Wanru Zhao,Royson Lee,Yihong Chen,Xinchi Qiu,Yan Gao,Hongxiang Fan,Nicholas Donald Lane
Breaking Physical and Linguistic Borders: Multilingual Federated Prompt Tuning for Low-Resource Languages
Pretrained large language models (LLMs) have emerged as a cornerstone in modern natural language processing, with their utility expanding to various applications and languages. However, the fine-tuning of multilingual LLMs, particularly for low-resource languages, is fraught with challenges steming from data-sharing restrictions (the physical border) and from the inherent linguistic differences (the linguistic border). These barriers hinder users of various languages, especially those in low-resource regions, from fully benefiting from the advantages of LLMs.

To overcome these challenges, we propose the Federated Prompt Tuning Paradigm for Multilingual Scenarios, which leverages parameter-efficient fine-tuning in a manner that preserves user privacy. We have designed a comprehensive set of experiments and introduced the concept of "language distance" to highlight the several strengths of this paradigm. Even under computational constraints, our method not only bolsters data efficiency but also facilitates mutual enhancements across languages, particularly benefiting low-resource ones. Compared to traditional local crosslingual transfer tuning methods, our approach achieves a 6.9\% higher accuracy, reduces the training parameters by over 99\%, and demonstrates stronger cross-lingual generalization. Such findings underscore the potential of our approach to promote social equality, ensure user privacy, and champion linguistic diversity.
****
Takuya Ito,Soham Dan,Mattia Rigotti,James Kozloski,Murray Campbell
On the generalization capacity of neural networks during generic multimodal reasoning
The advent of the Transformer has led to the development of large language models (LLM), which appear to demonstrate human-like capabilities. To assess the generality of this class of models and a variety of other base neural network architectures to multimodal domains, we evaluated and compared their capacity for multimodal generalization. We introduce a multimodal question-answer benchmark to evaluate three specific types of out-of-distribution (OOD) generalization performance: distractor generalization (generalization in the presence of distractors), systematic compositional generalization (generalization to new task permutations), and productive compositional generalization (generalization to more complex tasks with deeper dependencies). While we found that most architectures faired poorly on most forms of generalization (e.g., RNNs and standard Transformers), models that leveraged cross-attention mechanisms between input domains, such as the Perceiver, fared better. Our positive results demonstrate that for multimodal distractor and systematic generalization, cross-attention is an important mechanism to integrate multiple sources of information. On the other hand, all architectures failed in productive generalization, suggesting fundamental limitations of existing architectures for specific types of multimodal OOD generalization. These results demonstrate the strengths and limitations of specific architectural components underlying modern neural models for multimodal reasoning. Finally, we provide *Generic COG* (gCOG), a configurable benchmark with several multimodal generalization splits, for future studies to explore.
****
Christian Fabian,Kai Cui,Heinz Koeppl
Learning Mean Field Games on Sparse Graphs: A Hybrid Graphex Approach
Learning the behavior of large agent populations is an important task for numerous research areas. Although the field of multi-agent reinforcement learning (MARL) has made significant progress towards solving these systems, solutions for many agents often remain computationally infeasible and lack theoretical guarantees. Mean Field Games (MFGs) address both of these issues and can be extended to Graphon MFGs (GMFGs) to include network structures between agents. Despite their merits, the real world applicability of GMFGs is limited by the fact that grapho

ns only capture dense graphs. Since most empirically observed networks show some degree of sparsity, such as power law graphs, the GMFG framework is insufficient for capturing these network topologies. Thus, we introduce the novel concept of Graphex MFGs (GXMFGs) which builds on the graph theoretical concept of graphexes. Graphexes are the limiting objects to sparse graph sequences that also have other desirable features such as the small world property. Learning equilibria in these games is challenging due to the rich and sparse structure of the underlying graphs. To tackle these challenges, we design a new learning algorithm tailored to the GXMFG setup. This hybrid graphex learning approach leverages that the system mainly consists of a highly connected core and a sparse periphery. After defining the system and providing a theoretical analysis, we state our learning approach and demonstrate its learning capabilities on both synthetic graphs and real-world networks. This comparison shows that our GXMFG learning algorithm successfully extends MFGs to a highly relevant class of hard, realistic learning problems that are not accurately addressed by current MARL and MFG methods.
****

Siyuan Guo,Jonas Bernhard Wildberger,Bernhard Schölkopf
Out-of-Variable Generalisation for Discriminative Models
The ability of an agent to do well in new environments is a critical aspect of intelligence. In machine learning, this ability is known as $\textit{strong}$ or $\textit{out-of-distribution}$ generalization. However, merely considering differences in distributions is inadequate for fully capturing differences between learning environments. In the present paper, we investigate $\textit{out-of-variable}$ generalization, which pertains to an agent's generalization capabilities concerning environments with variables that were never jointly observed before. This skill closely reflects the process of animate learning: we, too, explore Nature by probing, observing, and measuring proper $\textit{subsets}$ of variables at any given time. Mathematically, $\textit{oov}$ generalization requires the efficient re-use of past marginal information, i.e., information over subsets of previously observed variables. We study this problem, focusing on prediction tasks across environments that contain overlapping, yet distinct, sets of causes. We show that after fitting a classifier, the residual distribution in one environment reveals the partial derivative of the true generating function with respect to the unobserved causal parent in that environment. We leverage this information and propose a method that exhibits non-trivial out-of-variable generalization performance when facing an overlapping, yet distinct, set of causal predictors. Code: https://github.com/syguo96/Out-of-Variable-Generalization
****

Xinlu Zhang,Shiyang Li,Xianjun Yang,Chenxin Tian,Yao Qin,Linda Ruth Petzold
Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting
Large language models (LLMs) demonstrate remarkable medical expertise, but data privacy concerns impede their direct use in healthcare environments. Although offering improved data privacy protection, domain-specific small language models (SLMs) often underperform LLMs, emphasizