM and highlight the need for fut
ure methodological development.
****

Teng Xiao,Yige Yuan,Mingxiao Li,Zhengyu Chen,Vasant G Honavar
On a Connection Between Imitation Learning and RLHF
This work studies the alignment of large language models with preference data fr
om an imitation learning perspective. We establish a close theoretical connecti
on between reinforcement learning from human feedback RLHF and imitation learnin
g (IL), revealing that RLHF implicitly performs imitation learning on the prefer
ence data distribution. Building on this connection, we propose DIL, a principle
d framework that directly optimizes the imitation learning objective. DIL provid
es a unified imitation learning perspective on alignment, encompassing existing

alignment algorithms as special cases while naturally introducing new variants.
By bridging IL and RLHF, DIL offers new insights into alignment with RLHF. Exten
sive experiments demonstrate that DIL outperforms existing methods on various ch
allenging benchmarks.
****
Julien Hermant,Marien Renaud,Jean-François Aujol,Charles Dossal,Aude Rondepierre
Gradient correlation is a key ingredient to accelerate SGD with momentum
Empirically, it has been observed that adding momentum to Stochastic Gradient De
scent (SGD) accelerates the convergence of the algorithm.
However, the literature has been rather pessimistic, even in the case of convex
functions, about the possibility of theoretically proving this observation.
We investigate the possibility of obtaining accelerated convergence of the Stoch
astic Nesterov Accelerated Gradient (SNAG), a momentum-based version of SGD, whe
n minimizing a sum of functions in a convex setting.
We demonstrate that the average correlation between gradients allows to verify t
he strong growth condition, which is the key ingredient to obtain acceleration w
ith SNAG.
Numerical experiments, both in linear regression and deep neural network optimiz
ation, confirm in practice our theoretical results.
****
Jiamu Zheng,Jinghuai Zhang,Tianyu Du,Xuhong Zhang,Jianwei Yin,Tao Lin
CollabEdit: Towards Non-destructive Collaborative Knowledge Editing
Collaborative learning of large language models (LLMs) has emerged as a
new paradigm for utilizing private data from different parties to guarantee
ef■ciency and privacy. Meanwhile, Knowledge Editing (KE) for LLMs has also
garnered increased attention due to its ability to manipulate the behaviors of
LLMs explicitly, yet leaves the collaborative KE case—in which knowledge
edits of multiple parties are aggregated in a privacy-preserving and continual
manner—unexamined. To this end, this manuscript dives into the ■rst investigatio
n
of collaborative KE, in which we start by carefully identifying the unique
three challenges therein, including knowledge overlap, knowledge con■ict, and
knowledge forgetting. We then propose a non-destructive collaborative KE
framework, COLLABEDIT, which employs a novel model merging mechanism
to mimic the global KE behavior while preventing the severe performance drop.
Extensive experiments on two canonical datasets demonstrate the superiority of
COLLABEDIT compared to other destructive baselines, and results shed light on
addressing three collaborative KE challenges and future applications. Our code i
s
available at [https://github.com/LINs-lab/CollabEdit](https://github.com/LINs-la
b/CollabEdit).
****
Navdeep Kumar,Yashaswini Murthy,Itai Shufaro,Kfir Yehuda Levy,R. Srikant,Shie Ma
nnor
Global Convergence of Policy Gradient in Average Reward MDPs
We present the first comprehensive finite-time global convergence analysis of po
licy gradient for infinite horizon average reward Markov decision processes (MDP
s). Specifically, we focus on ergodic tabular MDPs with finite state and action
spaces. Our analysis shows that the policy gradient iterates converge to the opt
imal policy at a sublinear rate of $O(\frac{1}{T})$, where T represents the nu
mber of iterations. Performance bounds for discounted reward MDPs cannot be easi
ly extended to average reward MDPs as the bounds grow proportional to the fifth
power of the effective horizon. Recent work on such extensions makes a smoothnes
s assumption that has not been verified. Thus, our primary contribution is in pr
oviding the first complete proof that the policy gradient algorithm converges gl
obally for average-reward MDPs, without such an assumption. We also obtain the c
orresponding finite-time performance guarantees. In contrast to the existing dis
counted reward performance bounds, our performance bounds have an explicit depen
dence on constants that capture the complexity of the underlying MDP. Motivated
by this observation, we reexamine and improve the existing performance bounds fo

r discounted reward MDPs. We also present simulations that empirically validate the result.
****
Seonghwan Park,Jaehyeon Jeong,Yongjun Kim,Jaeho Lee,Namhoon Lee
ZIP: An Efficient Zeroth-order Prompt Tuning for Black-box Vision-Language Models
Recent studies have introduced various approaches for prompt-tuning black-box vision-language models, referred to as black-box prompt-tuning (BBPT). While BBPT has demonstrated considerable potential, it is often found that many existing methods require an excessive number of queries (i.e., function evaluations), which poses a significant challenge in real-world scenarios where the number of allowed queries is limited. To tackle this issue, we propose Zeroth-order Intrinsic-dimensional Prompt-tuning (ZIP), a novel approach that enables efficient and robust prompt optimization in a purely black-box setting. The key idea of ZIP is to reduce the problem dimensionality and the variance of zeroth-order gradient estimates, such that the training is done fast with far less queries. We achieve this by re-parameterizing prompts in low-rank representations and designing intrinsic-dimensional clipping of estimated gradients. We evaluate ZIP on 13+ vision-language tasks in standard benchmarks and show that it achieves an average improvement of approximately 6% in few-shot accuracy and 48% in query efficiency compared to the best-performing alternative BBPT methods, establishing a new state of the art. Our ablation analysis further shows that the proposed clipping mechanism is robust and nearly optimal, without the need to manually select the clipping threshold, matching the result of expensive hyperparameter search.
****
Kenji Fukumizu,Taiji Suzuki,Noboru Isobe,Kazusato Oko,Masanori Koyama
Flow matching achieves almost minimax optimal convergence
Flow matching (FM) has gained significant attention as a simulation-free generative model. Unlike diffusion models, which are based on stochastic differential equations, FM employs a simpler approach by solving an ordinary differential equation with an initial condition from a normal distribution, thus streamlining the sample generation process. This paper discusses the convergence properties of FM in terms of the p-Wasserstein distance, a measure of distributional discrepancy. We establish that FM can achieve an almost minimax optimal convergence rate for $1 \leq p \leq 2$, presenting the first theoretical evidence that FM can reach convergence rates comparable to those of diffusion models. Our analysis extends existing frameworks by examining a broader class of mean and variance functions for the vector fields and identifies specific conditions necessary to attain these optimal rates.
****
Tianyu Chen,Shuai Lu,Shan Lu,Yeyun Gong,Chenyuan Yang,Xuheng Li,Md Rakib Hossain Misu,Hao Yu,Nan Duan,Peng CHENG,Fan Yang,Shuvendu K Lahiri,Tao Xie,Lidong Zhou
Automated Proof Generation for Rust Code via Self-Evolution
Ensuring correctness is crucial for code generation. Formal verification offers a
definitive assurance of correctness, but demands substantial human effort in proof
construction and hence raises a pressing need for automation. The primary obstacle lies in the severe lack of data—there is much fewer proofs than code snippets
for Large Language Models (LLMs) to train upon. In this paper, we introduce
SAFE, a framework that overcomes the lack of human-written proofs to enable
automated proof generation of Rust code. SAFE establishes a self-evolving cycle
where data synthesis and fine-tuning collaborate to enhance the model capability,
leveraging the definitive power of a symbolic verifier in telling correct proofs from
incorrect ones. SAFE also re-purposes the large number of synthesized incorrect
proofs to train the self-debugging capability of the fine-tuned models, empowering

them to fix incorrect proofs based on the verifier’s feedback. SAFE demonstrates superior efficiency and precision compared to GPT-4o. Through tens of thousands of synthesized proofs and the self-debugging mechanism, we improve the capability of open-source models, initially unacquainted with formal verification, to automatically write proofs for Rust code. This advancement leads to a significant improvement in performance, achieving a 52.52% accuracy rate in a benchmark crafted by human experts, a significant leap over GPT-4o’s performance of 14.39%.

****

Karish Grover,Haiyang Yu,Xiang song,Qi Zhu,Han Xie,Vassilis N. Ioannidis,Christos Faloutsos

Spectro-Riemannian Graph Neural Networks

Can integrating spectral and curvature signals unlock new potential in graph representation learning? Non-Euclidean geometries, particularly Riemannian manifolds such as hyperbolic (negative curvature) and spherical (positive curvature), offer powerful inductive biases for embedding complex graph structures like scale-free, hierarchical, and cyclic patterns. Meanwhile, spectral filtering excels at processing signal variations across graphs, making it effective in homophilic and heterophilic settings. Leveraging both can significantly enhance the learned representations. To this end, we propose Spectro-Riemannian Graph Neural Networks (CUSP) - the first graph representation learning paradigm that unifies both CUrvature (geometric) and SPectral insights. CUSP is a mixed-curvature spectral GNN that learns spectral filters to optimize node embeddings in products of constant curvature manifolds (hyperbolic, spherical, and Euclidean). Specifically, CUSP introduces three novel components: (a) Cusp Laplacian, an extension of the traditional graph Laplacian based on Ollivier-Ricci curvature, designed to capture the curvature signals better; (b) Cusp Filtering, which employs multiple Riemannian graph filters to obtain cues from various bands in the eigenspectrum; and (c) Cusp Pooling, a hierarchical attention mechanism combined with a curvature-based positional encoding to assess the relative importance of differently curved substructures in our graph. Empirical evaluation across eight homophilic and heterophilic datasets demonstrates the superiority of CUSP in node classification and link prediction tasks, with a gain of up to 5.3\% over state-of-the-art models.

****

Aditya Bhaskara,Ashok Cutkosky,Ravi Kumar,Manish Purohit

Descent with Misaligned Gradients and Applications to Hidden Convexity

We consider the problem of minimizing a convex objective given access to an oracle that outputs "misaligned" stochastic gradients, where the expected value of the output is guaranteed to be correlated with, but not necessarily equal to the true gradient of the objective. In the case where the misalignment (or bias) of the oracle changes slowly, we obtain an optimization algorithm that achieves the optimum iteration complexity of $\tilde O(\epsilon^{-2})$; for the more general case where the changes need not be slow, we obtain an algorithm with $\tilde O(\epsilon^{-3})$ iteration complexity. As an application of our framework, we consider optimization problems with a "hidden convexity" property, and obtain an algorithm with $O(\epsilon^{-3})$ iteration complexity.

****

Jianrong Lu,Zhiyu Zhu,Junhui Hou

ParaSolver: A Hierarchical Parallel Integral Solver for Diffusion Models

This paper explores the challenge of accelerating the sequential inference process of Diffusion Probabilistic Models (DPMs). We tackle this critical issue from a dynamic systems perspective, in which the inherent sequential nature is transformed into a parallel sampling process. Specifically, we propose a unified framework that generalizes the sequential sampling process of DPMs as solving a system of banded nonlinear equations. Under this generic framework, we reveal that the Jacobian of the banded nonlinear equations system possesses a unit-diagonal structure, enabling further approximation for acceleration. Moreover, we theoretically propose an effective initialization approach for parallel sampling methods.

Finally, we construct \textit{ParaSolver}, a hierarchical parallel sampling tec
hnique that enhances sampling speed without compromising quality. Extensive expe
riments show that ParaSolver achieves up to \textbf{12.1× speedup} in terms of w
all-clock time. The source code is publicly available at https://github.com/Jian
rong-Lu/ParaSolver.git.
****
Junxuan Wang,Xuyang Ge,Wentao Shu,Qiong Tang,Yunhua Zhou,Zhengfu He,Xipeng Qiu
Towards Universality: Studying Mechanistic Similarity Across Language Model Arch
itectures
The hypothesis of \textit{Universality} in interpretability suggests that differ
ent neural networks may converge to
implement similar algorithms on similar tasks. In this work, we investigate two
mainstream architectures
for language modeling, namely Transformers and Mambas, to explore the extent of
their mechanistic similarity.
We propose to use Sparse Autoencoders (SAEs) to isolate interpretable features f
rom these models and show
that most features are similar in these two models. We also validate the correla
tion between feature similarity
and~\univ. We then delve into the circuit-level analysis of Mamba models
and find that the induction circuits in Mamba are structurally analogous to thos
e in Transformers. We also identify a nuanced difference we call \emph{Off-by-On
e motif}: The information of one token is written into the
SSM state in its next position. Whilst interaction between tokens in Transformer
s does not exhibit such trend.
****
Jiajun Fan,Shuaike Shen,Chaoran Cheng,Yuxin Chen,Chumeng Liang,Ge Liu
Online Reward-Weighted Fine-Tuning of Flow Matching with Wasserstein Regularizat
ion
Recent advancements in reinforcement learning (RL) have achieved great success i
n fine-tuning diffusion-based generative models. However, fine-tuning continuous
 flow-based generative models to align with arbitrary user-defined reward functi
ons remains challenging, particularly due to issues such as policy collapse from
 overoptimization and the prohibitively high computational cost of likelihoods i
n continuous-time flows. In this paper, we propose an easy-to-use and theoretica
lly sound RL fine-tuning method, which we term Online Reward-Weighted Conditiona
l Flow Matching with Wasserstein-2 Regularization (ORW-CFM-W2). Our method integ
rates RL into the flow matching framework to fine-tune generative models with ar
bitrary reward functions, without relying on gradients of rewards or filtered da
tasets. By introducing an online reward-weighting mechanism, our approach guides
 the model to prioritize high-reward regions in the data manifold. To prevent po
licy collapse and maintain diversity, we incorporate Wasserstein-2 (W2) distance
 regularization into our method and derive a tractable upper bound for it in flo
w matching, effectively balancing exploration and exploitation of policy optimiz
ation. We provide theoretical analyses to demonstrate the convergence properties
 and induced data distributions of our method, establishing connections with tra
ditional RL algorithms featuring Kullback-Leibler (KL) regularization and offeri
ng a more comprehensive understanding of the underlying mechanisms and learning
behavior of our approach. Extensive experiments on tasks including target image
generation, image compression, and text-image alignment demonstrate the effectiv
eness of our method, where our method achieves optimal policy convergence while
allowing controllable trade-offs between reward maximization and diversity prese
rvation.
****
Ziqi Jiang,Zhen Wang,Long Chen
CLIPDrag: Combining Text-based and Drag-based Instructions for Image Editing
Precise and flexible image editing remains a fundamental challenge in computer v
ision. Based on the modified areas, most editing methods can be divided into two
 main types: global editing and local editing. In this paper, we choose the two
most common editing approaches (\ie text-based editing and drag-based editing) a

nd analyze their drawbacks. Specifically, text-based methods often fail to descr
ibe the desired modifications precisely, while drag-based methods suffer from am
biguity. To address these issues, we proposed \textbf{CLIPDrag}, a novel image e
diting method that is the first to combine text and drag signals for precise and
 ambiguity-free manipulations on diffusion models. To fully leverage these two s
ignals, we treat text signals as global guidance and drag points as local inform
ation. Then we introduce a novel global-local motion supervision method to integ
rate text signals into existing drag-based methods by adapting a pre-trained lan
guage-vision model like CLIP. Furthermore, we also address the problem of slow c
onvergence in CLIPDrag by presenting a fast point-tracking method that enforces
drag points moving toward correct directions. Extensive experiments demonstrate
that CLIPDrag outperforms existing single drag-based methods or text-based metho
ds.
****

Runzhi Tian,Yongyi Mao
Algorithmic Stability Based Generalization Bounds for Adversarial Training
In this paper, we present a novel stability analysis of adversarial training and
 prove generalization upper bounds in terms of an expansiveness property of adve
rsarial perturbations used during training and used for evaluation. These expans
iveness parameters appear not only govern the vanishing rate of the generalizati
on error but also govern its scaling constant. Our proof techniques do not rely
on artificial assumptions of the adversarial loss, as are typically used in prev
ious works. Our bound attributes the robust overfitting in PGD-based adversarial
 training to the sign function used in the PGD attack, resulting in a bad expans
iveness parameter. The peculiar choice of sign function in the PGD attack appear
s to impact adversarial training both in terms of (inner) optimization and in te
rms of generalization, as shown in this work. This aspect has been largely overl
ooked to date. Going beyond the sign-function based PGD attacks, we further show
 that poor expansiveness properties exist in a wide family of PGD-like iterative
 attack algorithms, which may highlight an intrinsic difficulty in adversarial t
raining.
****

Zirui Zhang,Wei Hao,Aroon Sankoh,William Lin,Emanuel Mendiola-Ortiz,Junfeng Yang
,Chengzhi Mao
I Can Hear You: Selective Robust Training for Deepfake Audio Detection
Recent advances in AI-generated voices have intensified the challenge of detecti
ng deepfake audio, posing risks for scams and the spread of disinformation. To t
ackle this issue, we establish the largest public voice dataset to date, named D
eepFakeVox-HQ, comprising 1.3 million samples, including 270,000 high-quality de
epfake samples from 14 diverse sources. Despite previously reported high accurac
y, existing deepfake voice detectors struggle with our diversely collected datas
et, and their detection success rates drop even further under realistic corrupti
ons and adversarial attacks. We conduct a holistic investigation into factors th
at enhance model robustness and show that incorporating a diversified set of voi
ce augmentations is beneficial. Moreover, we find that the best detection models
 often rely on high-frequency features, which are imperceptible to humans and ca
n be easily manipulated by an attacker. To address this, we propose the F-SAT: F
requency-Selective Adversarial Training method focusing on high-frequency compon
ents. Empirical results demonstrate that using our training dataset boosts basel
ine model performance (without robust training) by 33%, and our robust training
further improves accuracy by 7.7% on clean samples and by 29.3% on corrupted and
 attacked samples, over the state-of-the-art RawNet3 model.
****

Zeru Shi,Kai Mei,Mingyu Jin,Yongye Su,Chaoji Zuo,Wenyue Hua,Wujiang Xu,Yujie Ren
,Zirui Liu,Mengnan Du,Dong Deng,Yongfeng Zhang
From Commands to Prompts: LLM-based Semantic File System for AIOS
Large language models (LLMs) have demonstrated significant potential in the deve
lopment of intelligent LLM-based agents. However, when users use these agent app
lications to perform file operations, their interaction with the file system sti
ll remains the traditional paradigm: reliant on manual navigation through precis

e commands. This paradigm poses a bottleneck to the usability of these systems a s users are required to navigate complex folder hierarchies and remember cryptic file names. To address this limitation, we propose an LLM-based Semantic File S ystem (LSFS) for prompt-driven file management in LLM Agent Operating System (AI OS). Unlike conventional approaches, LSFS incorporates LLMs to enable users or a gents to interact with files through natural language prompts, facilitating semantic file management. At the macro-level, we develop a comprehensive API set to achieve semantic file management functionalities, such as semantic file retr ieval, file update summarization, and semantic file rollback). At the micro-leve l, we store files by constructing semantic indexes for them, design and implemen t syscalls of different semantic operations, e.g., CRUD (create, read, update, d elete),
group by, join. Our experiments show that LSFS can achieve at least 15% retrieva l accuracy improvement with 2.1× higher retrieval speed in the semantic file ret rieval task compared with the traditional file system. In the traditional keywor d-based file retrieval task (i.e., retrieving by string-matching), LSFS also per forms stably well, i.e., over 89% F1-score with improved usability, especially w hen the keyword conditions become more complex. Additionally, LSFS supports more advanced file management operations, i.e., semantic file rollback and file shar ing and achieves 100% success rates in these tasks, further suggesting the capab ility of LSFS . The code is available at https://github.com/agiresearch/AIOS-LSF S.
****

Bryan Bo Cao,Lawrence O'Gorman,Michael Coss,Shubham Jain
Few-Class Arena: A Benchmark for Efficient Selection of Vision Models and Datase t Difficulty Measurement
We propose Few-Class Arena (FCA), as a unified benchmark with focus on testing e fficient image classification models for few classes. A wide variety of benchmar k datasets with many classes (80-1000) have been created to assist Computer Visi on architectural evolution. An increasing number of vision models are evaluated with these many-class datasets. However, real-world applications often involve s ubstantially fewer classes of interest (2-10). This gap between many and few cla sses makes it difficult to predict performance of the few-class applications usi ng models trained on the available many-class datasets. To date, little has been offered to evaluate models in this Few-Class Regime. We conduct a systematic ev aluation of the ResNet family trained on ImageNet subsets from 2 to 1000 classes , and test a wide spectrum of Convolutional Neural Networks and Transformer arch itectures over ten datasets by using our newly proposed FCA tool. Furthermore, t o aid an up-front assessment of dataset difficulty and a more efficient selectio n of models, we incorporate a difficulty measure as a function of class similari ty. FCA offers a new tool for efficient machine learning in the Few-Class Regime , with goals ranging from a new efficient class similarity proposal, to lightwei ght model architecture design, to a new scaling law. FCA is user-friendly and ca n be easily extended to new models and datasets, facilitating future research wo rk. Our benchmark is available at https://github.com/bryanbocao/fca.
****

Parham Rezaei,Farzan Farnia,Cheuk Ting Li
Be More Diverse than the Most Diverse: Optimal Mixtures of Generative Models via Mixture-UCB Bandit Algorithms
The availability of multiple training algorithms and architectures for generativ e models requires a selection mechanism to form a single model over a group of w ell-trained generation models. The selection task is commonly addressed by ident ifying the model that maximizes an evaluation score based on the diversity and q uality of the generated data. However, such a best-model identification approach overlooks the possibility that a mixture of available models can outperform eac h individual model. In this work, we numerically show that a mixture of generati ve models on benchmark image datasets can indeed achieve a better evaluation sco re (based on FID and KID scores), compared to the individual models. This observ ation motivates the development of efficient algorithms for selecting the optima l mixture of the models. To address this, we formulate a quadratic optimization

problem to find an optimal mixture model achieving the maximum of kernel-based e
valuation scores including kernel inception distance (KID) and Rényi kernel entr
opy (RKE). To identify the optimal mixture of the models using the fewest possib
le sample queries, we view the selection task as a multi-armed bandit (MAB) prob
lem and propose the *Mixture Upper Confidence Bound (Mixture-UCB)* algorithm tha
t provably converges to the optimal mixture of the involved models. More broadly
, the proposed Mixture-UCB can be extended to optimize every convex quadratic fu
nction of the mixture weights in a general MAB setting. We prove a regret bound
for the Mixture-UCB algorithm and perform several numerical experiments to show
 the success of Mixture-UCB in finding the optimal mixture of text and image gen
erative models. The project code is available in the [Mixture-UCB Github reposit
ory](https://github.com/Rezaei-Parham/Mixture-UCB).
****
Vishwajeet Agrawal,Rattana Pukdee,Maria Florina Balcan,Pradeep Kumar Ravikumar
Learning from weak labelers as constraints
We study programmatic weak supervision, where in contrast to labeled data, we ha
ve access to \emph{weak labelers}, each of which either abstains or provides noi
sy labels corresponding to any input. Most previous approaches typically employ
latent generative models that model the joint distribution of the weak labels an
d the latent ``true'' label. The caveats are that this relies on assumptions tha
t may not always hold in practice such as conditional independence assumptions o
ver the joint distribution of the weak labelers and the latent true label, and m
ore general implicit inductive biases in the latent generative models. In this w
ork, we consider a more explicit form of side-information that can be leveraged
to denoise the weak labeler, namely the bounds on the average error of the weak
labelers. We then propose a novel but natural weak supervision objective that mi
nimizes a regularization functional subject to satisfying these bounds. This tur
ns out to be a difficult constrained optimization problem due to discontinuous a
ccuracy bound constraints. We provide a continuous optimization formulation for
this objective through an alternating minimization algorithm that iteratively co
mputes soft pseudo labels on the unlabeled data satisfying the constraints while
 being close to the model, and then updates the model on these labels until all
the constraints are satisfied. We follow this with a theoretical analysis of thi
s approach and provide insights into its denoising effects in training discrimin
ative models given multiple weak labelers. Finally, we demonstrate the superior
performance and robustness of our method on a popular weak supervision benchmark
.
****
Botong Zhang,Shuo Li,Osbert Bastani
Conformal Structured Prediction
Conformal prediction has recently emerged as a promising strategy for quantifyin
g the uncertainty of a predictive model; these algorithms modify the model to ou
tput sets of labels that are guaranteed to contain the true label with high prob
ability. However, existing conformal prediction algorithms have largely targeted
 classification and regression settings, where the structure of the prediction s
et has a simple form as a level set of the scoring function. However, for comple
x structured outputs such as text generation, these prediction sets might includ
e a large number of labels and therefore be hard for users to interpret. In this
 paper, we propose a general framework for conformal prediction in the structure
d prediction setting, that modifies existing conformal prediction algorithms to
output structured prediction sets that implicitly represent sets of labels. In a
ddition, we demonstrate how our approach can be applied in domains where the pre
diction sets can be represented as a set of nodes in a directed acyclic graph; f
or instance, for hierarchical labels such as image classification, a prediction
set might be a small subset of coarse labels implicitly representing the predict
ion set of all their more fine-descendants. We demonstrate how our algorithm can
 be used to construct prediction sets that satisfy a desired coverage guarantee
in several domains.
****
Howard Yen,Tianyu Gao,Minmin Hou,Ke Ding,Daniel Fleischer,Peter Izsak,Moshe Wass

erblat,Danqi Chen
HELMET: How to Evaluate Long-context Models Effectively and Thoroughly
Many benchmarks exist for evaluating long-context language models (LCLMs), yet d
evelopers often rely on synthetic tasks such as needle-in-a-haystack (NIAH) or a
n arbitrary subset of tasks. However, it remains unclear whether these benchmark
s reflect the diverse downstream applications of LCLMs, and such inconsistencies
 further complicate model comparison. We investigate the underlying reasons behi
nd these practices and find that existing benchmarks often provide noisy signals
 due to limited coverage of applications, insufficient context lengths, unreliab
le metrics, and incompatibility with base models. In this work, we introduce HEL
MET (How to Evaluate Long-context Models Effectively and Thoroughly), a comprehe
nsive benchmark encompassing seven diverse, application-centric categories. We a
lso address several issues in previous benchmarks by adding controllable lengths
 up to 128K tokens, model-based evaluation for reliable metrics, and few-shot pr
ompting for robustly evaluating base models. Consequently, we demonstrate that H
ELMET offers more reliable and consistent rankings of frontier LCLMs. Through a
comprehensive study of 59 LCLMs, we find that (1) synthetic tasks like NIAH do n
ot reliably predict downstream performance; (2) the diverse categories in HELMET
 exhibit distinct trends and low correlations with each other; and (3) while mos
t LCLMs achieve perfect NIAH scores, open-source models significantly lag behind
 closed ones when tasks require full-context reasoning or following complex inst
ructions---the gap widens as length increases. Finally, we recommend using our R
AG tasks for fast model development, as they are easy to run and better predict
other downstream performance; ultimately, we advocate for a holistic evaluation
across diverse tasks.
****
Yifan Feng,Chengwu Yang,Xingliang Hou,Shaoyi Du,Shihui Ying,Zongze Wu,Yue Gao
Beyond Graphs: Can Large Language Models Comprehend Hypergraphs?
Existing benchmarks like NLGraph and GraphQA evaluate LLMs on graphs by focusing
 mainly on pairwise relationships, overlooking the high-order correlations found
 in real-world data. Hypergraphs, which can model complex beyond-pairwise relati
onships, offer a more robust framework but are still underexplored in the contex
t of LLMs. To address this gap, we introduce LLM4Hypergraph, the first comprehen
sive benchmark comprising 21,500 problems across eight low-order, five high-orde
r, and two isomorphism tasks, utilizing both synthetic and real-world hypergraph
s from citation networks and protein structures. We evaluate six prominent LLMs,
 including GPT-4o, demonstrating our benchmark’s effectiveness in identifying mo
del strengths and weaknesses. Our specialized prompt- ing framework incorporates
 seven hypergraph languages and introduces two novel techniques, Hyper-BAG and H
yper-COT, which enhance high-order reasoning and achieve an average 4% (up to 9%
) performance improvement on structure classification tasks. This work establish
es a foundational testbed for integrating hypergraph computational capabilities
into LLMs, advancing their comprehension.
****
Noga Mudrik,Ryan Ly,Oliver Ruebel,Adam Shabti Charles
CREIMBO: Cross-Regional Ensemble Interactions in Multi-view Brain Observations
Modern recordings of neural activity provide diverse observations of neurons acr
oss brain areas, behavioral conditions, and subjects; presenting an exciting opp
ortunity to reveal the fundamentals of brain-wide dynamics. Current analysis met
hods, however, often fail to fully harness the richness of such data, as they pr
ovide either uninterpretable representations (e.g., via deep networks) or oversi
mplify models (e.g., by assuming stationary dynamics or analyzing each session i
ndependently). Here, instead of regarding asynchronous neural recordings that la
ck alignment in neural identity or brain areas as a limitation, we leverage thes
e diverse views into the brain to learn a unified model of neural dynamics. Spec
ifically, we assume that brain activity is driven by multiple hidden global sub-
circuits. These sub-circuits represent global basis interactions between neural
ensembles—functional groups of neurons—such that the time-varying decomposition
of these sub-circuits defines how the ensembles' interactions evolve over time n
on-stationarily and non-linearly.

We discover the neural ensembles underlying non-simultaneous observations, along with their non-stationary evolving interactions, with our new model, **CREIMBO** (**C**ross-**R**egional **E**nsemble **I**nteractions in **M**ulti-view **B**rain **O**bservations). CREIMBO identifies the hidden composition of per-session neural ensembles through novel graph-driven dictionary learning and models the ensemble dynamics on a low-dimensional manifold spanned by a sparse time-varying composition of the global sub-circuits. Thus, CREIMBO disentangles overlapping temporal neural processes while preserving interpretability due to the use of a shared underlying sub-circuit basis. Moreover, CREIMBO distinguishes session-specific computations from global (session-invariant) ones by identifying session covariates and variations in sub-circuit activations. We demonstrate CREIMBO's ability to recover true components in synthetic data, and uncover meaningful brain dynamics in human high-density electrode recordings, including cross-subject neural mechanisms as well as inter- vs. intra-region dynamical motifs. Furthermore, using mouse whole-brain recordings, we show CREIMBO's ability to discover dynamical interactions that capture task and behavioral variables and meaningfully align with the biological importance of the brain areas they represent.
****

Devon Jarvis,Richard Klein,Benjamin Rosman,Andrew M Saxe
Make Haste Slowly: A Theory of Emergent Structured Mixed Selectivity in Feature Learning ReLU Networks
In spite of finite dimension ReLU neural networks being a consistent factor behind recent deep learning successes, a theory of feature learning in these models remains elusive. Currently, insightful theories still rely on assumptions including the linearity of the network computations, unstructured input data and architectural constraints such as infinite width or a single hidden layer. To begin to address this gap we establish an equivalence between ReLU networks and Gated Deep Linear Networks, and use their greater tractability to derive dynamics of learning. We then consider multiple variants of a core task reminiscent of multi-task learning or contextual control which requires both feature learning and nonlinearity. We make explicit that, for these tasks, the ReLU networks possess an inductive bias towards latent representations which are *not* strictly modular or disentangled but are still highly structured and reusable between contexts. This effect is amplified with the addition of more contexts and hidden layers. Thus, we take a step towards a theory of feature learning in finite ReLU networks and shed light on how structured mixed-selective latent representations can emerge due to a bias for node-reuse and learning speed.
****

Daniel Rose,Oliver Wieder,Thomas Seidel,Thierry Langer
PharmacoMatch: Efficient 3D Pharmacophore Screening via Neural Subgraph Matching
The increasing size of screening libraries poses a significant challenge for the development of virtual screening methods for drug discovery, necessitating a re-evaluation of traditional approaches in the era of big data. Although 3D pharmacophore screening remains a prevalent technique, its application to very large datasets is limited by the computational cost associated with matching query pharmacophores to database molecules. In this study, we introduce PharmacoMatch, a novel contrastive learning approach based on neural subgraph matching. Our method reinterprets pharmacophore screening as an approximate subgraph matching problem and enables efficient querying of conformational databases by encoding query-target relationships in the embedding space. We conduct comprehensive investigations of the learned representations and evaluate PharmacoMatch as pre-screening tool in a zero-shot setting. We demonstrate significantly shorter runtimes and comparable performance metrics to existing solutions, providing a promising speed-up for screening very large datasets.
****

Jingyang Li,Jiachun Pan,Vincent Y. F. Tan,Kim-chuan Toh,Pan Zhou
Towards Understanding Why FixMatch Generalizes Better Than Supervised Learning
Semi-supervised learning (SSL), exemplified by FixMatch (Sohn et al., 2020), has shown significant generalization advantages over supervised learning (SL), particularly in the context of deep neural networks (DNNs). However, it is still unc

lear, from a theoretical standpoint, why FixMatch-like SSL algorithms generalize better than SL on DNNs. In this work, we present the first theoretical justification for the enhanced test accuracy observed in FixMatch-like SSL applied to DNNs by taking convolutional neural networks (CNNs) on classification tasks as an example. Our theoretical analysis reveals that the semantic feature learning processes in FixMatch and SL are rather different. In particular, FixMatch learns all the discriminative features of each semantic class, while SL only randomly captures a subset of features due to the well-known lottery ticket hypothesis. Furthermore, we show that our analysis framework can be applied to other FixMatch-like SSL methods, e.g., FlexMatch, FreeMatch, Dash, and SoftMatch. Inspired by our theoretical analysis, we develop an improved variant of FixMatch, termed Semantic-Aware FixMatch (SA-FixMatch). Experimental results corroborate our theoretical findings and the enhanced generalization capability of SA-FixMatch.
****
Zhenyu Zhu,Fanghui Liu,Volkan Cevher
How Gradient descent balances features: A dynamical analysis for two-layer neural networks
This paper investigates the fundamental regression task of learning k neurons (\emph{a.k.a.} teachers) from Gaussian input, using two-layer ReLU neural networks with width m (\emph{a.k.a.} students) and $m, k= \mathcal{O}(1)$, trained via gradient descent under proper initialization and a small step-size. Our analysis follows a three-phase structure: \emph{alignment} after weak recovery, \emph{tangential growth}, and \emph{local convergence}, providing deeper insights into the learning dynamics of gradient descent (GD). We prove the global convergence at the rate of $\mathcal{O}(T^{-3})$ for the zero loss of excess risk. Additionally, our results show that GD automatically groups and balances student neurons, revealing an implicit bias toward achieving the minimum ``balanced'' ℓ_2-norm in the solution. Our work extends beyond previous studies in exact-parameterization setting ($m = k = 1$, (Yehudai and Ohad, 2020)) and single-neuron setting ($m \geq k = 1$, (Xu and Du, 2023)). The key technical challenge lies in handling the interactions between multiple teachers and students during training, which we address by refining the alignment analysis in Phase 1 and introducing a new dynamic system analysis for tangential components in Phase 2. Our results pave the way for further research on optimizing neural network training dynamics and understanding implicit biases in more complex architectures.
****
Xin Lin,Shi Luo,Xiaojun Shan,Xiaoyu Zhou,Chao Ren,Lu Qi,Ming-Hsuan Yang,Nuno Vasconcelos
HQGS: High-Quality Novel View Synthesis with Gaussian Splatting in Degraded Scenes
3D Gaussian Splatting (3DGS) has shown promising results for Novel View Synthesis. However, while it is quite effective when based on high-quality images, its performance declines as image quality degrades, due to lack of resolution, motion blur, noise, compression artifacts, or other factors common in real-world data collection. While some solutions have been proposed for specific types of degradation, general techniques are still missing. To address the problem, we propose a robust HQGS that significantly enhances the 3DGS under various degradation scenarios. We first analyze that 3DGS lacks sufficient attention in some detailed regions in low-quality scenes, leading to the absence of Gaussian primitives in those areas and resulting in loss of detail in the rendered images. To address this issue, we focus on leveraging edge structural information to provide additional guidance for 3DGS, enhancing its robustness. First, we introduce an edge-semantic fusion guidance module that combines rich texture information from high-frequency edge-aware maps with semantic information from images. The fused features serve as prior guidance to capture detailed distribution across different regions, bringing more attention to areas with detailed edge information and allowing for a higher concentration of Gaussian primitives to be assigned to such areas. Additionally, we present a structural cosine similarity loss to complement pixel-level constraints, further improving the quality of the rendered images. Extensive experiments demonstrate that our method offers better robustness and achie

ves the best results across various degraded scenes. Source code and trained mod
els are publicly available at: \url{https://github.com/linxin0/HQGS}.
****
Hadar Sharvit,Raz Karl,Tsevi Beatus
A deep inverse-mapping model for a flapping robotic wing
In systems control, the dynamics of a system are governed by modulating its inpu
ts to achieve a desired outcome. For example, to control the thrust of a quad-co
pter propeller the controller modulates its rotation rate, relying on a straight
forward mapping between the input rotation rate and the resulting thrust. This m
apping can be inverted to determine the rotation rate needed to generate a desir
ed thrust. However, in complex systems, such as flapping-wing robots where intri
cate fluid motions are involved, mapping inputs (wing kinematics) to outcomes (a
erodynamic forces) is nontrivial and inverting this mapping for real-time contro
l is computationally impractical. Here, we report a machine-learning solution fo
r the inverse mapping of a flapping-wing system based on data from an experiment
al system we have developed. Our model learns the input wing motion required to
generate a desired aerodynamic force outcome. We used a sequence-to-sequence mod
el tailored for time-series data and augmented it with a novel adaptive-spectrum
 layer that implements representation learning in the frequency domain. To train
 our model, we developed a flapping wing system that simultaneously measures the
 wing's aerodynamic force and its 3D motion using high-speed cameras. We demonst
rate the performance of our system on an additional open-source dataset of a fla
pping wing in a different flow regime. Results show superior performance compare
d with more complex state-of-the-art transformer-based models, with 11\% improve
ment on the test datasets median loss. Moreover, our model shows superior infere
nce time, making it practical for onboard robotic control. Our open-source data
and framework may improve modeling and real-time control of systems governed by
complex dynamics, from biomimetic robots to biomedical devices.
****
Melanie Sclar,Jane Dwivedi-Yu,Maryam Fazel-Zarandi,Yulia Tsvetkov,Yonatan Bisk,Y
ejin Choi,Asli Celikyilmaz
Explore Theory of Mind: program-guided adversarial data generation for theory of
 mind reasoning
Do large language models (LLMs) have theory of mind? A plethora of papers and be
nchmarks have been introduced to evaluate if current models have been able to de
velop this key ability of social intelligence. However, all rely on limited data
sets with simple patterns that can potentially lead to problematic blind spots i
n evaluation and an overestimation of model capabilities. We introduce ExploreTo
M, the first framework to allow large-scale generation of diverse and challengin
g theory of mind data for robust training and evaluation. Our approach leverages
 an A* search over a custom domain-specific language to produce complex story st
ructures and novel, diverse, yet plausible scenarios to stress test the limits o
f LLMs. Our evaluation reveals that state-of-the-art LLMs, such as Llama-3.1-70B
 and GPT-4o, show accuracies as low as 0% and 9% on ExploreToM-generated data, h
ighlighting the need for more robust theory of mind evaluation. As our generatio
ns are a conceptual superset of prior work, fine-tuning on our data yields a 27-
point accuracy improvement on the classic ToMi benchmark (Le et al., 2019). Expl
oreToM also enables uncovering underlying skills and factors missing for models
to show theory of mind, such as unreliable state tracking or data imbalances, wh
ich may contribute to models' poor performance on benchmarks.
****
Yiqun Sun,Qiang Huang,Yixuan Tang,Anthony Kum Hoe Tung,Jun Yu
A General Framework for Producing Interpretable Semantic Text Embeddings
Semantic text embedding is essential to many tasks in Natural Language Processin
g (NLP). While black-box models are capable of generating high-quality embedding
s, their lack of interpretability limits their use in tasks that demand transpar
ency. Recent approaches have improved interpretability by leveraging domain-expe
rt-crafted or LLM-generated questions, but these methods rely heavily on expert
input or well-prompt design, which restricts their generalizability and ability
to generate discriminative questions across a wide range of tasks. To address th

ese challenges, we introduce \algo{CQG-MBQA} (Contrastive Question Generation - Multi-task Binary Question Answering), a general framework for producing interpretable semantic text embeddings across diverse tasks. Our framework systematically generates highly discriminative, low cognitive load yes/no questions through the \algo{CQG} method and answers them efficiently with the \algo{MBQA} model, resulting in interpretable embeddings in a cost-effective manner. We validate the effectiveness and interpretability of \algo{CQG-MBQA} through extensive experiments and ablation studies, demonstrating that it delivers embedding quality comparable to many advanced black-box models while maintaining inherently interpretability. Additionally, \algo{CQG-MBQA} outperforms other interpretable text embedding methods across various downstream tasks. The source code is available at \url{https://github.com/dukesun99/CQG-MBQA}.
****

Chandra Sekhar Mukherjee,Jiapeng Zhang
Balanced Ranking with Relative Centrality: A multi-core periphery perspective
Ranking of vertices in a graph for different objectives is one of the most fundamental tasks in computer science. It is known that traditional ranking algorithms can generate unbalanced ranking when the graph has underlying communities, resulting in loss of information, polarised opinions, and reduced diversity (Celis, Straszak \& Vishnoi [ICALP 2018]).

In this paper, we focus on *unsupervised ranking* on graphs and observe that popular centrality measure based ranking algorithms such as PageRank may often generate unbalanced ranking here as well. We address this issue by coining a new approach, which we term *relative centrality*. Our approach is based on an iterative graph-dependent local normalization of the centrality score, which promotes balancedness while maintaining the validity of the ranking.

We further quantify reasons behind this unbalancedness of centrality measures on a novel structure that we propose is called multi-core-periphery with communities (MCPC). We also provide theoretical and extensive simulation support for our approach towards resolving the unbalancedness in MCPC.

Finally, we consider graph embeddings of 11 single-cell datasets. We observe that top-ranked as per existing centrality measures are better separable into the ground truth communities. However, due to the unbalanced ranking, the top nodes often do not contain points from some communities. Here, our relative-centrality-based approach generates a ranking that provides a similar improvement in clusterability while providing significantly higher balancedness.
****

Jiashun Liu,Johan Samir Obando Ceron,Aaron Courville,Ling Pan
Neuroplastic Expansion in Deep Reinforcement Learning
The loss of plasticity in learning agents, analogous to the solidification of neural pathways in biological brains, significantly impedes learning and adaptation in reinforcement learning due to its non-stationary nature. To address this fundamental challenge, we propose a novel approach, *Neuroplastic Expansion* (NE), inspired by cortical expansion in cognitive science. NE maintains learnability and adaptability throughout the entire training process by dynamically growing the network from a smaller initial size to its full dimension. Our method is designed with three key components: (1) elastic neuron generation based on potential gradients, (2) dormant neuron pruning to optimize network expressivity, and (3) neuron consolidation via experience review to strike a balance in the plasticity-stability dilemma. Extensive experiments demonstrate that NE effectively mitigates plasticity loss and outperforms state-of-the-art methods across various tasks in MuJoCo and DeepMind Control Suite environments. NE enables more adaptive learning in complex, dynamic environments, which represents a crucial step towards transitioning deep reinforcement learning from static, one-time training paradigms to more flexible, continually adapting models.
****

Maresa Schröder,Valentyn Melnychuk,Stefan Feuerriegel

Differentially private learners for heterogeneous treatment effects
Patient data is widely used to estimate heterogeneous treatment effects and unde
rstand the effectiveness and safety of drugs. Yet, patient data includes highly
sensitive information that must be kept private. In this work, we aim to estimat
e
the conditional average treatment effect (CATE) from observational data under
differential privacy. Specifically, we present DP-CATE, a novel framework for
CATE estimation that is *Neyman-orthogonal* and ensures *differential privacy* o
f the estimates.
Our framework is highly general: it applies to any two-stage
CATE meta-learner with a Neyman-orthogonal loss function and any machine
learning model can be used for nuisance estimation. We further provide an extens
ion of our DP-CATE, where we employ RKHS regression to release the complete
CATE function while ensuring differential privacy. We demonstrate the effectiven
ess of DP-CATE across various experiments using synthetic and real-world
datasets. To the best of our knowledge, we are the first to provide a framework
for
CATE estimation that is doubly robust and differentially private.
****
Federico Arangath Joseph,Jerome Sieber,Melanie Zeilinger,Carmen Amo Alonso
Lambda-Skip Connections: the architectural component that prevents Rank Collapse
Rank collapse, a phenomenon where embedding vectors in sequence models
rapidly converge to a uniform token or equilibrium state, has recently gained at
-
tention in the deep learning literature. This phenomenon leads to reduced expres
-
sivity and potential training instabilities due to vanishing gradients. Empirica
l ev-
idence suggests that architectural components like skip connections, LayerNorm,
and MultiLayer Perceptrons (MLPs) play critical roles in mitigating rank collaps
e.
While this issue is well-documented for transformers, alternative sequence mod-
els, such as State Space Models (SSMs), which have recently gained prominence,
have not been thoroughly examined for similar vulnerabilities. This paper extend
s
the theory of rank collapse from transformers to SSMs using a unifying frame-
work that captures both architectures. We introduce a modification in the skip
connection component, termed lambda-skip connections, that provides guaran-
tees for rank collapse prevention. We present, via analytical results, a suffici
ent
condition to achieve the guarantee for all of the aforementioned architectures.
We
also study the necessity of this condition via ablation studies and analytical e
xam-
ples. To our knowledge, this is the first study that provides a general guarante
e to
prevent rank collapse, and that investigates rank collapse in the context of SSM
s,
offering valuable understanding for both theoreticians and practitioners. Finall
y,
we validate our findings with experiments demonstrating the crucial role of arch
i-
tectural components in preventing rank collapse.
****
Muhammed Emrullah Ildiz,Halil Alperen Gozeten,Ege Onur Taga,Marco Mondelli,Samet
Oymak
High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalizati
on and Scaling Laws
A growing number of machine learning scenarios rely on knowledge distillation wh
ere one uses the output of a surrogate model as labels to supervise the training

of a target model. In this work, we provide a sharp characterization of this pr
ocess for ridgeless, high-dimensional regression, under two settings: *(i)* mode
l shift, where the surrogate model is arbitrary, and *(ii)* distribution shift,
where the surrogate model is the solution of empirical risk minimization with ou
t-of-distribution data. In both cases, we characterize the precise risk of the t
arget model through non-asymptotic bounds in terms of sample size and data distr
ibution under mild conditions. As a consequence, we identify the form of the opt
imal surrogate model, which reveals the benefits and limitations of discarding w
eak features in a data-dependent fashion. In the context of weak-to-strong (W2S)
generalization, this has the interpretation that *(i)* W2S training, with the s
urrogate as the weak model, can provably outperform training with strong labels
under the same data budget, but *(ii)* it is unable to improve the data scaling
law. We validate our results on numerical experiments both on ridgeless regressi
on and on neural network architectures.
****
Jaeseong Lee,Taewoong Kang,Marcel Buehler,Min-Jung Kim,Sungwon Hwang,Junha Hyung
,Hyojin Jang,Jaegul Choo
SurFhead: Affine Rig Blending for Geometrically Accurate 2D Gaussian Surfel Head
Avatars
Recent advancements in head avatar rendering using Gaussian primitives have achi
eved significantly high-fidelity results. Although precise head geometry is cruc
ial for applications like mesh reconstruction and relighting, current methods st
ruggle to capture intricate geometric details and render unseen poses due to the
ir reliance on similarity transformations, which cannot handle stretch and shear
transforms essential for detailed deformations of geometry. To address this, we
propose SurFhead, a novel method that reconstructs riggable head geometry from
RGB videos using 2D Gaussian surfels, which offer well-defined geometric propert
ies, such as precise depth from fixed ray intersections and normals derived from
their surface orientation, making them advantageous over 3D counterparts. SurFh
ead ensures high-fidelity rendering of both normals and images, even in extreme
poses, by leveraging classical mesh-based deformation transfer and affine transf
ormation interpolation. SurFhead introduces precise geometric deformation and bl
ends surfels through polar decomposition of transformations, including those aff
ecting normals. Our key contribution lies in bridging classical graphics techniq
ues, such as mesh-based deformation, with modern Gaussian primitives, achieving
state-of-the-art geometry reconstruction and rendering quality. Unlike previous
avatar rendering approaches, SurFhead enables efficient reconstruction driven by
Gaussian primitives while preserving high-fidelity geometry.
****
Daniel M. Steinberg,Rafael Oliveira,Cheng Soon Ong,Edwin V. Bonilla
Variational Search Distributions
We develop VSD, a method for conditioning a generative model of discrete, combin
atorial designs on a rare desired class by efficiently evaluating a black-box (e
.g. experiment, simulation) in a batch sequential manner. We call this task acti
ve generation; we formalize active generation's requirements and desiderata, and
formulate a solution via variational inference. VSD uses off-the-shelf gradient
based optimization routines, can learn powerful generative models for desirable
designs, and can take advantage of scalable predictive models. We derive asympt
otic convergence rates for learning the true conditional generative distribution
of designs with certain configurations of our method. After illustrating the ge
nerative model on images, we empirically demonstrate that VSD can outperform exi
sting baseline methods on a set of real sequence-design problems in various prot
ein and DNA/RNA engineering tasks.
****
Jungwon Park,Jungmin Ko,Dongnam Byun,Jangwon Suh,Wonjong Rhee
Cross-Attention Head Position Patterns Can Align with Human Visual Concepts in T
ext-to-Image Generative Models
Recent text-to-image diffusion models leverage cross-attention layers, which hav
e been effectively utilized to enhance a range of visual generative tasks. Howev
er, our understanding of cross-attention layers remains somewhat limited. In thi

s study, we introduce a mechanistic interpretability approach for diffusion mode ls by constructing Head Relevance Vectors (HRVs) that align with human-specified visual concepts. An HRV for a given visual concept has a length equal to the to tal number of cross-attention heads, with each element representing the importan ce of the corresponding head for the given visual concept. To validate HRVs as i nterpretable features, we develop an ordered weakening analysis that demonstrate s their effectiveness. Furthermore, we propose concept strengthening and concept adjusting methods and apply them to enhance three visual generative tasks. Our results show that HRVs can reduce misinterpretations of polysemous words in imag e generation, successfully modify five challenging attributes in image editing, and mitigate catastrophic neglect in multi-concept generation. Overall, our work provides an advancement in understanding cross-attention layers and introduces new approaches for fine-controlling these layers at the head level.
****

Lu Li,Tianyu Zhang,Zhiqi Bu,Suyuchen Wang,Huan He,Jie Fu,Yonghui Wu,Jiang Bian,Y ong Chen,Yoshua Bengio

MAP: Low-compute Model Merging with Amortized Pareto Fronts via Quadratic Approx imation

Model merging has emerged as an effective approach to combining multiple single-task models into a multitask model. This process typically involves computing a weighted average of the model parameters without additional training. Existing m odel-merging methods focus on improving average task accuracy. However, interfer ence and conflicts between the objectives of different tasks can lead to trade-o ffs during the merging process. In real-world applications, a set of solutions w ith various trade-offs can be more informative, helping practitioners make decis ions based on diverse preferences. In this paper, we introduce a novel and low-c ompute algorithm, Model Merging with Amortized Pareto Front (MAP). MAP efficient ly identifies a Pareto set of scaling coefficients for merging multiple models, reflecting the trade-offs involved. It amortizes the substantial computational c ost of evaluations needed to estimate the Pareto front by using quadratic approx imation surrogate models derived from a preselected set of scaling coefficients. Experimental results on vision and natural language processing tasks demonstrat e that MAP can accurately identify the Pareto front, providing practitioners wit h flexible solutions to balance competing task objectives. We also introduce Bay esian MAP for scenarios with a relatively low number of tasks and Nested MAP for situations with a high number of tasks, further reducing the computational cost of evaluation.
****

Sheng Wang,Liheng Chen,Pengan CHEN,Jingwei Dong,Boyang XUE,Jiyue Jiang,Lingpeng Kong,Chuan Wu

MoS: Unleashing Parameter Efficiency of Low-Rank Adaptation with Mixture of Shar ds

The rapid scaling of large language models necessitates more lightweight finetun ing methods to reduce the explosive GPU memory overhead when numerous customized models are served simultaneously.

Targeting more parameter-efficient low-rank adaptation (LoRA), parameter sharing presents a promising solution. Empirically, our research into high-level sharin g principles highlights the indispensable role of differentiation in reversing t he detrimental effects of pure sharing.

Guided by this finding, we propose Mixture of Shards (MoS), incorporating both i nter-layer and intra-layer sharing schemes, and integrating four nearly cost-fre e differentiation strategies, namely subset selection, pair dissociation, vector sharding, and shard privatization. Briefly, it selects a designated number of s hards from global pools with a Mixture-of-Experts (MoE)-like routing mechanism b efore sequentially concatenating them to low-rank matrices.

Hence, it retains all the advantages of LoRA while offering enhanced parameter e fficiency, and effectively circumvents the drawbacks of peer parameter-sharing m ethods.

Our empirical experiments demonstrate approximately $8\times$ parameter savings in a standard LoRA setting. The ablation study confirms the significance of each

component.
Our insights into parameter sharing and MoS method may illuminate future develop
ments of more parameter-efficient finetuning methods.
The code is officially available at https://github.com/Forence1999/MoS.
****
Qiyuan Zhang,Yufei Wang,Tiezheng YU,Yuxin Jiang,Chuhan Wu,Liangyou Li,Yasheng Wa
ng,Xin Jiang,Lifeng Shang,Ruiming Tang,Fuyuan Lyu,Chen Ma
RevisEval: Improving LLM-as-a-Judge via Response-Adapted References
With significant efforts in recent studies, LLM-as-a-Judge has become a cost-eff
ective alternative to human evaluation for assessing text generation quality in
a wide range of tasks. However, there still remains a reliability gap between LL
M-as-a-Judge and human evaluation. One important reason is the lack of guided or
acles in the evaluation process. Motivated by the role of reference pervasively
used in classic text evaluation, we introduce RevisEval, a novel text generation
evaluation paradigm via the response-adapted references. RevisEval is driven by
the key observation that an ideal reference should maintain the necessary relev
ance to the response to be evaluated. Specifically, RevisEval leverages the text
revision capabilities of large language models (LLMs) to adaptively revise the
response, then treat the revised text as the reference (response-adapted referen
ce) for the subsequent evaluation. Extensive experiments demonstrate that RevisE
val outperforms traditional reference-free and reference-based evaluation paradi
gms that use LLM-as-a-Judge across NLG tasks and open-ended instruction-followin
g tasks. More importantly, our response-adapted references can further boost the
classical text metrics, e.g., BLEU and BERTScore, compared to traditional refer
ences and even rival the LLM-as-a-Judge. A detailed analysis is also conducted t
o confirm RevisEval's effectiveness in bias reduction, the impact of inference c
ost, and reference relevance.
****
Sagi Shaier,Francisco Pereira,Katharina von der Wense,Lawrence Hunter,Matt Jones
More Experts Than Galaxies: Conditionally-Overlapping Experts with Biologically-
Inspired Fixed Routing
The evolution of biological neural systems has led to both modularity and sparse
coding, which enables energy efficiency and robustness across the diversity of
tasks in the lifespan. In contrast, standard neural networks rely on dense, non-
specialized architectures, where all model parameters are simultaneously updated
to learn multiple tasks, leading to interference. Current sparse neural network
approaches aim to alleviate this issue but are hindered by limitations such as
1) trainable gating functions that cause representation collapse, 2) disjoint ex
perts that result in redundant computation and slow learning, and 3) reliance on
explicit input or task IDs that limit flexibility and scalability.
In this paper we propose Conditionally Overlapping Mixture of ExperTs (COMET), a
general deep learning method that addresses these challenges by inducing a modu
lar, sparse architecture with an exponential number of overlapping experts. COME
T replaces the trainable gating function used in Sparse Mixture of Experts with
a fixed, biologically inspired random projection applied to individual input rep
resentations. This design causes the degree of expert overlap to depend on input
similarity, so that similar inputs tend to share more parameters. This results
in faster learning per update step and improved out-of-sample generalization.
We demonstrate the effectiveness of COMET on a range of tasks, including image c
lassification, language modeling, and regression, using several popular deep lea
rning architectures.
****
Ben Jourdan,Gregory Schwartzman,Peter Macgregor,He Sun
Coreset Spectral Clustering
Coresets have become an invaluable tool for solving k-means and kernel k-mea
ns clustering problems on large datasets with small numbers of clusters. On the
other hand, spectral clustering works well on sparse graphs and has recently bee
n extended to scale efficiently to large numbers of clusters. We exploit the con
nection between kernel k-means and the normalised cut problem to combine the b
enefits of both. Our main result is a coreset spectral clustering algorithm for

graphs that clusters a coreset graph to infer a good labelling of the original g
raph. We prove that an α-approximation for the normalised cut problem on
the coreset graph is an $O(\alpha)$-approximation on the original. We also impro
ve the running time of the state-of-the-art coreset algorithm for kernel k-mea
ns on sparse kernels, from $\tilde{O}(nk)$ to $\tilde{O}(n\cdot \min (k, d_{avg}
))$, where d_{avg} is the average number of non-zero entries in each row of th
e $n\times n$ kernel matrix. Our experiments confirm our coreset algorithm is as
ymptotically faster on large real-world graphs with many clusters, and show that
 our clustering algorithm overcomes the main challenge faced by coreset kernel $
k$-means on sparse kernels which is getting stuck in local optima.
****
Jingcheng Ke,Jun-cheng Chen,I-hong Jhuo,Chia-Wen Lin,Yen-Yu Lin
Generation and Comprehension Hand-in-Hand: Vision-guided Expression Diffusion fo
r Boosting Referring Expression Generation and Comprehension
Referring expression generation (REG) and comprehension (REC) are vital and comp
lementary in joint visual and textual reasoning. Existing REC datasets typicall
y contain insufficient image-expression pairs for training, hindering the genera
lization of REC models to unseen referring expressions. Moreover, REG methods fr
equently struggle to bridge the visual and textual domains due to the limited ca
pacity, leading to low-quality and restricted diversity in expression generation
. To address these issues, we propose a novel VIsion-guided Expression Diffusion
 Model (VIE-DM) for the REG task, where diverse synonymous expressions adhering
to both image and text contexts of the target object are generated to augment RE
C datasets. VIE-DM consists of a vision-text condition (VTC) module and a transf
ormer decoder. Our VTC and token selection design effectively addresses the feat
ure discrepancy problem prevalent in existing REG methods. This enables us to ge
nerate high-quality, diverse synonymous expressions that can serve as augmented
data for REC model learning. Extensive experiments on five datasets demonstrate
the high quality and large diversity of our generated expressions. Furthermore,
the augmented image-expression pairs consistently enhance the performance of exi
sting REC models, achieving state-of-the-art results.
****
Xinyi Liu,Yujie Wang,Fangcheng Fu,Xupeng Miao,Shenhan Zhu,Xiaonan Nie,Bin CUI
NetMoE: Accelerating MoE Training through Dynamic Sample Placement
Mixture of Experts (MoE) is a widely used technique to expand model sizes for be
tter model quality while maintaining the computation cost constant. In a nutshel
l, an MoE model consists of multiple experts in each model layer and routes the
training tokens to only a fixed number of experts rather than all. In distribute
d training, as experts are distributed among different GPUs, All-to-All communic
ation is necessary to exchange the training tokens among the GPUs after each tim
e of expert routing. Due to the frequent and voluminous data exchanges, All-to-A
ll communication has become a notable challenge to training efficiency.

In this paper, we manage to accelerate All-to-All communication in MoE models fr
om the training sample perspective, which is unexplored so far. In particular, w
e put forward the observation that tokens in the same training sample have certa
in levels of locality in expert routing. Motivated by this, we develop NetMoE, w
hich takes such locality into account and dynamically rearranges the placement o
f training samples to minimize All-to-All communication costs. Specifically, we
model the All-to-All communication given the sample placement and formulate an i
nteger programming problem to deduce the optimal placement in polynomial time. E
xperiments with 32 GPUs show that NetMoE achieves a maximum efficiency improveme
nt of $1.67 \times$ compared with current MoE training frameworks.
****
Omar Shaikh,Michelle S. Lam,Joey Hejna,Yijia Shao,Hyundong Justin Cho,Michael S.
 Bernstein,Diyi Yang
Aligning Language Models with Demonstrated Feedback
Language models are aligned to emulate the collective voice of many, resulting i
n outputs that align with no one in particular. Steering LLMs away from generic
output is possible through supervised finetuning or RLHF, but requires prohibiti

vely large datasets for new ad-hoc tasks. We argue that it is instead possible t
o align an LLM to a specific setting by leveraging a very small number (<10) o
f demonstrations as feedback. Our method, Demonstration ITerated Task Optimizati
on (DITTO), directly aligns language model outputs to a user's demonstrated beha
viors. Derived using ideas from online imitation learning, DITTO cheaply generat
es online comparison data by treating users' demonstrations as preferred over ou
tput from the LLM and its intermediate checkpoints. We evaluate DITTO's ability
to learn fine-grained style and task alignment across domains such as news artic
les, emails, and blog posts. Additionally, we conduct a user study soliciting a
range of demonstrations from participants ($N=16$). Across our benchmarks and us
er study, we find that win-rates for DITTO outperform few-shot prompting, superv
ised fine-tuning, and other self-play methods by an average of 19\% points. By u
sing demonstrations as feedback directly, DITTO offers a novel method for effect
ive customization of LLMs.
****
Shilin Xu,Haobo Yuan,Qingyu Shi,Lu Qi,Jingbo Wang,Yibo Yang,Yining Li,Kai Chen,Y
unhai Tong,Bernard Ghanem,Xiangtai Li,Ming-Hsuan Yang
RMP-SAM: Towards Real-Time Multi-Purpose Segment Anything
Recent segmentation methods, which adopt large-scale data training and transform
er architecture, aim to create one foundation model that can perform multiple ta
sks.
    However, most of these methods rely on heavy encoder and decoder frameworks,
 hindering their performance in real-time scenarios.
    To explore real-time segmentation, recent advancements primarily focus on se
mantic segmentation within specific environments, such as autonomous driving. Ho
wever, they often overlook the generalization ability of these models across div
erse scenarios.
    Therefore, to fill this gap, this work explores a novel real-time segmentati
on setting called real-time multi-purpose segmentation.
    It contains three fundamental sub-tasks: interactive segmentation, panoptic
segmentation, and video instance segmentation.
    Unlike previous methods, which use a specific design for each task, we aim t
o use only a single end-to-end model to accomplish all these tasks in real-time.
    To meet real-time requirements and balance multi-task learning, we present a
 novel dynamic convolution-based method, Real-Time Multi-Purpose SAM (RMP-SAM).
    It contains an efficient encoder and an efficient decoupled adapter to perfo
rm prompt-driven decoding.
    Moreover, we further explore different training strategies and one new adapt
er design to boost co-training performance further.
    We benchmark several strong baselines by extending existing works to support
 our multi-purpose segmentation.
    Extensive experiments demonstrate that RMP-SAM is effective and generalizes
well on proposed benchmarks and other specific semantic tasks.
    Our implementation of RMP-SAM achieves the optimal balance between accuracy
and speed for these tasks. The code is released at
    \url{https://github.com/xushilin1/RAP-SAM}
****
Jixun Yao,Hexin Liu,Chen Chen,Yuchen Hu,EngSiong Chng,Lei Xie
GenSE: Generative Speech Enhancement via Language Models using Hierarchical Mode
ling
Semantic information refers to the meaning conveyed through words, phrases, and
contextual relationships within a given linguistic structure. Humans can leverag
e semantic information, such as familiar linguistic patterns and contextual cues
, to reconstruct incomplete or masked speech signals in noisy environments. Howe
ver, existing speech enhancement (SE) approaches often overlook the rich semanti
c information embedded in speech, which is crucial for improving intelligibility
, speaker consistency, and overall quality of enhanced speech signals. To enrich
 the SE model with semantic information, we employ language models as an efficie
nt semantic learner and propose a comprehensive framework tailored for language
model-based speech enhancement, called GenSE. Specifically, we approach SE as a

conditional language modeling task rather than a continuous signal regression pr
oblem defined in existing works. This is achieved by tokenizing speech signals i
nto semantic tokens using a pre-trained self-supervised model and into acoustic
tokens using a custom-designed single-quantizer neural codec model. To improve t
he stability of language model predictions, we propose a hierarchical modeling m
ethod that decouples the generation of clean semantic tokens and clean acoustic
tokens into two distinct stages. Moreover, we introduce a token chain prompting
mechanism during the acoustic token generation stage to ensure timbre consistenc
y throughout the speech enhancement process. Experimental results on benchmark d
atasets demonstrate that our proposed approach outperforms state-of-the-art SE s
ystems in terms of speech quality and generalization capability. Codes and demos
 are publicly available at https://anonymous.4open.science/w/gen-se-7F52/.
****

Seanie Lee,Minsu Kim,Lynn Cherif,David Dobre,Juho Lee,Sung Ju Hwang,Kenji Kawagu
chi,Gauthier Gidel,Yoshua Bengio,Nikolay Malkin,Moksh Jain
Learning Diverse Attacks on Large Language Models for Robust Red-Teaming and Saf
ety Tuning
Red-teaming, or identifying prompts that elicit harmful responses, is a critical
 step in ensuring the safe and responsible deployment of large language models (
LLMs). Developing effective protection against many modes of attack prompts requ
ires discovering diverse attacks. Automated red-teaming typically uses reinforce
ment learning to fine-tune an attacker language model to generate prompts that e
licit undesirable responses from a target LLM, as measured, for example, by an a
uxiliary toxicity classifier. We show that even with explicit regularization to
favor novelty and diversity, existing approaches suffer from mode collapse or fa
il to generate effective attacks. As a flexible and probabilistically principled
 alternative, we propose to use GFlowNet fine-tuning, followed by a secondary sm
oothing phase, to train the attacker model to generate *diverse* and *effective*
 attack prompts. We find that the attacks generated by our method are effective
against a wide range of target LLMs, both with and without safety tuning, and tr
ansfer well between target LLMs. Finally, we demonstrate that models safety-tune
d using a dataset of red-teaming prompts generated by our method are robust to a
ttacks from other RL-based red-teaming approaches.
****

Daiwei Chen,Yi Chen,Aniket Rege,Zhi Wang,Ramya Korlakai Vinayak
PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment
Foundation models trained on internet-scale data benefit from extensive alignmen
t to human preferences before deployment. However, existing methods typically as
sume a homogeneous preference shared by all individuals, overlooking the diversi
ty inherent in human values. In this work, we propose a general reward modeling
framework for pluralistic alignment (PAL), which incorporates diverse preference
s from the ground up. PAL has a modular design that leverages commonalities acro
ss users while catering to individual personalization, enabling efficient few-sh
ot localization of preferences for new users. Extensive empirical evaluation dem
onstrates that PAL matches or outperforms state-of-the-art methods on both text-
to-text and text-to-image tasks: on Reddit TL;DR Summary, PAL is 1.7% more accur
ate for seen users and 36% more accurate for unseen users compared to the previo
us best method, with 100× less parameters. On Pick-a-Pic v2, PAL is 2.5% more ac
curate than the best method with 156× fewer learned parameters. Finally, we prov
ide theoretical analysis for generalization of rewards learned via PAL framework
 showcasing the reduction in number of samples needed per user.
****

Paul Brunzema,Mikkel Jordahn,John Willes,Sebastian Trimpe,Jasper Snoek,James Har
rison
Bayesian Optimization via Continual Variational Last Layer Training
Gaussian Processes (GPs) are widely seen as the state-of-the-art surrogate model
s for Bayesian optimization (BO) due to their ability to model uncertainty and t
heir performance on tasks where correlations are easily captured (such as those
defined by Euclidean metrics) and their ability to be efficiently updated online
. However, the performance of GPs depends on the choice of kernel, and kernel se

lection for complex correlation structures is often difficult or must be made be spoke. While Bayesian neural networks (BNNs) are a promising direction for highe r capacity surrogate models, they have so far seen limited use due to poor perfo rmance on some problem types. In this paper, we propose an approach which shows competitive performance on many problem types, including some that BNNs typicall y struggle with. We build on variational Bayesian last layers (VBLLs), and conne ct training of these models to exact conditioning in GPs. We exploit this connec tion to develop an efficient online training algorithm that interleaves conditio ning and optimization. Our findings suggest that VBLL networks significantly out perform GPs and other BNN architectures on tasks with complex input correlations , and match the performance of well-tuned GPs on established benchmark tasks.
****

luca alessandro silva,Barthelemy Meynard-Piganeau,Carlo Lucibello,Christoph Fein auer

Fast Uncovering of Protein Sequence Diversity from Structure

We present InvMSAFold, an inverse folding method for generating protein sequence s optimized for diversity and speed. For a given structure, InvMSAFold generates the parameters of a pairwise probability distribution over the space of sequenc es, capturing the amino acid covariances observed in Multiple Sequence Alignment s (MSA) of homologous proteins. This allows for the efficient generation of high ly diverse protein sequences while preserving structural and functional integrit y.

We demonstrate that this increased diversity in sampled sequences translates int o greater variability in biochemical properties, highlighting the exciting poten tial of our method for applications such as protein design. The orders of magnit ude improvement in sampling speed compared to existing methods unlocks new possi bilities for high-throughput in virtual screening.
****

Klaus-Rudolf Kladny,Bernhard Schölkopf,Michael Muehlebach

Conformal Generative Modeling with Improved Sample Efficiency through Sequential Greedy Filtering

Generative models lack rigorous statistical guarantees with respect to their pre dictions. In this work, we propose Sequential Conformal Prediction for Generativ e Models (SCOPE-Gen), a sequential conformal prediction method producing predict ion sets that satisfy a rigorous statistical guarantee called conformal admissib ility control. This guarantee means that the prediction sets contain at least on e admissible (or valid) example, with high probability. To this end, our method first samples an initial set of i.i.d. examples from a black box generative mode l. Then, this set is iteratively pruned via so-called greedy filters. As a conse quence of the iterative generation procedure, admissibility of the final predict ion set factorizes as a Markov chain, where each factor can be controlled separa tely, using conformal prediction. In comparison to prior work, our method demons trates a large reduction in the number of admissibility evaluations during calib ration. This is crucial e.g. in safety-critical applications, where these evalua tions must be conducted manually by domain experts and are therefore costly and time consuming. We highlight the advantages of our method in terms of admissibil ity evaluations and cardinality of the prediction set through experiments in nat ural language generation and molecular graph extension tasks.
****

Laura Ruis,Maximilian Mozes,Juhan Bae,Siddhartha Rao Kamalakara,Dwaraknath Gnane shwar,Acyr Locatelli,Robert Kirk,Tim Rocktäschel,Edward Grefenstette,Max Bartolo

Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models

The capabilities and limitations of Large Language Models (LLMs) have been sketc hed out in great detail in recent years, providing an intriguing yet conflicting picture. On the one hand, LLMs demonstrate a general ability to solve problems. On the other hand, they show surprising reasoning gaps when compared to humans, casting doubt on the robustness of their generalisation strategies. The sheer v olume of data used in the design of LLMs has precluded us from applying the meth od traditionally used to measure generalisation: train-test set separation. To o vercome this, we study what kind of generalisation strategies LLMs employ when p

erforming reasoning tasks by investigating the pretraining data they rely on. Fo
r two models of different sizes (7B and 35B) and 2.5B of their pretraining token
s, we identify what documents influence the model outputs for three simple mathe
matical reasoning tasks and contrast this to the data that are influential for a
nswering factual questions. We find that, while the models rely on mostly distin
ct sets of data for each factual question, a document often has a similar influe
nce across different reasoning questions within the same task, indicating the pr
esence of procedural knowledge. We further find that the answers to factual ques
tions often show up in the most influential data. However, for reasoning questio
ns the answers usually do not show up as highly influential, nor do the answers
to the intermediate reasoning steps. When we characterise the top ranked documen
ts for the reasoning questions qualitatively, we confirm that the influential do
cuments often contain procedural knowledge, like demonstrating how to obtain a s
olution using formulae or code. Our findings indicate that the approach to reaso
ning the models use is unlike retrieval, and more like a generalisable strategy
that synthesises procedural knowledge from documents doing a similar form of rea
soning.
****

Peitian Zhang,Zheng Liu,Shitao Xiao,Ninglu Shao,Qiwei Ye,Zhicheng Dou
Long Context Compression with Activation Beacon
Long context compression is a critical research problem due to its significance
in reducing the high computational and memory costs associated with LLMs. In thi
s paper, we propose Activation Beacon, a plug-in module for transformer-based LL
Ms that targets effective, efficient, and flexible compression of long contexts.
 To achieve this, our method introduces the following technical designs.
1) We directly compress the activations (i.e. keys and values at every layer), r
ather than leveraging soft prompts to relay information (which constitute a majo
r bottleneck to encapsulate the complex information within long contexts).
2) We tailor the compression workflow, where each fine-grained input unit is pro
gressively compressed, enabling high-quality compression and efficient computati
on during both training and inference.
3) We train the model through compression-based auto-regression, making full use
 of plain texts and instructional data to optimize the model's compression perfo
rmance.
4) During training, we randomly sample a compression ratio at each step, teachin
g the model to support a wide range of compression configurations.

Extensive evaluations are conducted on various long-context tasks whose lengths
(e.g., 128K) may far exceed the maximum training length (20K), such as document
understanding, few-shot learning, and Needle-in-a-Haystack. Whilst existing meth
ods struggle to handle these challenging tasks, Activation Beacon maintains a co
mparable performance to the uncompressed baseline across various scenarios,
achieving a 2x acceleration in inference time and an 8x reduction of memory cost
s for KV cache.
****

Ishan Amin,Sanjeev Raja,Aditi S. Krishnapriyan
Towards Fast, Specialized Machine Learning Force Fields: Distilling Foundation M
odels via Energy Hessians
The foundation model (FM) paradigm is transforming Machine Learning Force Fields
 (MLFFs), leveraging general-purpose representations and scalable training to pe
rform a variety of computational chemistry tasks. Although MLFF FMs have begun t
o close the accuracy gap relative to first-principles methods, there is still a
strong need for faster inference speed. Additionally, while research is increasi
ngly focused on general-purpose models which transfer across chemical space, pra
ctitioners typically only study a small subset of systems at a given time. At te
st time, MLFFs must also obey physical constraints unique to the downstream use
case, such as energy conservation for molecular dynamics simulations. This under
scores the need for fast, specialized MLFFs relevant to specific downstream appl
ications, which preserve test-time physical soundness while maintaining train-ti
me scalability. In this work, we introduce a method for transferring general-pur

pose representations from MLFF foundation models to smaller, faster MLFFs specia
lized to specific regions of chemical space. We formulate our approach as an arc
hitecture-agnostic knowledge distillation procedure, where the smaller "student"
 MLFF is trained to match the Hessians of the energy predictions of the "teacher
" foundation model. We demonstrate our approach across multiple recent foundatio
n models, large-scale datasets, chemical subsets, and downstream tasks. Our spec
ialized MLFFs can be up to 20 times faster than the original foundation model, w
hile retaining, and in some cases exceeding, its performance and that of undisti
lled models. We also show that distilling from a teacher model with a direct for
ce parameterization into a student model trained with conservative forces (i.e.,
 computed as derivatives of the potential energy) successfully leverages the rep
resentations from the large-scale teacher for improved accuracy, while maintaini
ng energy conservation during test-time molecular dynamics simulations. More bro
adly, our work suggests a new paradigm for MLFF development, in which foundation
 models are released along with smaller, specialized simulation ``engines" for c
ommon chemical subsets. The implementation of our method is available at https:/
/github.com/ASK-Berkeley/MLFF-distill.
****

Yongxian Wei,Zixuan Hu,Li Shen,Zhenyi Wang,Chun Yuan,Dacheng Tao
Open-Vocabulary Customization from CLIP via Data-Free Knowledge Distillation
Vision-language models such as CLIP have demonstrated strong zero-shot performan
ce, but their considerable size and inefficient inference limit customizable dep
loyment for users. While knowledge distillation is a solution, it still requires
 the original data, which is not always available due to copyrights and privacy
concerns. For many users seeking open-vocabulary customization, Data-Free Knowle
dge Distillation (DFKD) emerges as a promising direction. Upon rethinking DFKD,
we find that existing methods fail on CLIP due to their heavy reliance on BatchN
orm layers, which are unexpectedly unusable in CLIP. Based on our findings, we a
dopt image-text matching to achieve DFKD for CLIP, enabling customization based
on arbitrary class texts. This involves (i) inversing a surrogate dataset from C
LIP based on text prompts; and (ii) distilling a student model from CLIP using t
he surrogate dataset. Specifically, we introduce style dictionary diversificatio
n to enhance the diversity of synthetic images. To prevent uncontrollable semant
ics introduced by diversification, we propose a class consistency maintaining st
rategy to ensure the consistency of synthetic images. Based on synthetic images
with various styles, we further propose meta knowledge distillation to train the
 student model with good generalization ability. Moreover, we introduce a simple
 yet effective method to enable customization based on few example images. Compr
ehensive experiments showcase the superiority of our approach across twelve cust
omized tasks, achieving a 9.33\% improvement compared to existing DFKD methods.
****

Shaozhe Hao,Xuantong LIU,Xianbiao Qi,Shihao Zhao,Bojia Zi,Rong Xiao,Kai Han,Kwan
-Yee K. Wong
BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Re
presentation Capabilities
We introduce BiGR, a novel conditional image generation model using compact bina
ry latent codes for generative training, focusing on enhancing both generation a
nd representation capabilities. BiGR is the first conditional generative model t
hat unifies generation and discrimination within the same framework.
BiGR features a binary tokenizer, a masked modeling mechanism, and a binary tran
scoder for binary code prediction.
Additionally, we introduce a novel entropy-ordered sampling method to enable eff
icient image generation.
Extensive experiments validate BiGR's superior performance in generation quality
, as measured by FID-50k, and representation capabilities, as evidenced by linea
r-probe accuracy.
Moreover, BiGR showcases zero-shot generalization across various vision tasks, e
nabling applications such as image inpainting, outpainting, editing, interpolati
on, and enrichment, without the need for structural modifications. Our findings
suggest that BiGR unifies generative and discriminative tasks effectively, pavin

g the way for further advancements in the field. We further enable BiGR to perfo
rm text-to-image generation, showcasing its potential for broader applications.
****
Nguyen Hung-Quang,Ngoc-Hieu Nguyen,The-Anh Ta,Thanh Nguyen-Tang,Kok-Seng Wong,Ho
ang Thanh-Tung,Khoa D Doan
Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attack
s
Deep neural networks are vulnerable to backdoor attacks, a type of adversarial
attack that poisons the training data to manipulate the behavior of models train
ed on such data.
Clean-label attacks are a more stealthy form of backdoor attacks that can perfor
m the attack without changing the labels of poisoned data.
Early works on clean-label attacks added triggers to a random subset of the trai
ning set, ignoring the fact that samples contribute unequally to the attack's su
ccess. This results in high poisoning rates and low attack success rates.
To alleviate the problem, several supervised learning-based sample selection str
ategies have been proposed.
However, these methods assume access to the entire labeled training set and requ
ire training, which is expensive and may not always be practical.
This work studies a new and more practical (but also more challenging) threat mo
del where the attacker only provides data for the target class (e.g., in face re
cognition systems) and has no knowledge of the victim model or any other classes
 in the training set.
We study different strategies for selectively poisoning a small set of training
samples in the target class to boost the attack success rate in this setting.
Our threat model poses a serious threat in training machine learning models with
 third-party datasets, since the attack can be performed effectively with limite
d information. Experiments on benchmark datasets illustrate the effectiveness of
 our strategies in improving clean-label backdoor attacks.
****
Yutong Yin,Zhaoran Wang
Are Transformers Able to Reason by Connecting Separated Knowledge in Training Da
ta?
Humans exhibit remarkable compositional reasoning by integrating knowledge from
various sources. For example, if someone learns (B = f(A)) from one source and
 (C = g(B)) from another, they can deduce (C=g(B)=g(f(A))) even without enco
untering (ABC) together, showcasing the generalization ability of human intell
igence. In this paper, we introduce a synthetic learning task, "FTCT" (Fragment
ed at Training, Chained at Testing), to validate the potential of Transformers
in replicating this skill and interpret its inner mechanism. During training, da
ta consist of separated knowledge fragments from an overall causal graph. In tes
ting, Transformers must combine these fragments to infer complete causal traces.
 Our findings demonstrate that few-shot Chain-of-Thought prompting enables Trans
formers to perform compositional reasoning on FTCT by revealing correct combinat
ions of fragments, even if such combinations were absent in training data. Furth
ermore, the emergence of compositional reasoning ability is strongly correlated
with model complexity and training-testing data similarity. We propose, both the
oretically and empirically, that Transformers learn an underlying generalizable
program from training, enabling effective compositional reasoning during testing
.
****
Songyuan Zhang,Oswin So,Mitchell Black,Chuchu Fan
Discrete GCBF Proximal Policy Optimization for Multi-agent Safe Optimal Control
Control policies that can achieve high task performance and satisfy safety const
raints are desirable for any system, including multi-agent systems (MAS). One pr
omising technique for ensuring the safety of MAS is distributed control barrier
functions (CBF). However, it is difficult to design distributed CBF-based polici
es for MAS that can tackle unknown discrete-time dynamics, partial observability
, changing neighborhoods, and input constraints, especially when a distributed h
igh-performance nominal policy that can achieve the task is unavailable. To tack

le these challenges, we propose **DGPPO**, a new framework that *simultaneously* learns both a *discrete* graph CBF which handles neighborhood changes and input constraints, and a distributed high-performance safe policy for MAS with unknown discrete-time dynamics.
We empirically validate our claims on a suite of multi-agent tasks spanning three different simulation engines. The results suggest that, compared with existing methods, our DGPPO framework obtains policies that achieve high task performance (matching baselines that ignore the safety constraints), and high safety rates (matching the most conservative baselines), with a *constant* set of hyperparameters across all environments.
****
Kotaro Yoshida,Yuji Naraki,Takafumi Horie,Ryosuke Yamaki,Ryotaro Shimizu,Yuki Saito,Julian McAuley,Hiroki Naganuma
Mastering Task Arithmetic: τJp as a Key Indicator for Weight Disentanglement
Model-editing techniques using task arithmetic have rapidly gained attention. Through task arithmetic, simply through arithmetic operations on the weights of pre-trained and fine-tuned models create desired models, such as multi-task models, models in which specific tasks are unsolvable, or domain-transferred models. However, task arithmetic faces challenges, such as poor reproducibility and the high cost associated with adjusting coefficients in the arithmetic operations on model parameters, which have limited its practical success. In this paper, we present three key contributions in the context of task addition and task negation within task arithmetic. First, we propose a new metric called τJp which is based on the product of the task vector (τ) and the Jacobian of the pre-trained model with respect to its weights. We show that τJp has a causal relationship with the interference that occurs from arithmetic operations. Second, we show that introducing regularization to minimize τJp significantly mitigates interference between task inference, which leads to the elimination of coefficient tuning and improved accuracy on each task.
Third, in the context of incremental learning, we demonstrate that our τJp regularization achieves more robust performance in environments where access to future tasks is unavailable, thus validating the scalability of the approach.
Finally, we demonstrate that the τJp regularizer further reinforces the performance of task arithmetic by leveraging publicly available fine-tuned models, offering practical benefits for real-world applications.
Our code is available at https://github.com/katoro8989/tau-Jp_Task_Arithmetic
****
Yuyang Zhao,Chung-Ching Lin,Kevin Lin,Zhiwen Yan,Linjie Li,Zhengyuan Yang,Jianfeng Wang,Gim Hee Lee,Lijuan Wang
GenXD: Generating Any 3D and 4D Scenes
Recent developments in 2D visual generation have been remarkably successful. However, 3D and 4D generation remain challenging in real-world applications due to the lack of large-scale 4D data and effective model design. In this paper, we propose to jointly investigate general 3D and 4D generation by leveraging camera and object movements commonly observed in daily life. Due to the lack of real-world 4D data in the community, we first propose a data curation pipeline to obtain camera poses and object motion strength from videos. Based on this pipeline, we introduce a large-scale real-world 4D scene dataset: CamVid-30K. By leveraging all the 3D and 4D data, we develop our framework, GenXD, which allows us to produce any 3D or 4D scene. We propose multiview-temporal modules, which disentangle camera and object movements, to seamlessly learn from both 3D and 4D data. Additionally, GenXD employs masked latent conditions to support a variety of conditioning views. GenXD can generate videos that follow the camera trajectory as well as consistent 3D views that can be lifted into 3D representations. We perform extensive evaluations across various real-world and synthetic datasets, demonstrating GenXD's effectiveness and versatility compared to previous methods in 3D and 4D generation.
****
Yuxin Guo,Shuailei Ma,Shijie Ma,Xiaoyi Bao,Chen-Wei Xie,Kecheng Zheng,Tingyu Wen

g,Siyang Sun,Yun Zheng,Wei Zou
Aligned Better, Listen Better For Audio-Visual Large Language Models
Audio is essential for multimodal video understanding. On the one hand, video in herently contains audio and audio supplies complementary information to the visu al modality. Besides, video large language models (Video-LLMs) can encounter man y audio-centric settings. However, existing Video-LLMs and Audio-Visual Large La nguage Models (AV-LLMs) exhibit deficiencies in exploiting audio information, le ading to weak understanding and hallucination. To solve the issues, we delve int o the model architecture and data aspects. (1) From the architectural perspectiv e, we propose a fine-grained AV-LLM, namely Dolphin. The concurrent alignment of audio and visual modalities in both temporal and spatial dimensions ensures a c omprehensive and accurate understanding of videos. Specifically, we devise an au dio-visual multi-scale adapter for multi-scale information aggregation, which ac hieves spatial alignment. For temporal alignment, we propose audio-visual interl eaved merging. (2) From the data perspective, we curate an audio-visual caption \& instruction-tuning dataset, called AVU. It comprises 5.2 million diverse, ope n-ended data tuples (video, audio, question, answer) and introduces a novel data partitioning strategy. Extensive experiments show our model not only achieves r emarkable performance in audio-visual understanding, but also mitigates hallucin ations. Our codes and dataset will be made publicly available.
****
Patrick Emami,Zhaonan Li,Saumya Sinha,Truc Nguyen
SysCaps: Language Interfaces for Simulation Surrogates of Complex Systems
Surrogate models are used to predict the behavior of complex energy systems that are too expensive to simulate with traditional numerical methods.
Our work introduces the use of language descriptions, which we call "system capt ions" or SysCaps, to interface with such surrogates.
We argue that interacting with surrogates through text, particularly natural lan guage, makes these models more accessible for both experts and non-experts.
We introduce a lightweight multimodal text and timeseries regression model and a training pipeline that uses large language models (LLMs) to synthesize high-qua lity captions from simulation metadata.
Our experiments on two real-world simulators of buildings and wind farms show th at our SysCaps-augmented surrogates have better accuracy on held-out systems tha n traditional methods while enjoying new generalization abilities, such as handl ing semantically related descriptions of the same test system.
Additional experiments also highlight the potential of SysCaps to unlock languag e-driven design space exploration and to regularize training through prompt augm entation.
****
Jianan Zhao,Zhaocheng Zhu,Mikhail Galkin,Hesham Mostafa,Michael M. Bronstein,Jia n Tang
Fully-inductive Node Classification on Arbitrary Graphs
One fundamental challenge in graph machine learning is generalizing to new graph s. Many existing methods following the inductive setup can generalize to test gr aphs with new structures, but assuming the feature and label spaces remain the s ame as the training ones.
This paper introduces a fully-inductive setup, where models should perform infer ence on arbitrary test graphs with new structures, feature and label spaces. We propose GraphAny as the first attempt at this challenging setup. GraphAny models inference on a new graph as an analytical solution to a LinearGNN, which can be naturally applied to graphs with any feature and label spaces. To further build a stronger model with learning capacity, we fuse multiple LinearGNN predictions with learned inductive attention scores. Specifically, the attention module is carefully parameterized as a function of the entropy-normalized distance feature s between pairs of LinearGNN predictions to ensure generalization to new graphs. Empirically, GraphAny trained on a single Wisconsin dataset with only 120 label ed nodes can generalize to 30 new graphs with an average accuracy of 67.26\%, su rpassing not only all inductive baselines, but also strong transductive methods trained separately on each of the 30 test graphs.

****
Chen Xu,Yuxin Li,Wenjie Wang,Liang Pang,Jun Xu,Tat-Seng Chua
Bridging Jensen Gap for Max-Min Group Fairness Optimization in Recommendation
Group max-min fairness (MMF) is commonly used in fairness-aware recommender systems (RS) as an optimization objective, as it aims to protect marginalized item groups and ensures a fair competition platform. However, our theoretical analysis indicates that integrating MMF constraint violates the assumption of sample independence during optimization, causing the loss function to deviate from linear additivity. Such nonlinearity property introduces the Jensen gap between the model's convergence point and the optimal point if mini-batch sampling is applied. Both theoretical and empirical studies show that as the mini-batch size decreases and the group size increases, the Jensen gap will widen accordingly. Some methods using heuristic re-weighting or debiasing strategies have the potential to bridge the Jensen gap. However, they either lack theoretical guarantees or suffer from heavy computational costs. To overcome these limitations, we first theoretically demonstrate that the MMF-constrained objective can be essentially reformulated as a group-weighted optimization objective. Then we present an efficient and effective algorithm named FairDual, which utilizes a dual optimization technique to minimize Jensen gap. Our theoretical analysis demonstrates that FairDual can achieve a sub-linear convergence rate to the globally optimal solution and the Jensen gap can be well bounded under a mini-batch sampling strategy with random shuffle. Extensive experiments conducted using six large-scale RS backbone models on three publicly available datasets demonstrate that FairDual outperforms all baselines in terms of both accuracy and fairness.
****
Jungwoo Park,Ahn Young Jin,Kee-Eung Kim,Jaewoo Kang
Monet: Mixture of Monosemantic Experts for Transformers
Understanding the internal computations of large language models (LLMs) is crucial for aligning them with human values and preventing undesirable behaviors like toxic content generation. However, mechanistic interpretability is hindered by *polysemanticity*—where individual neurons respond to multiple, unrelated concepts. While Sparse Autoencoders (SAEs) have attempted to disentangle these features through sparse dictionary learning, they have compromised LLM performance due to reliance on post-hoc reconstruction loss. To address this issue, we introduce **Mixture of Monosemantic Experts for Transformers (Monet)** architecture, which incorporates sparse dictionary learning directly into end-to-end Mixture-of-Experts pretraining. Our novel expert decomposition method enables scaling the expert count to 262,144 per layer while total parameters scale proportionally to the square root of the number of experts. Our analyses demonstrate mutual exclusivity of knowledge across experts and showcase the parametric knowledge encapsulated within individual experts. Moreover, **Monet** allows knowledge manipulation over domains, languages, and toxicity mitigation without degrading general performance. Our pursuit of transparent LLMs highlights the potential of scaling expert counts to enhance mechanistic interpretability and directly resect the internal knowledge to fundamentally adjust model behavior.
****
Suin Lee,Daeshik Kim
TexTailor: Customized Text-aligned Texturing via Effective Resampling
We present TexTailor, a novel method for generating consistent object textures from textual descriptions. Existing text-to-texture synthesis approaches utilize depth-aware diffusion models to progressively generate images and synthesize textures across predefined multiple viewpoints. However, these approaches lead to a gradual shift in texture properties across viewpoints due to (1) insufficient integration of previously synthesized textures at each viewpoint during the diffusion process and (2) the autoregressive nature of the texture synthesis process. Moreover, the predefined selection of camera positions, which does not account for the object's geometry, limits the effective use of texture information synthesized from different viewpoints, ultimately degrading overall texture consistency. In TexTailor, we address these issues by (1) applying a resampling scheme that repeatedly integrates information from previously synthesized textures within

the diffusion process, and (2) fine-tuning a depth-aware diffusion model on the se resampled textures. During this process, we observed that using only a few tr aining images restricts the model's original ability to generate high-fidelity i mages aligned with the conditioning, and therefore propose an performance preser vation loss to mitigate this issue. Additionally, we improve the synthesis of vi ew-consistent textures by adaptively adjusting camera positions based on the obj ect's geometry. Experiments on a subset of the Objaverse dataset and the ShapeNe t car dataset demonstrate that TexTailor outperforms state-of-the-art methods in synthesizing view-consistent textures.
****

Aleksandar Makelov,Georg Lange,Neel Nanda
Towards Principled Evaluations of Sparse Autoencoders for Interpretability and C ontrol
Disentangling model activations into human-interpretable features is a central problem in interpretability. Sparse autoencoders (SAEs) have recently attracted much attention as a scalable unsupervised approach to this problem. However, our imprecise understanding of ground-truth features in realistic scenarios makes it difficult to measure the success of SAEs. To address this challenge, we propose to evaluate SAEs on specific tasks by comparing them to supervised feature dictionaries computed with knowledge of the concepts relevant to the task.

Specifically, we suggest that it is possible to (1) compute supervised sparse feature dictionaries that disentangle model computations for a specific task; (2) use them to evaluate and contextualize the degree of disentanglement and control offered by SAE latents on this task. Importantly, we can do this in a way that is agnostic to whether the SAEs have learned the exact ground-truth features or a different but similarly useful representation.

As a case study, we apply this framework to the indirect object identification (IOI) task using GPT-2 Small, with SAEs trained on either the IOI or OpenWebText datasets. We find that SAEs capture interpretable features for the IOI task, and that more recent SAE variants such as Gated SAEs and Top-K SAEs are competitive with supervised features in terms of disentanglement and control over the model. We also exhibit, through this setup and toy models, some qualitative phenomena in SAE training illustrating feature splitting and the role of feature magnitudes in solutions preferred by SAEs.
****

Jie Ren,Pin-Yu Chen,Ren Wang
Revisiting Mode Connectivity in Neural Networks with Bezier Surface
Understanding the loss landscapes of neural networks (NNs) is critical for optim izing model performance. Previous research has identified the phenomenon of mode connectivity on curves, where two well-trained NNs can be connected by a contin uous path in parameter space where the path maintains nearly constant loss. In t his work, we extend the concept of mode connectivity to explore connectivity on surfaces, significantly broadening its applicability and unlocking new opportuni ties. While initial attempts to connect models via linear surfaces in parameter space were unsuccessful, we propose a novel optimization technique that consiste ntly discovers Bézier surfaces with low-loss and high-accuracy connecting multip le NNs in a nonlinear manner. We further demonstrate that even without optimizat ion, mode connectivity exists in certain cases of Bézier surfaces, where the mod els are carefully selected and combined linearly. This approach provides a deepe r and more comprehensive understanding of the loss landscape and offers a novel way to identify models with enhanced performance for model averaging and output ensembling. We demonstrate the effectiveness of our method on CIFAR-10, CIFAR-10 0, and Tiny-ImageNet datasets using VGG16, ResNet18, and ViT architectures.
****

Merey Ramazanova,Alejandro Pardo,Bernard Ghanem,Motasem Alfarra
Test-Time Adaptation for Combating Missing Modalities in Egocentric Videos
Understanding videos that contain multiple modalities is crucial, especially in

egocentric videos, where combining various sensory inputs significantly improves tasks like action recognition and moment localization. However, real-world applications often face challenges with incomplete modalities due to privacy concerns, efficiency needs, or hardware issues. Current methods, while effective, often necessitate retraining the model entirely to handle missing modalities, making them computationally intensive, particularly with large training datasets. In this study, we propose a novel approach to address this issue at test time without requiring retraining. We frame the problem as a test-time adaptation task, where the model adjusts to the available unlabeled data at test time. Our method, MiDl~(Mutual information with self-Distillation), encourages the model to be insensitive to the specific modality source present during testing by minimizing the mutual information between the prediction and the available modality. Additionally, we incorporate self-distillation to maintain the model's original performance when both modalities are available. MiDl represents the first self-supervised, online solution for handling missing modalities exclusively at test time. Through experiments with various pretrained models and datasets, MiDl demonstrates substantial performance improvement without the need for retraining.
****
Canfer Akbulut,Kevin Robinson,Maribeth Rauh,Isabela Albuquerque,Olivia Wiles,Laura Weidinger,Verena Rieser,Yana Hasson,Nahema Marchal,Iason Gabriel,William Isaac,Lisa Anne Hendricks
Century: A Framework and Dataset for Evaluating Historical Contextualisation of Sensitive Images
How do multi-modal generative models describe images of recent historical events and figures, whose legacies may be nuanced, multifaceted, or contested? This task necessitates not only accurate visual recognition, but also socio-cultural knowledge and cross-modal reasoning. To address this evaluation challenge, we introduce Century -- a novel dataset of sensitive historical images. This dataset consists of 1,500 images from recent history, created through an automated method combining knowledge graphs and language models with quality and diversity criteria created from the practices of museums and digital archives. We demonstrate through automated and human evaluation that this method produces a set of images that depict events and figures that are diverse across topics and represents all regions of the world.
We additionally propose an evaluation framework for evaluating the historical contextualisation capabilities along dimensions of accuracy, thoroughness, and objectivity. We demonstrate this approach by using Century to evaluate four foundation models, scoring performance using both automated and human evaluation. We find that historical contextualisation of sensitive images poses a significant challenge for modern multi-modal foundation models, and offer practical recommendations for how developers can use Century to evaluate improvements to models and applications.
****
Shuai Tan,Biao Gong,Xiang Wang,Shiwei Zhang,DanDan Zheng,Ruobing Zheng,Kecheng Zheng,Jingdong Chen,Ming Yang
Animate-X: Universal Character Image Animation with Enhanced Motion Representation
Character image animation, which generates high-quality videos from a reference image and target pose sequence, has seen significant progress in recent years. However, most existing methods only apply to human figures, which usually do not generalize well on anthropomorphic characters commonly used in industries like gaming and entertainment. Our in-depth analysis suggests to attribute this limitation to their insufficient modeling of motion, which is unable to comprehend the movement pattern of the driving video, thus imposing a pose sequence rigidly onto the target character. To this end, this paper proposes $\texttt{Animate-X}$, a universal animation framework based on LDM for various character types (collectively named $\texttt{X}$), including anthropomorphic characters. To enhance motion representation, we introduce the Pose Indicator, which captures comprehensive motion pattern from the driving video through both implicit and explicit manner. The former leverages CLIP visual features of a driving video to extract its g

ist of motion, like the overall movement pattern and temporal relations among mo
tions, while the latter strengthens the generalization of LDM by simulating poss
ible inputs in advance that may arise during inference. Moreover, we introduce a
 new Animated Anthropomorphic Benchmark ($\texttt{$A^2$Bench}$) to evaluate the
performance of $\texttt{Animate-X}$ on universal and widely applicable animation
 images. Extensive experiments demonstrate the superiority and effectiveness of
$\texttt{Animate-X}$ compared to state-of-the-art methods.
****
Zimu Lu,Aojun Zhou,Ke Wang,Houxing Ren,Weikang Shi,Junting Pan,Mingjie Zhan,Hong
sheng Li
MathCoder2: Better Math Reasoning from Continued Pretraining on Model-translated
 Mathematical Code
Code has been shown to be effective in enhancing the mathematical reasoning abil
ities of large language models due to its precision and accuracy. Previous works
 involving continued mathematical pretraining often include code that utilizes
math-related packages, which are primarily designed for fields such as engineeri
ng, machine learning, signal processing, or module testing, rather than being di
rectly focused on mathematical reasoning. In this paper, we introduce a novel me
thod for generating mathematical code accompanied with corresponding reasoning s
teps for continued pretraining. Our approach begins with the construction of a h
igh-quality mathematical continued pretraining dataset by incorporating math-rel
ated web data, code using mathematical packages, math textbooks, and synthetic d
ata. Next, we construct reasoning steps by extracting LaTeX expressions, the con
ditions needed for the expressions, and the results of the expressions from the
previously collected dataset. Based on this extracted information, we generate c
orresponding code to accurately capture the mathematical reasoning process. Appe
nding the generated code to each reasoning step results in data consisting of pa
ired natural language reasoning steps and their corresponding code. Combining th
is data with the original dataset results in a 19.2B-token high-performing mathe
matical pretraining corpus, which we name MathCode-Pile. Training several popula
r base models with this corpus significantly improves their mathematical abiliti
es, leading to the creation of the MathCoder2 family of models. All of our data
processing and training code is open-sourced, ensuring full transparency and eas
y reproducibility of the entire data collection and training pipeline.
****
Kha Pham,Hung Le,Man Ngo,Truyen Tran
Rapid Selection and Ordering of In-Context Demonstrations via Prompt Embedding C
lustering
While Large Language Models (LLMs) excel at in-context learning (ICL) using just
 a few demonstrations, their performances are sensitive to demonstration orders.
 The reasons behind this sensitivity remain poorly understood. In this paper, we
 investigate the prompt embedding space to bridge the gap between the order sens
itivity of ICL with inner workings of decoder-only LLMs, uncovering the clusteri
ng property: prompts sharing the first and last demonstrations have closer embed
dings, with first-demonstration clustering usually being stronger in practice. W
e explain this property through extensive theoretical analyses and empirical evi
dences. Our finding suggests that the positional encoding and the causal attenti
on mask are key contributors to the clustering phenomenon. Leveraging this clust
ering insight, we introduce Cluster-based Search, a novel method that accelerate
s the selection and ordering of demonstrations in self-adaptive ICL settings. Ou
r approach substantially decreases the time complexity from factorial to quadrat
ic, saving 92% to nearly 100% execution time while maintaining comparable perfor
mance to exhaustive search.
****
Arthur Jacot,Peter Súkeník,Zihan Wang,Marco Mondelli
Wide Neural Networks Trained with Weight Decay Provably Exhibit Neural Collapse
Deep neural networks (DNNs) at convergence consistently represent the training d
ata in the last layer via a geometric structure referred to as neural collapse.
This empirical evidence has spurred a line of theoretical research aimed at prov
ing the emergence of neural collapse, mostly focusing on the unconstrained featu

res model. Here, the features of the penultimate layer are free variables, which makes the model data-agnostic and puts into question its ability to capture DNN training. Our work addresses the issue, moving away from unconstrained features and studying DNNs that end with at least two linear layers. We first prove generic guarantees on neural collapse that assume \emph{(i)} low training error and balancedness of
linear layers (for within-class variability collapse), and \emph{(ii)} bounded conditioning of the features before the linear part (for orthogonality of class-means, and their alignment with weight matrices). The balancedness refers to the fact that $W_{\ell+1}^\top W_{\ell+1}\approx W_\ell W_\ell ^\top$ for any pair of
consecutive weight matrices
of the linear part, and the bounded conditioning requires a well-behaved ratio between largest and smallest non-zero singular values of the features. We then show that such assumptions hold for gradient descent training with weight decay: \emph{(i)} for networks with a wide first layer, we prove low training error and balancedness, and \emph{(ii)} for solutions that are either nearly optimal or stable under large learning rates, we additionally prove the bounded conditioning. Taken together, our results are the first to show neural collapse in the end-to-end training of DNNs.
****
Joshua Russell,Ignacio Gavier,Devdhar Patel,Edward Rietman,Hava T Siegelmann
Optimizing Neural Network Representations of Boolean Networks
Neural networks are known to be universal computers for Boolean functions. Recent advancements in hardware have significantly reduced matrix multiplication times, making neural network simulation both fast and efficient. Consequently, functions defined by complex Boolean networks are increasingly viable candidates for simulation through their neural network representation. Prior research has introduced a general method for deriving neural network representations of Boolean networks. However, the resulting neural networks are often suboptimal in terms of the number of neurons and connections, leading to slower simulation performance. Optimizing them while preserving functional equivalence --lossless optimization-- is an NP-hard problem, and current methods only provide lossy solutions. In this paper, we present a deterministic algorithm to optimize such neural networks in terms of neurons and connections while preserving functional equivalence. Moreover, to accelerate the compression of the neural network, we introduce an objective-aware algorithm that exploits representations that are shared among subproblems of the overall optimization. We demonstrate experimentally that we are able to reduce connections and neurons by up to 70% and 60%, respectively, in comparison to state-of-the-art. We also find that our objective-aware algorithm results in consistent speedups in optimization time, achieving up to 34.3x and 5.9x speedup relative to naive and caching solutions, respectively. Our methods are of practical relevance to applications such as high-throughput circuit simulation and placing neurosymbolic systems on the same hardware architecture.
****
Zachary Ankner,Cody Blakeney,Kartik Sreenivasan,Max Marion,Matthew L Leavitt,Mansheej Paul
Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models
In this work, we investigate whether small language models can determine high-quality subsets of large-scale text datasets that improve the performance of larger language models. While existing work has shown that pruning based on the perplexity of a larger model can yield high-quality data, we investigate whether smaller models can be used for perplexity-based pruning and how pruning is affected by the domain composition of the data being pruned. We demonstrate that for multiple dataset compositions, perplexity-based pruning of pretraining data can significantly improve downstream task performance: pruning based on perplexities computed with a 125 million parameter model improves the average performance on downstream tasks of a 3 billion parameter model by up to 2.04 and achieves up to a 1.45× reduction in pretraining steps to reach commensurate baseline performance.

Furthermore, we demonstrate that such perplexity-based data pruning also yields downstream performance gains in the over-trained and data-constrained regimes.
****
Alessandro Canevaro,Julian Schmidt,Mohammad Sajad Marvi,Hang Yu,Georg Martius,Julian Jordan
Advancing Out-of-Distribution Detection via Local Neuroplasticity
In the domain of machine learning, the assumption that training and test data share the same distribution is often violated in real-world scenarios, requiring effective out-of-distribution (OOD) detection.
This paper presents a novel OOD detection method that leverages the unique local neuroplasticity property of Kolmogorov-Arnold Networks (KANs).
Unlike traditional multilayer perceptrons, KANs exhibit local plasticity, allowing them to preserve learned information while adapting to new tasks.
Our method compares the activation patterns of a trained KAN against its untrained counterpart to detect OOD samples.
We validate our approach on benchmarks from image and medical domains, demonstrating superior performance and robustness compared to state-of-the-art techniques.
These results underscore the potential of KANs in enhancing the reliability of machine learning systems in diverse environments.
****
Sungmin Cha,Sungjun Cho,Dasol Hwang,Moontae Lee
Towards Robust and Parameter-Efficient Knowledge Unlearning for LLMs
Large Language Models (LLMs) have demonstrated strong reasoning and memorization capabilities via pretraining on massive textual corpora. However, this poses risk of privacy and copyright violations, highlighting the need for efficient machine unlearning methods that remove sensitive data without retraining from scratch. While Gradient Ascent (GA) is commonly used to unlearn by reducing the likelihood of generating unwanted content, it leads to unstable optimization and catastrophic forgetting of retrained knowledge. We find that combining GA with low-rank adaptation results in poor trade-offs between computational cost and generative performance. To address these challenges, we propose two novel techniques for robust and efficient unlearning for LLMs. First, we introduce Inverted Hinge Loss, which suppresses unwanted tokens while maintaining fluency by boosting the probability of the next most likely token. Second, we develop a data-adaptive initialization for LoRA adapters via low-rank approximation weighted with relative Fisher information, thereby focusing updates on parameters critical for removing targeted knowledge. Experiments on the Training Data Extraction Challenge dataset using GPT-Neo models as well as on the TOFU benchmark with Phi-1.5B and Llama2-7B models demonstrate that our approach effectively removes sensitive information while maintaining reasoning and generative capabilities with minimal impact. Our implementation can be found in https://github.com/csm9493/efficient-llm-unlearning.
****
Gen Luo,Yiyi Zhou,Yuxin Zhang,Xiawu Zheng,Xiaoshuai Sun,Rongrong Ji
Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models
In existing multimodal large language models (MLLMs), image resolution plays a significant role for granular visual recognition. However, directly increasing image resolution leads to expensive computational cost for MLLMs. In this paper, we reveal that a combination of low- and high-resolution visual features can efficiently mitigate this shortcoming. Based on this principle, we propose a novel and efficient method for MLLMs, termed Mixture-of-Resolution Adaptation (MRA). In particular, MRA adopts two visual pathways for images of different resolutions, where high-resolution visual information is embedded into the low-resolution pathway via the novel mixture-of-resolution adapters (MR-Adapters). This design also greatly reduces the input sequence length of MLLMs. To validate MRA, we apply it to a recent MLLM called LLaVA, and term the new model LLaVA-HR. We conduct extensive experiments on 17 vision-language (VL) tasks, which show that LLaVA-HR outperforms existing MLLMs on 15 VL tasks, e.g., +5.2\% on TextVQA.

More importantly, both training and inference of LLaVA-HR remain efficient with MRA, e.g., 20 training hours and faster inference speed than LLaVA-NeXT. Source codes are released at: https://github.com/luogen1996/LLaVA-HR.
****

Zekun Wang,Mingyang Yi,Shuchen Xue,Zhenguo Li,Ming Liu,Bing Qin,Zhi-Ming Ma
Improved Diffusion-based Generative Model with Better Adversarial Robustness
Diffusion Probabilistic Models (DPMs) have achieved significant success in generative tasks. However, their training and sampling processes suffer from the issue of distribution mismatch. During the denoising process, the input data distributions differ between the training and inference stages, potentially leading to inaccurate data generation. To obviate this, we analyze the training objective of DPMs and theoretically demonstrate that this mismatch can be alleviated through Distributionally Robust Optimization (DRO), which is equivalent to performing robustness-driven Adversarial Training (AT) on DPMs. Furthermore, for the recently proposed Consistency Model (CM), which distills the inference process of the DPM, we prove that its training objective also encounters the mismatch issue. Fortunately, this issue can be mitigated by AT as well. Based on these insights, we propose to conduct efficient AT on both DPM and CM. Finally, extensive empirical studies validate the effectiveness of AT in diffusion-based models. The code is available at https://github.com/kugwzk/AT_Diff.
****

Shiyu Wang,Jiawei LI,Xiaoming Shi,Zhou Ye,Baichuan Mo,Wenze Lin,Ju Shengtong,Zhixuan Chu,Ming Jin
TimeMixer++: A General Time Series Pattern Machine for Universal Predictive Analysis
Time series analysis plays a critical role in numerous applications, supporting tasks such as forecasting, classification, anomaly detection, and imputation. In this work, we present the time series pattern machine (TSPM), a model designed to excel in a broad range of time series tasks through powerful representation and pattern extraction capabilities. Traditional time series models often struggle to capture universal patterns, limiting their effectiveness across diverse tasks. To address this, we define multiple scales in the time domain and various resolutions in the frequency domain, employing various mixing strategies to extract intricate, task-adaptive time series patterns. Specifically, we introduce TimeMixer++, a general-purpose TSPM that processes multi-scale time series using (1) multi-resolution time imaging (MRTI), (2) time image decomposition (TID), (3) multi-scale mixing (MCM), and (4) multi-resolution mixing (MRM) to extract comprehensive temporal patterns. MRTI transforms multi-scale time series into multi-resolution time images, capturing patterns across both temporal and frequency domains. TID leverages dual-axis attention to extract seasonal and trend patterns, while MCM hierarchically aggregates these patterns across scales. MRM adaptively integrates all representations across resolutions. TimeMixer++ achieves state-of-the-art performance across 8 time series analytical tasks, consistently surpassing both general-purpose and task-specific models. Our work marks a promising step toward the next generation of TSPMs, paving the way for further advancements in time series analysis.
****

Georgy Ponimatkin,Martin Cífka,Tomas Soucek,Médéric Fourmy,Yann Labbé,Vladimir Petrik,Josef Sivic
6D Object Pose Tracking in Internet Videos for Robotic Manipulation
We seek to extract a temporally consistent 6D pose trajectory of a manipulated object from an Internet instructional video. This is a challenging set-up for current 6D pose estimation methods due to uncontrolled capturing conditions, subtle but dynamic object motions, and the fact that the exact mesh of the manipulated object is not known. To address these challenges, we present the following contributions. First, we develop a new method that estimates the 6D pose of any object in the input image without prior knowledge of the object itself. The method proceeds by (i) retrieving a CAD model similar to the depicted object from a large-scale model database, (ii) 6D aligning the retrieved CAD model with the input image, and (iii) grounding the absolute scale of the object with respect to the

scene. Second, we extract smooth 6D object trajectories from Internet videos by carefully tracking the detected objects across video frames. The extracted object trajectories are then retargeted via trajectory optimization into the configuration space of a robotic manipulator. Third, we thoroughly evaluate and ablate our 6D pose estimation method on YCB-V and HOPE-Video datasets as well as a new dataset of instructional videos manually annotated with approximate 6D object trajectories. We demonstrate significant improvements over existing state-of-the-art RGB 6D pose estimation methods. Finally, we show that the 6D object motion estimated from Internet videos can be transferred to a 7-axis robotic manipulator both in a virtual simulator as well as in a real world set-up. We also successfully apply our method to egocentric videos taken from the EPIC-KITCHENS dataset, demonstrating potential for Embodied AI applications.
****

Zhenyu Pan,Haozheng Luo,Manling Li,Han Liu
Chain-of-Action: Faithful and Multimodal Question Answering through Large Language Models
We present a Chain-of-Action (CoA) framework for multimodal and retrieval-augmented Question-Answering (QA). Compared to the literature, CoA overcomes two major challenges of current QA applications: (i) unfaithful hallucination that is inconsistent with real-time or domain facts and (ii) weak reasoning performance over compositional information. Our key contribution is a novel reasoning-retrieval mechanism that decomposes a complex question into a reasoning chain via systematic prompting and pre-designed actions. Methodologically, we propose three types of domain-adaptable `Plug-and-Play' actions for retrieving real-time information from heterogeneous sources. We also propose a multi-reference faith score to verify conflicts in the answers.
In addition, our system demonstrates that detecting the knowledge boundaries of LLMs can significantly reduce both LLM interaction frequency and tokens usage in QA tasks. Empirically, we exploit both public benchmarks and a Web3 case study to demonstrate the capability of CoA over other methods.
****

Laziz Abdullaev,Tan Minh Nguyen
Transformer Meets Twicing: Harnessing Unattended Residual Information
Transformer-based deep learning models have achieved state-of-the-art performance across numerous language and vision tasks. While the self-attention mechanism, a core component of transformers, has proven capable of handling complex data patterns, it has been observed that the representational capacity of the attention matrix degrades significantly across transformer layers, thereby hurting its overall performance. In this work, we leverage the connection between self-attention computations and low-pass non-local means (NLM) smoothing filters and propose the Twicing Attention, a novel attention mechanism that uses *kernel twicing procedure* in nonparametric regression to alleviate the low-pass behavior of associated NLM smoothing with compelling theoretical guarantees. This approach enables the extraction and reuse of meaningful information retained in the residuals following the imperfect smoothing operation at each layer. Our proposed method offers two key advantages over standard self-attention: 1) a provably slower decay of representational capacity and 2) improved accuracy across various data modalities and tasks. We empirically demonstrate the performance gains of our model over baseline transformers on multiple tasks and benchmarks, including image classification and language modeling, on both clean and corrupted data.
****

Shilin Lu,Zihan Zhou,Jiayou Lu,Yuanzhi Zhu,Adams Wai-Kin Kong
Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances
Current image watermarking methods are vulnerable to advanced image editing techniques enabled by large-scale text-to-image models. These models can distort embedded watermarks during editing, posing significant challenges to copyright protection. In this work, we introduce W-Bench, the first comprehensive benchmark designed to evaluate the robustness of watermarking methods against a wide range of image editing techniques, including image regeneration, global editing, local

editing, and image-to-video generation. Through extensive evaluations of eleven representative watermarking methods against prevalent editing techniques, we demonstrate that most methods fail to detect watermarks after such edits. To address this limitation, we propose VINE, a watermarking method that significantly enhances robustness against various image editing techniques while maintaining high image quality. Our approach involves two key innovations: (1) we analyze the frequency characteristics of image editing and identify that blurring distortions exhibit similar frequency properties, which allows us to use them as surrogate attacks during training to bolster watermark robustness; (2) we leverage a large-scale pretrained diffusion model SDXL-Turbo, adapting it for the watermarking task to achieve more imperceptible and robust watermark embedding. Experimental results show that our method achieves outstanding watermarking performance under various image editing techniques, outperforming existing methods in both image quality and robustness. Code is available at https://github.com/Shilin-LU/VINE
****

Yongxin Guo,Jingyu Liu,Mingda Li,Qingbin Liu,Xi Chen,Xiaoying Tang
TRACE: Temporal Grounding Video LLM via Causal Event Modeling
Video Temporal Grounding (VTG) is a crucial capability for video understanding models and plays a vital role in downstream tasks such as video browsing and editing.
To effectively handle various tasks simultaneously and enable zero-shot prediction, there is a growing trend in employing video LLMs for VTG tasks. However, current video LLM-based methods rely exclusively on natural language generation, lacking the ability to model the clear structure inherent in videos, which restricts their effectiveness in tackling VTG tasks. To address this issue, this paper first formally introduces causal event modeling framework, which represents video LLM outputs as sequences of events, and predict the current event using previous events, video inputs, and textural instructions. Each event consists of three components: timestamps, salient scores, and textual captions. We then propose a novel task-interleaved video LLM called TRACE to effectively implement the causal event modeling framework in practice.
The TRACE process visual frames, timestamps, salient scores, and text as distinct tasks, employing various encoders and decoding heads for each. Task tokens are arranged in an interleaved sequence according to the causal event modeling framework's formulation.
Extensive experiments on various VTG tasks and datasets demonstrate the superior performance of TRACE compared to state-of-the-art video LLMs. Our model and code are avaliable at \url{https://github.com/gyxxyg/TRACE}.
****

Ziyu Chen,Jiawei Yang,Jiahui Huang,Riccardo de Lutio,Janick Martinez Esturo,Boris Ivanovic,Or Litany,Zan Gojcic,Sanja Fidler,Marco Pavone,Li Song,Yue Wang
OmniRe: Omni Urban Scene Reconstruction
We introduce OmniRe, a comprehensive system for efficiently creating high-fidelity digital twins of dynamic real-world scenes from on-device logs. Recent methods using neural fields or Gaussian Splatting primarily focus on vehicles, hindering a holistic framework for all dynamic foregrounds demanded by downstream applications, e.g., the simulation of human behavior. OmniRe extends beyond vehicle modeling to enable accurate, full-length reconstruction of diverse dynamic objects in urban scenes. Our approach builds scene graphs on 3DGS and constructs multiple Gaussian representations in canonical spaces that model various dynamic actors, including vehicles, pedestrians, cyclists, and others. OmniRe allows holistically reconstructing any dynamic object in the scene, enabling advanced simulations (~60 Hz) that include human-participated scenarios, such as pedestrian behavior simulation and human-vehicle interaction. This comprehensive simulation capability is unmatched by existing methods. Extensive evaluations on the Waymo dataset show that our approach outperforms prior state-of-the-art methods quantitatively and qualitatively by a large margin. We further extend our results to 5 additional popular driving datasets to demonstrate its generalizability on common urban scenes. Code and results are available at [omnire](https://ziyc.github.io/omnire/).

****
Han Yu,Hanrui Lyu,YiXun Xu,Charlie Windolf,Eric Kenji Lee,Fan Yang,Andrew M Shel
ton,Olivier Winter,International Brain Laboratory,Eva L Dyer,Chandramouli Chandr
asekaran,Nicholas A. Steinmetz,Liam Paninski,Cole Lincoln Hurwitz
In vivo cell-type and brain region classification via multimodal contrastive lea
rning
Current electrophysiological approaches can track the activity of many neurons,
yet it is usually unknown which cell-types or brain areas are being recorded wit
hout further molecular or histological analysis. Developing accurate and scalabl
e algorithms for identifying the cell-type and brain region of recorded neurons
is thus crucial for improving our understanding of neural computation. In this w
ork, we develop a multimodal contrastive learning approach for neural data that
can be fine-tuned for different downstream tasks, including inference of cell-ty
pe and brain location. We utilize multimodal contrastive learning to jointly emb
ed the activity autocorrelations and extracellular waveforms of individual neuro
ns. We demonstrate that our embedding approach, Neuronal Embeddings via MultimOd
al Contrastive Learning (NEMO), paired with supervised fine-tuning, achieves sta
te-of-the-art cell-type classification for two opto-tagged datasets and brain re
gion classification for the public International Brain Laboratory Brain-wide Map
 dataset. Our method represents a promising step towards accurate cell-type and
brain region classification from electrophysiological recordings.
****
Yicong Li,Kuanjiu Zhou,Shuo Yu,Qiang Zhang,Renqiang Luo,Xiaodong Li,Feng Xia
Factor Graph-based Interpretable Neural Networks
Comprehensible neural network explanations are foundations for a better understa
nding of decisions, especially when the input data are infused with malicious pe
rturbations. Existing solutions generally mitigate the impact of perturbations t
hrough adversarial training, yet they fail to generate comprehensible explanatio
ns under unknown perturbations. To address this challenge, we propose AGAIN, a f
actor graph-based interpretable neural network, which is capable of generating c
omprehensible explanations under unknown perturbations. Instead of retraining li
ke previous solutions, the proposed AGAIN directly integrates logical rules by w
hich logical errors in explanations are identified and rectified during inferenc
e. Specifically, we construct the factor graph to express logical rules between
explanations and categories. By treating logical rules as exogenous knowledge, A
GAIN can identify incomprehensible explanations that violate real-world logic. F
urthermore, we propose an interactive intervention switch strategy rectifying ex
planations based on the logical guidance from the factor graph without learning
perturbations, which overcomes the inherent limitation of adversarial training-b
ased methods in defending only against known perturbations. Additionally, we the
oretically demonstrate the effectiveness of employing factor graph by proving th
at the comprehensibility of explanations is strongly correlated with factor grap
h. Extensive experiments are conducted on three datasets and experimental result
s illustrate the superior performance of AGAIN compared to state-of-the-art base
lines.
****
Jiahai Feng,Stuart Russell,Jacob Steinhardt
Monitoring Latent World States in Language Models with Propositional Probes
Language models (LMs) are susceptible to bias, sycophancy, backdoors, and other
tendencies that lead to unfaithful responses to the input context. Interpreting
internal states of LMs could help monitor and correct unfaithful behavior. We hy
pothesize that LMs faithfully represent their input contexts in a latent world m
odel, and we seek to extract these latent world states as logical propositions.
For example, given the input context ``Greg is a nurse. Laura is a physicist.'',
 we aim to decode the propositions WorksAs(Greg, nurse) and WorksAs(Laura, physi
cist) from the model's internal activations. To do so we introduce _propositiona
l probes_, which compositionally extract lexical concepts from token activations
 and bind them into propositions. Key to this is identifying a _binding subspace
_ in which bound tokens have high similarity (Greg $\leftrightarrow$ nurse) but
unbound ones do not (Greg $\not\leftrightarrow$ physicist). Despite only being t

rained on linguistically simple English templates, we find that propositional pr
obes generalize to inputs written as short stories and translated to Spanish. Mo
reover, in three settings where LMs respond unfaithfully to the input context---
prompt injections, backdoor attacks, and gender bias--- the decoded propositions
 remain faithful. This suggests that LMs often encode a faithful world model but
 decode it unfaithfully, which motivates the search for better interpretability
tools for monitoring LMs.
****
Yingzi Ma,Jiongxiao Wang,Fei Wang,Siyuan Ma,Jiazhao Li,Jinsheng Pan,Xiujun Li,Fu
rong Huang,Lichao Sun,Bo Li,Yejin Choi,Muhao Chen,Chaowei Xiao
Benchmarking Vision Language Model Unlearning via Fictitious Facial Identity Dat
aset
Machine unlearning has emerged as an effective strategy for forgetting specific
information in the training data. However, with the increasing integration of vi
sual data, privacy concerns in Vision Language Models (VLMs) remain underexplore
d. To address this, we introduce Facial Identity Unlearning Benchmark (FIUBench)
, a novel VLM unlearning benchmark designed to robustly evaluate the effectivene
ss of unlearning algorithms under the Right to be Forgotten setting. Specificall
y, we formulate the VLM unlearning task via constructing the Fictitious Facial I
dentity VQA dataset and apply a two-stage evaluation pipeline that is designed t
o precisely control the sources of information and their exposure levels. In ter
ms of evaluation, since VLM supports various forms of ways to ask questions with
 the same semantic meaning, we also provide robust evaluation metrics including
 membership inference attacks and carefully designed adversarial privacy attacks
 to evaluate the performance of algorithms. Through the evaluation of four base
line VLM unlearning algorithms within FIUBench, we find that all methods remain
limited in their unlearning performance, with significant trade-offs between mod
el utility and forget quality. Furthermore, our findings also highlight the impo
rtance of privacy attacks for robust evaluations. We hope FIUBench will drive pr
ogress in developing more effective VLM unlearning algorithms.
****
Eduard Gorbunov,Nazarii Tupitsa,Sayantan Choudhury,Alen Aliev,Peter Richtárik,Sa
muel Horváth,Martin Taká■
Methods for Convex (L_0,L_1)-Smooth Optimization: Clipping, Acceleration, and
Adaptivity
Due to the non-smoothness of optimization problems in Machine Learning, generali
zed smoothness assumptions have been gaining a lot of attention in recent years.
 One of the most popular assumptions of this type is (L_0,L_1)-smoothness (Zha
ng et al., 2020). In this paper, we focus on the class of (strongly) convex $(L_
0,L_1)$-smooth functions and derive new convergence guarantees for several exist
ing methods. In particular, we derive improved convergence rates for Gradient De
scent with (Smoothed) Gradient Clipping and for Gradient Descent with Polyak Ste
psizes. In contrast to the existing results, our rates do not rely on the standa
rd smoothness assumption and do not suffer from the exponential dependency on th
e initial distance to the solution. We also extend these results to the stochast
ic case under the over-parameterization assumption, propose a new accelerated me
thod for convex (L_0,L_1)-smooth optimization, and derive new convergence rat
es for Adaptive Gradient Descent (Malitsky and Mishchenko, 2020).
****
Mingyang Chen,Jiawei Du,Bo Huang,Yi Wang,Xiaobo Zhang,Wei Wang
Influence-Guided Diffusion for Dataset Distillation
Dataset distillation aims to streamline the training process by creating a compa
ct yet effective dataset for a much larger original dataset. However, existing m
ethods often struggle with distilling large, high-resolution datasets due to pro
hibitive resource costs and limited performance, primarily stemming from sample-
wise optimizations in the pixel space. Motivated by the remarkable capabilities
of diffusion generative models in learning target dataset distributions and cont
rollably sampling high-quality data tailored to user needs, we propose framing d
ataset distillation as a controlled diffusion generation task aimed at generatin
g data specifically tailored for effective training purposes. By establishing a

correlation between the overarching objective of dataset distillation and the tr
ajectory influence function, we introduce the Influence-Guided Diffusion (IGD) s
ampling framework to generate training-effective data without the need to retrai
n diffusion models. An efficient guided function is designed by leveraging the t
rajectory influence function as an indicator to steer diffusions to produce data
 with influence promotion and diversity enhancement. Extensive experiments show
that the training performance of distilled datasets generated by diffusions can
be significantly improved by integrating with our IGD method and achieving state
-of-the-art performance in distilling ImageNet datasets. Particularly, an except
ional result is achieved on the ImageNet-1K, reaching 60.3\% at IPC=50. Our code
 is available at https://github.com/mchen725/DD_IGD.
****
Chen Bo Calvin Zhang,Zhang-Wei Hong,Aldo Pacchiano,Pulkit Agrawal
ORSO: Accelerating Reward Design via Online Reward Selection and Policy Optimiza
tion
Reward shaping is critical in reinforcement learning (RL), particularly for comp
lex tasks where sparse rewards can hinder learning. However, choosing effective
shaping rewards from a set of reward functions in a computationally efficient ma
nner remains an open challenge. We propose Online Reward Selection and Policy Op
timization (ORSO), a novel approach that frames the selection of shaping reward
function as an online model selection problem. ORSO automatically identifies per
formant shaping reward functions without human intervention with provable regret
 guarantees. We demonstrate ORSO's effectiveness across various continuous contr
ol tasks. Compared to prior approaches, ORSO significantly reduces the amount of
 data required to evaluate a shaping reward function, resulting in superior data
 efficiency and a significant reduction in computational time (up to 8×). ORSO c
onsistently identifies high-quality reward functions outperforming prior methods
 by more than 50% and on average identifies policies as performant as the ones l
earned using manually engineered reward functions by domain experts.
****
Pier Giuseppe Sessa,Robert Dadashi-Tazehozi,Leonard Hussenot,Johan Ferret,Nino V
ieillard,Alexandre Rame,Bobak Shahriari,Sarah Perrin,Abram L. Friesen,Geoffrey C
ideron,Sertan Girgin,Piotr Stanczyk,Andrea Michi,Danila Sinopalnikov,Sabela Ramo
s Garea,Amélie Héliou,Aliaksei Severyn,Matthew Hoffman,Nikola Momchev,Olivier Ba
chem
BOND: Aligning LLMs with Best-of-N Distillation
Reinforcement learning from human feedback (RLHF) is a key driver of quality and
 safety in state-of-the-art large language models.
Yet, a surprisingly simple and strong inference-time strategy is Best-of-N sampl
ing that selects the best generation among N candidates.
In this paper, we propose Best-of-N Distillation (BOND), a novel RLHF algorithm
that seeks to emulate Best-of-N but without its significant computational overhe
ad at inference time. Specifically, BOND is a distribution matching algorithm th
at forces the distribution of generations from the policy to get closer to the B
est-of-N distribution. We use the Jeffreys divergence (a linear combination of f
orward and backward KL) to balance between mode-covering and mode-seeking behavi
or, and derive an iterative formulation that utilizes a moving anchor for effici
ency. We demonstrate the effectiveness of our approach and several design choice
s through experiments on abstractive summarization and Gemma models.
****
Ekdeep Singh Lubana,Kyogo Kawaguchi,Robert P. Dick,Hidenori Tanaka
A Percolation Model of Emergence: Analyzing Transformers Trained on a Formal Lan
guage
Increase in data, size, or compute can lead to sudden learning of specific capab
ilities by a neural network---a phenomenon often called "emergence". Beyond scie
ntific understanding, establishing the causal factors underlying such emergent c
apabilities is crucial to enable risk regulation frameworks for AI. In this work
, we seek inspiration from study of emergent properties in other fields and prop
ose a phenomenological definition for the concept in the context of neural netwo
rks. Our definition implicates the acquisition of general regularities underlyin

g the data-generating process as a cause of sudden performance growth for specif
ic, narrower tasks. We empirically investigate this definition by proposing an e
xperimental system grounded in a context-sensitive formal language, and find tha
t Transformers trained to perform tasks on top of strings from this language ind
eed exhibit emergent capabilities. Specifically, we show that once the language'
s underlying grammar and context-sensitivity inducing regularities are learned b
y the model, performance on narrower tasks suddenly begins to improve. We then a
nalogize our network's learning dynamics with the process of percolation on a bi
partite graph, establishing a formal phase transition model that predicts the sh
ift in the point of emergence observed in our experiments when intervening on th
e data regularities. Overall, our experimental and theoretical frameworks yield
a step towards better defining, characterizing, and predicting emergence in neur
al networks.
****

Feng yan,Weixin Luo,Yujie Zhong,Yiyang Gan,Lin Ma
CO-MOT: Boosting End-to-end Transformer-based Multi-Object Tracking via Coopetit
ion Label Assignment and Shadow Sets
Existing end-to-end Multi-Object Tracking (e2e-MOT) methods have not surpassed n
on-end-to-end tracking-by-detection methods. One possible reason lies in the tra
ining label assignment strategy that consistently binds the tracked objects with
 tracking queries and assigns few newborns to detection queries. Such an assignm
ent, with one-to-one bipartite matching, yields an unbalanced training, _i.e._,
scarce positive samples for detection queries, especially for an enclosed scene
with the majority of the newborns at the beginning of videos. As such, e2e-MOT w
ill incline to generate a tracking terminal without renewal or re-initialization
, compared to other tracking-by-detection methods.
To alleviate this problem, we propose **Co-MOT**, a simple yet effective method
to facilitate e2e-MOT by a novel coopetition label assignment with a shadow conc
ept. Specifically, we add tracked objects to the matching targets for detection
queries when performing the label assignment for training the intermediate decod
ers. For query initialization, we expand each query by a set of shadow counterpa
rts with limited disturbance to itself.
With extensive ablation studies, Co-MOT achieves superior performances without e
xtra costs, _e.g._, 69.4% HOTA on DanceTrack and 52.8% TETA on BDD100K. Impressi
vely, Co-MOT only requires 38% FLOPs of MOTRv2 with comparable performances, res
ulting in the 1.4× faster inference speed. Source code is publicly available at
[GitHub](https://github.com/BingfengYan/CO-MOT).
****

Itai Shufaro,Nadav Merlis,Nir Weinberger,Shie Mannor
On Bits and Bandits: Quantifying the Regret-Information Trade-off
In many sequential decision problems, an agent performs a repeated task. He then
 suffers regret and obtains information that he may use in the following rounds.
 However, sometimes the agent may also obtain information and avoid suffering re
gret by querying external sources. We study the trade-off between the informatio
n an agent accumulates and the regret it suffers. We invoke information-theoreti
c methods for obtaining regret lower bounds, that also allow us to easily re-der
ive several known lower bounds. We introduce the first Bayesian regret lower bou
nds that depend on the information an agent accumulates. We also prove regret up
per bounds using the amount of information the agent accumulates. These bounds s
how that information measured in bits, can be traded off for regret, measured in
 reward. Finally, we demonstrate the utility of these bounds in improving the pe
rformance of a question-answering task with large language models, allowing us t
o obtain valuable insights.
****

Xinhao Zhong,Bin Chen,Hao Fang,Xulin Gu,Shu-Tao Xia,EN-HUI YANG
Going Beyond Feature Similarity: Effective Dataset distillation based on Class-a
ware Conditional Mutual Information
Dataset distillation (DD) aims to minimize the time and memory consumption neede
d for training deep neural networks on large datasets, by creating a smaller syn
thetic dataset that has similar performance to that of the full real dataset. Ho

wever, current dataset distillation methods often result in synthetic datasets t
hat are excessively difficult for networks to learn from, due to the compression
 of a substantial amount of information from the original data through metrics m
easuring feature similarity, e,g., distribution matching (DM). In this work, we
introduce conditional mutual information (CMI) to assess the class-aware complex
ity of a dataset and propose a novel method by minimizing CMI. Specifically, we
minimize the distillation loss while constraining the class-aware complexity of
the synthetic dataset by minimizing its empirical CMI from the feature space of
pre-trained networks, simultaneously. Conducting on a thorough set of experiment
s, we show that our method can serve as a general regularization method to exist
ing DD methods and improve the performance and training efficiency.
****
Utsav Singh,Vinay P. Namboodiri
PEAR: Primitive Enabled Adaptive Relabeling for Boosting Hierarchical Reinforcem
ent Learning
Hierarchical reinforcement learning (HRL) has the potential to solve complex lon
g horizon tasks using temporal abstraction and increased exploration. However, h
ierarchical agents are difficult to train due to inherent non-stationarity. We p
resent primitive enabled adaptive relabeling (PEAR), a two-phase approach where
we first perform adaptive relabeling on a few expert demonstrations to generate
efficient subgoal supervision, and then jointly optimize HRL agents by employing
 reinforcement learning (RL) and imitation learning (IL). We perform theoretical
 analysis to bound the sub-optimality of our approach and derive a joint optimiz
ation framework using RL and IL. Since PEAR utilizes only a few expert demonstra
tions and considers minimal limiting assumptions on the task structure, it can b
e easily integrated with typical off-policy \RL algorithms to produce a practica
l HRL approach. We perform extensive experiments on challenging environments and
 show that PEAR is able to outperform various hierarchical and non-hierarchical
baselines and achieve upto 80% success rates in complex sparse robotic control t
asks where other baselines typically fail to show significant progress. We also
perform ablations to thoroughly analyze the importance of our various design cho
ices. Finally, we perform real world robotic experiments on complex tasks and de
monstrate that PEAR consistently outperforms the baselines.
****
Sherwin Bahmani,Ivan Skorokhodov,Aliaksandr Siarohin,Willi Menapace,Guocheng Qia
n,Michael Vasilkovsky,Hsin-Ying Lee,Chaoyang Wang,Jiaxu Zou,Andrea Tagliasacchi,
David B. Lindell,Sergey Tulyakov
VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control
Modern text-to-video synthesis models demonstrate coherent, photorealistic gener
ation of complex videos from a text description. However, most existing models l
ack fine-grained control over camera movement, which is critical for downstream
applications related to content creation, visual effects, and 3D vision. Recentl
y, new methods demonstrate the ability to generate videos with controllable came
ra poses---these techniques leverage pre-trained U-Net-based diffusion models th
at explicitly disentangle spatial and temporal generation. Still, no existing ap
proach enables camera control for new, transformer-based video diffusion models
that process spatial and temporal information jointly. Here, we propose to tame
video transformers for 3D camera control using a ControlNet-like conditioning me
chanism that incorporates spatiotemporal camera embeddings based on Plucker coor
dinates. The approach demonstrates state-of-the-art performance for controllable
 video generation after fine-tuning on the RealEstate10K dataset. To the best of
 our knowledge, our work is the first to enable camera control for transformer-b
ased video diffusion models.
****
Andrew Ying,Zhichen Zhao,Ronghui Xu
Incremental Causal Effect for Time to Treatment Initialization
We consider time to treatment initialization. This can commonly occur in prevent
ive medicine, such as disease screening and vaccination; it can also occur with
non-fatal health conditions such as HIV infection without the onset of AIDS. Whi
le traditional causal inference focused on ‘when to treat’ and its effects, incl

uding their possible dependence on subject characteristics, we consider the incr
emental causal effect when the intensity of time to treatment initialization is
intervened upon. We provide identification of the incremental causal effect with
out the commonly required positivity assumption, as well as an estimation framew
ork using inverse probability weighting. We illustrate our approach via simulati
on, and apply it to a rheumatoid arthritis study to evaluate the incremental eff
ect of time to start methotrexate on joint pain.
****

Vibhhu Sharma,Bryan Wilder

Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resour
ces

Machine learning is increasingly used to select which individuals receive limite
d-resource interventions in domains such as human services, education, developme
nt, and more. However, it is often not apparent what the right quantity is for m
odels to predict. In particular, policymakers rarely have access to data from a
randomized controlled trial (RCT) that would enable accurate estimates of treatm
ent effects -- which individuals would benefit more from the intervention. Obser
vational data is more likely to be available, creating a substantial risk of bia
s in treatment effect estimates. Practitioners instead commonly use a technique
termed "risk-based targeting" where the model is just used to predict each indi
vidual's status quo outcome (an easier, non-causal task). Those with higher pred
icted risk are offered treatment. There is currently almost no empirical evidenc
e to inform which choices lead to the most effect machine learning-informed targ
eting strategies in social domains. In this work, we use data from 5 real-world
RCTs in a variety of domains to empirically assess such choices. We find that wh
en treatment effects can be estimated reliably (which we simulate by using direc
t outcome observations), treatment effect based targeting substantially outperfo
rms risk-based targeting, even when treatment effect estimates are biased. Moreo
ver, these results hold even when the policymaker has strong normative preferenc
es for assisting higher-risk individuals. However, when treatment effects must b
e predicted from features alone (as is always the case in practice), performance
can degrade significantly due to limited data making it difficult to learn accu
rate mappings from features to treatment effects. Our results suggest treatment
effect targeting has significant potential benefits, but realizing these benefit
s requires careful attention to model training and validation.
****

Tam Le,Jerome Malick

Universal generalization guarantees for Wasserstein distributionally robust mode
ls

Distributionally robust optimization has emerged as an attractive way to train r
obust machine learning models, capturing data uncertainty and distribution shift
s. Recent statistical analyses have proved that generalization guarantees of rob
ust models based on the Wasserstein distance have generalization guarantees that
do not suffer from the curse of dimensionality. However, these results are eith
er approximate, obtained in specific cases, or based on assumptions difficult to
verify in practice. In contrast, we establish exact generalization guarantees t
hat cover a wide range of cases, with arbitrary transport costs and parametric l
oss functions, including deep learning objectives with nonsmooth activations. We
complete our analysis with an excess bound on the robust objective and an exten
sion to Wasserstein robust models with entropic regularizations.
****

Stephane Rivaud,Louis Fournier,Thomas Pumir,Eugene Belilovsky,Michael Eickenberg
,Edouard Oyallon

PETRA: Parallel End-to-end Training with Reversible Architectures

Reversible architectures have been shown to be capable of performing on par with
their non-reversible architectures, being applied in deep learning for memory s
avings and generative modeling. In this work, we show how reversible architectur
es can solve challenges in parallelizing deep model training. We introduce PETRA
, a novel alternative to backpropagation for parallelizing gradient computations
. PETRA facilitates effective model parallelism by enabling stages (i.e., a set

of layers) to compute independently on different devices, while only needing to communicate activations and gradients between each other. By decoupling the forward and backward passes and keeping a single updated version of the parameters, the need for weight stashing is also removed. We develop a custom autograd-like training framework for PETRA, and we demonstrate its effectiveness on standard computer vision benchmarks, achieving competitive accuracies comparable to backpropagation using ResNet-18, ResNet-34, and ResNet-50 models.
****
Benjamin Frederick Spector,Simran Arora,Aaryan Singhal,Arjun Parthasarathy,Daniel Y Fu,Christopher Re
ThunderKittens: Simple, Fast, and $\textit{Adorable}$ Kernels
The challenge of mapping AI architectures to GPU hardware is creating a critical bottleneck in AI progress. Despite substantial efforts, hand-written custom kernels fail to meet their theoretical performance thresholds, even on well-established operations like linear attention. The diverse capabilities of GPUs suggests we might we need a wide variety of techniques to achieve high performance. However, our work explores if a small number of key abstractions can drastically simplify the process. We present ThunderKittens (TK), a framework for writing performant AI kernels while remaining easy to use. Our abstractions map to the three levels of the GPU hierarchy: (1) at the warp-level, we provide 16x16 matrix tiles as basic data structures and PyTorch-like operations, (2) at the thread-block level, we provide templates for asynchronously overlapping operations, and (3) at the grid-level, TK helps hide block launch, tear-down, and memory costs. We show the value of TK by providing simple & diverse kernels that match or outperform prior art. We match CuBLAS and FlashAttention-3 on GEMM and attention inference performance and outperform the strongest baselines by $10-40$\% on attention backwards, $8\times$ on state space models, and $14\times$ on linear attention.
****
Emily Cheng,Diego Doimo,Corentin Kervadec,Iuri Macocco,Lei Yu,Alessandro Laio,Marco Baroni
Emergence of a High-Dimensional Abstraction Phase in Language Transformers
A language model (LM) is a mapping from a linguistic context to an output token. However, much remains to be known about this mapping, including how its geometric properties relate to its function. We take a high-level geometric approach to its analysis, observing, across five pre-trained transformer-based LMs and three input datasets, a distinct phase characterized by high intrinsic dimensionality. During this phase, representations (1) correspond to the first full linguistic abstraction of the input; (2) are the first to viably transfer to downstream tasks; (3) predict each other across different LMs. Moreover, we find that an earlier onset of the phase strongly predicts better language modelling performance. In short, our results suggest that a central high-dimensionality phase underlies core linguistic processing in many common LM architectures.
****
Zheng Wei Lim,Nitish Gupta,Honglin Yu,Trevor Cohn
Mufu: Multilingual Fused Learning for Low-Resource Translation with LLM
Multilingual large language models (LLMs) are great translators, but this is largely limited to high-resource languages. For many LLMs, translating in and out of low-resource languages remains a challenging task. To maximize data efficiency in this low-resource setting, we introduce Mufu, which includes a selection of automatically generated multilingual candidates and an instruction to correct inaccurate translations in the prompt. Mufu prompts turn a translation task into a postediting one, and seek to harness the LLM’s reasoning capability with auxiliary translation candidates, from which the model is required to assess the input quality, align the semantics cross-lingually, copy from relevant inputs and override instances that are incorrect. Our experiments on En-XX translations over the Flores-200 dataset show LLMs finetuned against Mufu-style prompts are robust to poor quality auxiliary translation candidates, achieving performance superior to NLLB 1.3B distilled model in 64% of low- and very-low-resource language pairs. We then distill these models to reduce inference cost, while maintaining on average 3.1 chrF improvement over finetune-only baseline in low-resource translat

ions.
****
Qiang Liu,Huiqiao Fu,Kaiqiang Tang,Chunlin Chen,Daoyi Dong
PN-GAIL: Leveraging Non-optimal Information from Imperfect Demonstrations
Imitation learning aims at constructing an optimal policy by emulating expert de
monstrations. However, the prevailing approaches in this domain typically presum
e that the demonstrations are optimal, an assumption that seldom holds true in t
he complexities of real-world applications. The data collected in practical scen
arios often contains imperfections, encompassing both optimal and non-optimal ex
amples. In this study, we propose Positive-Negative Generative Adversarial Imita
tion Learning (PN-GAIL), a novel approach that falls within the framework of Gen
erative Adversarial Imitation Learning (GAIL). PN-GAIL innovatively leverages no
n-optimal information from imperfect demonstrations, allowing the discriminator
to comprehensively assess the positive and negative risks associated with these
demonstrations. Furthermore, it requires only a small subset of labeled confiden
ce scores. Theoretical analysis indicates that PN-GAIL deviates from the non-opt
imal data while mimicking imperfect demonstrations. Experimental results demonst
rate that PN-GAIL surpasses conventional baseline methods in dealing with imperf
ect demonstrations, thereby significantly augmenting the practical utility of im
itation learning in real-world contexts. Our codes are available at https://gith
ub.com/QiangLiuT/PN-GAIL.
****
SUBBA REDDY OOTA,Khushbu Pahwa,mounika marreddy,Maneesh Kumar Singh,Manish Gupta
,Bapi Raju Surampudi
Multi-modal brain encoding models for multi-modal stimuli
Despite participants engaging in unimodal stimuli, such as watching images or si
lent videos, recent work has demonstrated that multi-modal Transformer models ca
n predict visual brain activity impressively well, even with incongruent modalit
y representations. This raises the question of how accurately these multi-modal
models can predict brain activity when participants are engaged in multi-modal s
timuli. As these models grow increasingly popular, their use in studying neural
activity provides insights into how our brains respond to such multi-modal natur
alistic stimuli, i.e., where it separates and integrates information across moda
lities through a hierarchy of early sensory regions to higher cognition (languag
e regions). We investigate this question by using multiple unimodal and two type
s of multi-modal models—cross-modal and jointly pretrained—to determine which ty
pe of models is more relevant to fMRI brain activity when participants are engag
ed in watching movies (videos with audio). We observe that both types of multi-m
odal models show improved alignment in several language and visual regions. This
 study also helps in identifying which brain regions process unimodal versus mul
ti-modal information. We further investigate the contribution of each modality t
o multi-modal alignment by carefully removing unimodal features one by one from
multi-modal representations, and find that there is additional information beyon
d the unimodal embeddings that is processed in the visual and language regions.
Based on this investigation, we find that while for cross-modal models, their br
ain alignment is partially attributed to the video modality; for jointly pretrai
ned models, it is partially attributed to both the video and audio modalities. T
hese findings serve as strong motivation for the neuro-science community to inve
stigate the interpretability of these models for deepening our understanding of
multi-modal information processing in brain.
****
Hannes Stark,Bowen Jing,Tomas Geffner,Jason Yim,Tommi Jaakkola,Arash Vahdat,Kars
ten Kreis
ProtComposer: Compositional Protein Structure Generation with 3D Ellipsoids
We develop ProtComposer to generate protein structures conditioned on spatial pr
otein layouts that are specified via a set of 3D ellipsoids capturing substructu
re shapes and semantics. At inference time, we condition on ellipsoids that are
hand-constructed, extracted from existing proteins, or from a statistical model,
 with each option unlocking new capabilities. Hand-specifying ellipsoids enables
 users to control the location, size, orientation, secondary structure, and appr

oximate shape of protein substructures. Conditioning on ellipsoids of existing p
roteins enables redesigning their substructure's connectivity or editing substru
cture properties. By conditioning on novel and diverse ellipsoid layouts from a
simple statistical model, we improve protein generation with expanded Pareto fro
ntiers between designability, novelty, and diversity. Further, this enables samp
ling designable proteins with a helix-fraction that matches PDB proteins, unlike
 existing generative models that commonly oversample conceptually simple helix b
undles. Code is available at https://github.com/NVlabs/protcomposer.
****

Zhi Gao,Bofei Zhang,Pengxiang Li,Xiaojian Ma,Tao Yuan,Yue Fan,Yuwei Wu,Yunde Jia
,Song-Chun Zhu,Qing Li

Multi-modal Agent Tuning: Building a VLM-Driven Agent for Efficient Tool Usage

The advancement of large language models (LLMs) prompts the development of multi
-modal agents, which are used as a controller to call external tools, providing
a feasible way to solve practical tasks. In this paper, we propose a multi-modal
 agent tuning method that automatically generates multi-modal tool-usage data an
d tunes a vision-language model (VLM) as the controller for powerful tool-usage
reasoning. To preserve the data quality, we prompt the GPT-4o mini model to gene
rate queries, files, and trajectories, followed by query-file and trajectory ver
ifiers. Based on the data synthesis pipeline, we collect the MM-Traj dataset tha
t contains 20K tasks with trajectories of tool usage. Then, we develop the T3-Ag
ent via Trajectory Tuning on VLMs for Tool usage using MM-Traj. Evaluations on t
he GTA and GAIA benchmarks show that the T3-Agent consistently achieves improvem
ents on two popular VLMs: MiniCPM-V-8.5B and Qwen2-VL-7B, which outperforms untr
ained VLMs by 20%, showing the effectiveness of the proposed data synthesis pipe
line, leading to high-quality data for tool-usage capabilities.
****

Zecheng Li,Wengang Zhou,Weichao Zhao,Kepeng Wu,Hezhen Hu,Houqiang Li

Uni-Sign: Toward Unified Sign Language Understanding at Scale

Sign language pre-training has gained increasing attention for its ability to en
hance performance across various sign language understanding (SLU) tasks. Howeve
r, existing methods often suffer from a gap between pre-training and fine-tuning
, leading to suboptimal results. To address this, we propose Uni-Sign, a unified
 pre-training framework that eliminates the gap between pre-training and downstr
eam SLU tasks through a large-scale generative pre-training strategy and a novel
 fine-tuning paradigm. First, we introduce CSL-News, a large-scale Chinese Sign
Language (CSL) dataset containing 1,985 hours of video paired with textual annot
ations, which enables effective large-scale pre-training. Second, Uni-Sign unifi
es SLU tasks by treating downstream tasks as a single sign language translation
(SLT) task during fine-tuning, ensuring seamless knowledge transfer between pre-
training and fine-tuning. Furthermore, we incorporate a prior-guided fusion (PGF
) module and a score-aware sampling strategy to efficiently fuse pose and RGB in
formation, addressing keypoint inaccuracies and improving computational efficien
cy. Extensive experiments across multiple SLU benchmarks demonstrate that Uni-S
ign achieves state-of-the-art performance across multiple downstream SLU tasks.
Dataset and code are available at github.com/ZechengLi19/Uni-Sign.
****

Runyi Zhao,Sheng Xu,Bo Yue,Guiliang Liu

Toward Exploratory Inverse Constraint Inference with Generative Diffusion Verifi
ers

An important prerequisite for safe control is aligning the policy with the under
lying constraints in the environment. In many real-world applications, due to th
e difficulty of manually specifying these constraints, existing works have propo
sed recovering constraints from expert demonstrations by solving the Inverse Con
straint Learning (ICL) problem. However, ICL is inherently ill-posed, as multipl
e constraints can equivalently explain the experts' preferences, making the opti
mal solutions not uniquely identifiable. In this work, instead of focusing solel
y on a single constraint, we propose the novel approach of Exploratory ICL (ExIC
L). The goal of ExICL is to recover a diverse set of feasible constraints, there
by providing practitioners the flexibility to select the most appropriate constr

aint based on the practical needs of deployment. To achieve this goal, we design a generative diffusion verifier that guides the trajectory generation process using the probabilistic representation of an optimal constrained policy. By comparing these decisions with those made by expert agents, we can efficiently verify a candidate constraint. Driven by the verification feedback, ExICL implements an exploratory constraint update mechanism that strategically facilitates diversity within the collection of feasible constraints. Our empirical results demonstrate that ExICL can seamlessly and reliably generalize across different tasks and environments. The code is available at https://github.com/ZhaoRunyi/ExICL.
****

Huiqing QI,Xiaoliu Luo,Tingting Li,Fang Li
Edge-aware Image Smoothing with Relative Wavelet Domain Representation
Image smoothing is a fundamental technique in image processing, designed to eliminate perturbations and textures while preserving dominant structures. It plays a pivotal role in numerous high-level computer vision tasks. More recently, both traditional and deep learning-based smoothing methods have been developed. However, existing algorithms frequently encounter issues such as gradient reversals and halo artifacts. Furthermore, the smoothing strength of deep learning-based models, once trained, cannot be adjusted for adapting different complexity levels of textures. These limitations stem from the inability of previous approaches to achieve an optimal balance between smoothing intensity and edge preservation. Consequently, image smoothing while maintaining edge integrity remains a significant challenge. To address these challenges, we propose a novel edge-aware smoothing model that leverages a relative wavelet domain representation. Specifically, by employing wavelet transformation, we introduce a new measure, termed Relative Wavelet Domain Representation (RWDR), which effectively distinguishes between textures and structures. Additionally, we present an innovative edge-aware scale map that is incorporated into the adaptive bilateral filter, facilitating mutual guidance in the smoothing process. This paper provides complete theoretical derivations for solving the proposed non-convex optimization model. Extensive experiments substantiate that our method has a competitive superiority with previous algorithms in edge-preserving and artifact removal. Visual and numerical comparisons further validate the effectiveness and efficiency of our approach in several applications of image smoothing.
****

Samuel Kessler,Tam Le,Vu Nguyen
SAVA: Scalable Learning-Agnostic Data Valuation
Selecting data for training machine learning models is crucial since large, web-scraped, real datasets contain noisy artifacts that affect the quality and relevance of individual data points. These noisy artifacts will impact model performance. We formulate this problem as a data valuation task, assigning a value to data points in the training set according to how similar or dissimilar they are to a clean and curated validation set. Recently, *LAVA* (Just et al., 2023) demonstrated the use of optimal transport (OT) between a large noisy training dataset and a clean validation set, to value training data efficiently, without the dependency on model performance. However, the *LAVA* algorithm requires the entire dataset as an input, this limits its application to larger datasets. Inspired by the scalability of stochastic (gradient) approaches which carry out computations on *batches* of data points instead of the entire dataset, we analogously propose *SAVA*, a scalable variant of *LAVA* with its computation on batches of data points. Intuitively, *SAVA* follows the same scheme as *LAVA* which leverages the hierarchically defined OT for data valuation. However, while *LAVA* processes the whole dataset, *SAVA* divides the dataset into batches of data points, and carries out the OT problem computation on those batches. Moreover, our theoretical derivations on the trade-off of using entropic regularization for OT problems include refinements of prior work. We perform extensive experiments, to demonstrate that *SAVA* can scale to large datasets with millions of data points and does not trade off data valuation performance. Our Github repository is available at \url{https://github.com/skezle/sava}.
****

Michael Chen,A. Pavan,N. V. Vinodchandran,Ruosong Wang,Lin Yang
Regret-Optimal List Replicable Bandit Learning: Matching Upper and Lower Bounds
This paper investigates *list replicability* [Dixon et al., 2023] in the context of multi-armed (also linear) bandits (MAB). We define an algorithm A for MAB to be (ℓ,δ)-list replicable if with probability at least $1-\delta$, A has at most ℓ traces in independent executions even with different random bits, where a trace means sequence of arms played during an execution. For k-armed bandits, although the total number of traces can be $\Omega(k^T)$ for a time horizon T, we present several surprising upper bounds that either independent of or logarithmic of T: (1) a $(2^{k},\delta)$-list replicable algorithm with near-optimal regret, $\widetilde{O}({\sqrt{kT}})$, (2) a $(O(k/\delta),\delta)$-list replicable algorithm with regret $\widetilde{O}\left(\frac{k}{\delta}\sqrt{kT}\right)$, (3) a $((k+1)^{B-1}, \delta)$-list replicable algorithm with regret $\widetilde{O}(k^{\frac{3}{2}}T^{{\frac{1}{2}}+2^{-(B+1)}})$ for any integer $B>1$. On the other hand, for the sublinear regret regime, we establish a matching lower bound on the list complexity (parameter ℓ). We prove that there is no $(k-1,\delta)$-list replicable algorithm with $o(T)$-regret. This is optimal in list complexity in the sub-linear regret regime as there is a $(k, 0)$-list replicable algorithm with $O(T^{2/3})$-regret. We further show that for linear bandits with d-dimensional features, there is a $\widetilde{O}(d^2T^{1/2+2^{-(B+1)}})$-regret algorithm with $((2d+1)^{B-1},\delta)$-list replicability, for $B>1$, even when the number of possible arms can be infinite.
****
Jingbo Sun,Songjun Tu,qichao Zhang,Haoran Li,Xin Liu,Yaran Chen,Ke Chen,Dongbin Zhao
Unsupervised Zero-Shot Reinforcement Learning via Dual-Value Forward-Backward Representation
Online unsupervised reinforcement learning (URL) can discover diverse skills via reward-free pre-training and exhibits impressive downstream task adaptation abilities through further fine-tuning.
However, online URL methods face challenges in achieving zero-shot generalization, i.e., directly applying pre-trained policies to downstream tasks without additional planning or learning.
In this paper, we propose a novel Dual-Value Forward-Backward representation (DVFB) framework with a contrastive entropy intrinsic reward to achieve both zero-shot generalization and fine-tuning adaptation in online URL.
On the one hand, we demonstrate that poor exploration in forward-backward representations can lead to limited data diversity in online URL, impairing successor measures, and ultimately constraining generalization ability.
To address this issue, the DVFB framework learns successor measures through a skill value function while promoting data diversity through an exploration value function, thus enabling zero-shot generalization.
On the other hand, and somewhat surprisingly, by employing a straightforward dual-value fine-tuning scheme combined with a reward mapping technique, the pre-trained policy further enhances its performance through fine-tuning on downstream tasks, building on its zero-shot performance.
Through extensive multi-task generalization experiments, DVFB demonstrates both superior zero-shot generalization (outperforming on all 12 tasks) and fine-tuning adaptation (leading on 10 out of 12 tasks) abilities, surpassing state-of-the-art URL methods.
****
Benedikt Alkin,Lukas Miklautz,Sepp Hochreiter,Johannes Brandstetter
MIM-Refiner: A Contrastive Learning Boost from Intermediate Pre-Trained Masked Image Modeling Representations
We introduce MIM (Masked Image Modeling)-Refiner, a contrastive learning boost for pre-trained MIM models. MIM-Refiner is motivated by the insight that strong representations within MIM models generally reside in intermediate layers. Accordingly, MIM-Refiner leverages multiple instance discrimination (ID) heads that are connected to different intermediate layers. In each head, a nearest neighbor ID objective constructs clusters that capture semantic information which improves

performance on downstream tasks, including off-the-shelf and fine-tuning settin
gs.

The refinement process is short and simple - yet highly effective. Within a few
epochs, we refine the features of MIM models from subpar to state-of-the-art, o
ff-the-shelf features. Refining a ViT-H, pre-trained with data2vec 2.0 on ImageN
et-1K, sets a new state-of-the-art in linear probing (84.7\%) and low-shot class
ification among models that are pre-trained on ImageNet-1K. MIM-Refiner efficien
tly combines the advantages of MIM and ID objectives, enabling scaling ID object
ives to billion parameter models using relatively little compute. MIM-Refiner co
mpares favorably against previous state-of-the-art SSL models on various benchma
rks such as low-shot classification, long-tailed classification and semantic seg
mentation.
****
Mazda Moayeri,Sriram Balasubramanian,Samyadeep Basu,Priyatham Kattakinda,Atoosa
Chegini,Robert Brauneis,Soheil Feizi
Rethinking Artistic Copyright Infringements In the Era Of Text-to-Image Generati
ve Models
The advent of text-to-image generative models has led artists to worry that thei
r individual styles may be copied, creating a pressing need to reconsider the la
ck of protection for artistic styles under copyright law. This requires answerin
g challenging questions, like what defines style and what constitutes style infr
ingment. In this work, we build on prior legal scholarship to develop an automat
ic and interpretable framework to \emph{quantitatively} assess style infringemen
t. Our methods hinge on a simple logical argument: if an artist's works can cons
istently be recognized as their own, then they have a unique style. Based on thi
s argument, we introduce ArtSavant, a practical (i.e., efficient and easy to und
erstand) tool to (i) determine the unique style of an artist by comparing it to
a reference corpus of works from hundreds of artists, and (ii) recognize if the
identified style reappears in generated images. We then apply ArtSavant in an em
pirical study to quantify the prevalence of artistic style copying across 3 popu
lar text-to-image generative models, finding that under simple prompting, $20\\%
$ of 372 prolific artists studied appear to have their styles be at risk of co
pying by today's generative models. Our findings show that prior legal arguments
can be operationalized in quantitative ways, towards more nuanced examination o
f the issue of artistic style infringements.
****
Hyungi Lee,Seungyoo Lee,Juho Lee
Variational Bayesian Pseudo-Coreset
The success of deep learning requires large datasets and extensive training, whi
ch can create significant computational challenges. To address these challenges,
pseudo-coresets, small learnable datasets that mimic the entire data, have been
proposed. Bayesian Neural Networks, which offer predictive uncertainty and prob
abilistic interpretation for deep neural networks, also face issues with large-s
cale datasets due to their high-dimensional parameter space. Prior works on Baye
sian Pseudo-Coresets (BPC) attempt to reduce the computational load for computin
g weight posterior distribution by a small number of pseudo-coresets but suffer
from memory inefficiency during BPC training and sub-optimal results. To overcom
e these limitations, we propose Variational Bayesian Pseudo-Coreset (VBPC), a no
vel approach that utilizes variational inference to efficiently approximate the
posterior distribution, reducing memory usage and computational costs while impr
oving performance across benchmark datasets.
****
Jea Kwon,Sungjun Lim,Kyungwoo Song,C. Justin Lee
Brain-inspired L_p-Convolution benefits large kernels and aligns better with v
isual cortex
Convolutional Neural Networks (CNNs) have profoundly influenced the field of com
puter vision, drawing significant inspiration from the visual processing mechani
sms inherent in the brain. Despite sharing fundamental structural and representa
tional similarities with the biological visual system, differences in local conn

ectivity patterns within CNNs open up an interesting area to explore. In this wo rk, we explore whether integrating biologically observed receptive fields (RFs) can enhance model performance and foster alignment with brain representations. W e introduce a novel methodology, termed L_p-convolution, which employs the mul tivariate L_p-generalized normal distribution as an adaptable L_p-masks, to reconcile disparities between artificial and biological RFs. L_p-masks finds t he optimal RFs through task-dependent adaptation of conformation such as distort ion, scale, and rotation. This allows L_p-convolution to excel in tasks that r equire flexible RF shapes, including not only square-shaped regular RFs but also horizontal and vertical ones. Furthermore, we demonstrate that L_p-convolutio n with biological RFs significantly enhances the performance of large kernel CNN s possibly by introducing structured sparsity inspired by L_p-generalized norm al distribution in convolution. Lastly, we present that neural representations o f CNNs align more closely with the visual cortex when -convolution is close to b iological RFs.
****

Yilun Hao,Yang Zhang,Chuchu Fan
Planning Anything with Rigor: General-Purpose Zero-Shot Planning with LLM-based Formalized Programming
While large language models (LLMs) have recently demonstrated strong potential i n solving planning problems, there is a trade-off between flexibility and comple xity. LLMs, as zero-shot planners themselves, are still not capable of directly generating valid plans for complex planning problems such as multi-constraint or long-horizon tasks. On the other hand, many frameworks aiming to solve complex planning problems often rely on task-specific preparatory efforts, such as task-specific in-context examples and pre-defined critics/verifiers, which limits the ir cross-task generalization capability. In this paper, we tackle these challeng es by observing that the core of many planning problems lies in optimization pro blems: searching for the optimal solution (best plan) with goals subject to cons traints (preconditions and effects of decisions). With LLMs' commonsense, reason ing, and programming capabilities, this opens up the possibilities of a universa l LLM-based approach to planning problems. Inspired by this observation, we prop ose LLMFP, a general-purpose framework that leverages LLMs to capture key inform ation from planning problems and formally formulate and solve them as optimizati on problems from scratch, with no task-specific examples needed. We apply LLMFP to 9 planning problems, ranging from multi-constraint decision making to multi-s tep planning problems, and demonstrate that LLMFP achieves on average 83.7\% and 86.8\% optimal rate across 9 tasks for GPT-4o and Claude 3.5 Sonnet, significan tly outperforming the best baseline (direct planning with OpenAI o1-preview) wit h 37.6\% and 40.7\% improvements. We also validate components of LLMFP with abla tion experiments and analyzed the underlying success and failure reasons.
****

Antonio Sclocchi,Alessandro Favero,Noam Itzhak Levi,Matthieu Wyart
Probing the Latent Hierarchical Structure of Data via Diffusion Models
High-dimensional data must be highly structured to be learnable. Although the co mpositional and hierarchical nature of data is often put forward to explain lear nability, quantitative measurements establishing these properties are scarce. Li kewise, accessing the latent variables underlying such a data structure remains a challenge. In this work, we show that forward-backward experiments in diffusio n-based models, where data is noised and then denoised to generate new samples, are a promising tool to probe the latent structure of data. We predict in simple hierarchical models that, in this process, changes in data occur by correlated chunks, with a length scale that diverges at a noise level where a phase transit ion is known to take place. Remarkably, we confirm this prediction in both text and image datasets using state-of-the-art diffusion models. Our results show how latent variable changes manifest in the data and establish how to measure these effects in real data using diffusion models.
****

Sung Woong Cho,Hwijae Son
Physics-Informed Deep Inverse Operator Networks for Solving PDE Inverse Problems

Inverse problems involving partial differential equations (PDEs) can be seen as discovering a mapping from measurement data to unknown quantities, often framed within an operator learning approach. However, existing methods typically rely on large amounts of labeled training data, which is impractical for most real-world applications. Moreover, these supervised models may fail to capture the underlying physical principles accurately. To address these limitations, we propose a novel architecture called Physics-Informed Deep Inverse Operator Networks (PI-DIONs), which can learn the solution operator of PDE-based inverse problems without any labeled training data. We extend the stability estimates established in the inverse problem literature to the operator learning framework, thereby providing a robust theoretical foundation for our method. These estimates guarantee that the proposed model, trained on a finite sample and grid, generalizes effectively across the entire domain and function space. Extensive experiments are conducted to demonstrate that PI-DIONs can effectively and accurately learn the solution operators of the inverse problems without the need for labeled data.
****

Youssef Shehata,Benjamin Holzschuh,Nils Thuerey
Improved Sampling Of Diffusion Models In Fluid Dynamics With Tweedie's Formula
State-of-the-art Denoising Diffusion Probabilistic Models (DDPMs) rely on an expensive sampling process with a large Number of Function Evaluations (NFEs) to provide high-fidelity predictions. This computational bottleneck renders diffusion models less appealing as surrogates for the spatio-temporal prediction of physics-based problems with long rollout horizons. We propose Truncated Sampling Models, enabling single-step and few-step sampling with elevated fidelity by simple truncation of the diffusion process, reducing the gap between DDPMs and deterministic single-step approaches. We also introduce a novel approach, Iterative Refinement, to sample pre-trained DDPMs by reformulating the generative process as a refinement process with few sampling steps. Both proposed methods enable significant improvements in accuracy compared to DDPMs, DDIMs, and EDMs with NFEs $\leq$ 10 on a diverse set of experiments, including incompressible and compressible turbulent flow and airfoil flow uncertainty simulations. Our proposed methods provide stable predictions for long rollout horizons in time-dependent problems and are able to learn all modes of the data distribution in steady-state problems with high uncertainty.
****

Qianlan Yang,Yu-Xiong Wang
RTDiff: Reverse Trajectory Synthesis via Diffusion for Offline Reinforcement Learning
In offline reinforcement learning (RL), managing the distribution shift between the learned policy and the static offline dataset is a persistent challenge that can result in overestimated values and suboptimal policies. Traditional offline RL methods address this by introducing conservative biases that limit exploration to well-understood regions, but they often overly restrict the agent's generalization capabilities. Recent work has sought to generate trajectories using generative models to augment the offline dataset, yet these methods still struggle with overestimating synthesized data, especially when out-of-distribution samples are produced. To overcome this issue, we propose RTDiff, a novel diffusion-based data augmentation technique that synthesizes trajectories *in reverse*, moving from unknown to known states. Such reverse generation naturally mitigates the risk of overestimation by ensuring that the agent avoids planning through unknown states. Additionally, reverse trajectory synthesis allows us to generate longer, more informative trajectories that take full advantage of diffusion models' generative strengths while ensuring reliability. We further enhance RTDiff by introducing flexible trajectory length control and improving the efficiency of the generation process through noise management. Our empirical results show that RTDiff significantly improves the performance of several state-of-the-art offline RL algorithms across diverse environments, achieving consistent and superior results by effectively overcoming distribution shift.
****

Minjun Zhu,Yixuan Weng,Linyi Yang,Yue Zhang

Personality Alignment of Large Language Models
Aligning large language models (LLMs) typically aim to reflect general human values and behaviors, but they often fail to capture the unique characteristics and preferences of individual users. To address this gap, we introduce the concept of Personality Alignment. This approach tailors LLMs' responses and decisions to match the specific preferences of individual users or closely related groups. Inspired by psychometrics, we created the Personality Alignment with Personality Inventories (PAPI) dataset, which includes data from over 320,000 real subjects across multiple personality assessments - including both the Big Five Personality Factors and Dark Triad traits. This comprehensive dataset enables quantitative evaluation of LLMs' alignment capabilities across both positive and potentially problematic personality dimensions. Recognizing the challenges of personality alignments—such as limited personal data, diverse preferences, and scalability requirements—we developed an activation intervention optimization method. This method enhances LLMs' ability to efficiently align with individual behavioral preferences using minimal data and computational resources. Remarkably, our method, PAS, achieves superior performance while requiring only 1/5 of the optimization time compared to DPO, offering practical value for personality alignment. Our work paves the way for future AI systems to make decisions and reason in truly personality ways, enhancing the relevance and meaning of AI interactions for each user and advancing human-centered artificial intelligence. The dataset and code are released at https://github.com/zhu-minjun/PAlign.
****
Kyle Vedder,Neehar Peri,Ishan Khatri,Siyi Li,Eric Eaton,Mehmet Kemal Kocamaz,Yue Wang,Zhiding Yu,Deva Ramanan,Joachim Pehserl
Neural Eulerian Scene Flow Fields
We reframe scene flow as the task of estimating a continuous space-time ordinary differential equation (ODE) that describes motion for an entire observation sequence, represented with a neural prior. Our method, EulerFlow, optimizes this neural prior estimate against several multi-observation reconstruction objectives, enabling high quality scene flow estimation via self-supervision on real-world data. EulerFlow works out-of-the-box without tuning across multiple domains, including large-scale autonomous driving scenes and dynamic tabletop settings. Remarkably, EulerFlow produces high quality flow estimates on small, fast moving objects like birds and tennis balls, and exhibits emergent 3D point tracking behavior by solving its estimated ODE over long-time horizons. On the Argoverse 2 2024 Scene Flow Challenge, EulerFlow outperforms all prior art, surpassing the next-best unsupervised method by more than 2.5 times, and even exceeding the next-best supervised method by over 10%. See https://vedder.io/eulerflow for interactive visuals.
****
Mintong Kang,Chejian Xu,Bo Li
AdvWave: Stealthy Adversarial Jailbreak Attack against Large Audio-Language Models
Recent advancements in large audio-language models (LALMs) have enabled speech-based user interactions, significantly enhancing user experience and accelerating the deployment of LALMs in real-world applications. However, ensuring the safety of LALMs is crucial to prevent risky outputs that may raise societal concerns or violate AI regulations. Despite the importance of this issue, research on jailbreaking LALMs remains limited due to their recent emergence and the additional technical challenges they present compared to attacks on DNN-based audio models. Specifically, the audio encoders in LALMs, which involve discretization operations, often lead to gradient shattering, hindering the effectiveness of attacks relying on gradient-based optimizations. The behavioral variability of LALMs further complicates the identification of effective (adversarial) optimization targets. Moreover, enforcing stealthiness constraints on adversarial audio waveforms introduces a reduced, non-convex feasible solution space, further intensifying the challenges of the optimization process. To overcome these challenges, we develop AdvWave, the first jailbreak framework against LALMs. We propose a dual-phase optimization method that addresses gradient shattering, enabling effective en

d-to-end gradient-based optimization. Additionally, we develop an adaptive adver
sarial target search algorithm that dynamically adjusts the adversarial optimiza
tion target based on the response patterns of LALMs for specific queries. To ens
ure that adversarial audio remains perceptually natural to human listeners, we d
esign a classifier-guided optimization approach that generates adversarial noise
 resembling common urban sounds. Extensive evaluations on multiple advanced LALM
s demonstrate that AdvWave outperforms baseline methods, achieving a 40\% higher
 average jailbreak attack success rate. Both audio stealthiness metrics and huma
n evaluations confirm that adversarial audio generated by AdvWave is indistingui
shable from natural sounds. We believe AdvWave will inspire future research aimi
ng to enhance the safety alignment of LALMs, supporting their responsible deploy
ment in real-world scenarios.
****

Hyun Ryu,Gyeongman Kim,Hyemin S. Lee,Eunho Yang
Divide and Translate: Compositional First-Order Logic Translation and Verificati
on for Complex Logical Reasoning
Complex logical reasoning tasks require a long sequence of reasoning, which a la
rge language model (LLM) with chain-of-thought prompting still falls short. To a
lleviate this issue, neurosymbolic approaches incorporate a symbolic solver. Spe
cifically, an LLM only translates a natural language problem into a satisfiabili
ty (SAT) problem that consists of first-order logic formulas, and a sound symbol
ic solver returns a mathematically correct solution. However, we discover that L
LMs have difficulties to capture complex logical semantics hidden in the natural
 language during translation. To resolve this limitation, we propose a Compositi
onal First-Order Logic Translation. An LLM first parses a natural language sente
nce into newly defined logical dependency structures that consist of an atomic s
ubsentence and its dependents, then sequentially translate the parsed subsentenc
es. Since multiple logical dependency structures and sequential translations are
 possible for a single sentence, we also introduce two Verification algorithms t
o ensure more reliable results. We utilize an SAT solver to rigorously compare s
emantics of generated first-order logic formulas and select the most probable on
e. We evaluate the proposed method, dubbed CLOVER, on seven logical reasoning be
nchmarks and show that it outperforms the previous neurosymbolic approaches and
achieves new state-of-the-art results.
****

Peiran Xu,Yadong MU
Weakly-Supervised Affordance Grounding Guided by Part-Level Semantic Priors
In this work, we focus on the task of weakly supervised affordance grounding, wh
ere a model is trained to identify affordance regions on objects using human-obj
ect interaction images and egocentric object images without dense labels.
Previous works are mostly built upon class activation maps, which are effective
for semantic segmentation but may not be suitable for locating actions and funct
ions. Leveraging recent advanced foundation models, we develop a supervised trai
ning pipeline based on pseudo labels. The pseudo labels are generated from an of
f-the-shelf part segmentation model, guided by a mapping from affordance to part
 names.
Furthermore, we introduce three key enhancements to the baseline model: a label
refining stage, a fine-grained feature alignment process, and a lightweight reas
oning module. These techniques harness the semantic knowledge of static objects
embedded in off-the-shelf foundation models to improve affordance learning, effe
ctively bridging the gap between objects and actions.
Extensive experiments demonstrate that the performance of the proposed model has
 achieved a breakthrough improvement over existing methods.
****

Zitong Yang,Neil Band,Shuangping Li,Emmanuel Candes,Tatsunori Hashimoto
Synthetic continued pretraining
Pretraining on large-scale, unstructured internet text enables language models t
o acquire a significant amount of world knowledge.
However, this knowledge acquisition is data-inefficient---to learn a fact, model
s must be trained on hundreds to thousands of diverse representations of it.

This poses a challenge when adapting a pretrained model to a small corpus of domain-specific documents, where each fact may appear rarely or only once.
We propose to bridge this gap with synthetic continued pretraining: using the small domain-specific corpus to synthesize a large corpus more amenable to learning, and then performing continued pretraining on the synthesized corpus.
We instantiate this proposal with EntiGraph, a synthetic data augmentation algorithm that extracts salient entities from the source corpus and then generates diverse text by drawing connections between those entities.
Synthetic continued pretraining with EntiGraph enables a language model to answer questions and follow generic instructions related to the source documents without access to them.
If the source documents are instead available at inference time, we show that the knowledge acquired through our approach compounds with retrieval-augmented generation.
To better understand these results, we build a simple mathematical model of EntiGraph, and show how synthetic data augmentation can "rearrange" knowledge to enable more data-efficient learning.
****
Jacob Morrison,Clara Na,Jared Fernandez,Tim Dettmers,Emma Strubell,Jesse Dodge
Holistically Evaluating the Environmental Impact of Creating Language Models
As the performance of artificial intelligence systems has dramatically increased, so too has the environmental impact of creating these systems. While many model developers release estimates of the power consumption and carbon emissions from the final training runs for their latest models, there is comparatively little transparency into the impact of model development, hardware manufacturing, and total water usage throughout. In this work, we estimate the real-world environmental impact of developing a series of language models, ranging from 20 million to 13 billion active parameters, trained on up to 5.6 trillion tokens each. When accounting for hardware manufacturing, model development, and our final training runs, we find that our series of models released **493 metric tons** of carbon emissions, equivalent to powering about 98 homes in the United States for one year, and consumed **2.769 million liters of water**, equivalent to about 24.5 years of water usage by a person in the United States, even though our data center is extremely water-efficient. We measure and report the environmental impact of our model development; to the best of our knowledge we are the first to do so for LLMs, and we find that model development, the impact of which is generally not disclosed by most model developers, amounted to **~50%** of that of training. By looking at detailed time series data for power consumption, we also find that power usage throughout training is not consistent, fluctuating between ~15% and ~85% of our hardware's maximum power draw, with negative implications for grid-scale planning as demand continues to grow. We close with a discussion on the continued difficulty of estimating the environmental impact of AI systems, and key takeaways for model developers and the public at large.
****
Youssef Allouah,Rachid Guerraoui,Nirupam Gupta,Ahmed Jellouli,Geovani Rizk,John Stephan
Adaptive Gradient Clipping for Robust Federated Learning
Robust federated learning aims to maintain reliable performance despite the presence of adversarial or misbehaving workers. While state-of-the-art (SOTA) robust distributed gradient descent (Robust-DGD) methods were proven theoretically optimal, their empirical success has often relied on pre-aggregation gradient clipping.
However, existing static clipping strategies yield inconsistent results: enhancing robustness against some attacks while being ineffective or even detrimental against others.
To address this limitation, we propose a principled adaptive clipping strategy, Adaptive Robust Clipping (ARC), which dynamically adjusts clipping thresholds based on the input gradients. We prove that ARC not only preserves the theoretical robustness guarantees of SOTA Robust-DGD methods but also provably improves asymptotic convergence when the model is well-initialized. Extensive experiments on

benchmark image classification tasks confirm these theoretical insights, demonstrating that ARC significantly enhances robustness, particularly in highly heterogeneous and adversarial settings.
****
Simon Heilig,Alessio Gravina,Alessandro Trenta,Claudio Gallicchio,Davide Bacciu
Port-Hamiltonian Architectural Bias for Long-Range Propagation in Deep Graph Networks
The dynamics of information diffusion within graphs is a critical open issue that heavily influences graph representation learning, especially when considering long-range propagation. This calls for principled approaches that control and regulate the degree of propagation and dissipation of information throughout the neural flow. Motivated by this, we introduce port-Hamiltonian Deep Graph Networks, a novel framework that models neural information flow in graphs by building on the laws of conservation of Hamiltonian dynamical systems. We reconcile under a single theoretical and practical framework both non-dissipative long-range propagation and non-conservative behaviors, introducing tools from mechanical systems to gauge the equilibrium between the two components. Our approach can be applied to general message-passing architectures, and it provides theoretical guarantees on information conservation in time. Empirical results prove the effectiveness of our port-Hamiltonian scheme in pushing simple graph convolutional architectures to state-of-the-art performance in long-range benchmarks.
****
Yecheng Wu,Zhuoyang Zhang,Junyu Chen,Haotian Tang,Dacheng Li,Yunhao Fang,Ligeng Zhu,Enze Xie,Hongxu Yin,Li Yi,Song Han,Yao Lu
VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation
VILA-U is a Unified foundation model that integrates Video, Image, Language understanding and generation. Traditional visual language models (VLMs) use separate modules for understanding and generating visual content, which can lead to misalignment and increased complexity. In contrast, VILA-U employs a single autoregressive next-token prediction framework for both tasks, eliminating the need for additional components like diffusion models. This approach not only simplifies the model but also achieves near state-of-the-art performance in visual language understanding and generation. The success of VILA-U is attributed to two main factors: the unified vision tower that aligns discrete visual tokens with textual inputs during pretraining, which enhances visual perception, and autoregressive image generation can achieve similar quality as diffusion models with high-quality dataset. This allows VILA-U to perform comparably to more complex models using a fully token-based autoregressive framework.
****
Siyu Luan,Zhenyi Wang,Li Shen,Zonghua Gu,Chao Wu,Dacheng Tao
Dynamic Neural Fortresses: An Adaptive Shield for Model Extraction Defense
Model extraction aims to acquire a pre-trained black-box model concealed behind a black-box API.
Existing defense strategies against model extraction primarily concentrate on preventing the unauthorized extraction of API functionality. However, two significant challenges still need to be solved: (i) Neural network architecture of the API constitutes a form of intellectual property that also requires protection; (ii) The current practice of allocating the same network architecture to both attack and benign queries results in substantial resource wastage. To address these challenges, we propose a novel \textit{Dynamic Neural Fortresses} (DNF) defense method, employing a dynamic Early-Exit neural network, deviating from the conventional fixed architecture. Firstly, we facilitate the random exit of attack queries from the network at earlier layers. This strategic exit point selection significantly reduces the computational cost for attack queries. Furthermore, the random exit of attack queries from earlier layers introduces increased uncertainty for attackers attempting to discern the exact architecture, thereby enhancing architectural protection. On the contrary, we aim to facilitate benign queries to exit at later layers, preserving model utility, as these layers typically yield meaningful information.

Extensive experiments on defending against various model extraction scenarios an
d datasets demonstrate the effectiveness of DNF, achieving a notable 2$\times$ i
mprovement in efficiency and an impressive reduction of up to 12\% in clone mode
l accuracy compared to SOTA defense methods. Additionally, DNF provides strong p
rotection against neural architecture theft, effectively safeguarding network ar
chitecture from being stolen.
****

Zaid Khan,Elias Stengel-Eskin,Jaemin Cho,Mohit Bansal
DataEnvGym: Data Generation Agents in Teacher Environments with Student Feedback
The process of creating training data to teach models is currently driven by hum
ans, who manually analyze model weaknesses and plan how to create data that impr
oves a student model. Recent approaches using large language models (LLMs) as an
notators reduce human annotation effort, but still require humans to interpret f
eedback from evaluations and control the LLM to produce data the student needs.
Automating this labor-intensive process by creating autonomous data generation a
gents – or teachers – is desirable, but requires environments that can simulate
the feedback-driven, iterative, closed loop of data creation. To enable rapid an
d scalable testing for such agents and their modules, we introduce DataEnvGym, a
 testbed of teacher environments for data generation agents. DataEnvGym frames d
ata generation as a sequential decision-making task, involving an agent consisti
ng of a data generation policy (which generates a plan for creating training dat
a) and a data generation engine (which transforms the plan into data), inside an
 environment that provides feedback from a student. The agent’s end goal is to i
mprove student model performance. Students are iteratively trained and evaluated
 on generated data, with their feedback (in the form of errors or weak skills) b
eing reported to the agent after each iteration. As a general-purpose testbed, D
ataEnvGym includes multiple instantiations of teacher environments across three
levels of structure in the state representation and action space, with varying l
evels of scaffolding support. More structured environments are based on automati
cally-inferred skills and offer a higher degree of interpretability and control
over the curriculum. We support developing and testing data generation agents in
 four diverse tasks covering text, images, and actions (mathematics, programming
, visual question answering, and tool-use) and test multiple student and teacher
 models. We find that example agents in our teaching environments can iterativel
y improve students across diverse tasks and settings. Moreover, we show that env
ironments can teach different skill levels and can be used to test variants of k
ey modules, pointing to directions of future work in improving data generation a
gents, engines, and feedback mechanisms. Project page: https://DataEnvGym.github
.io.
****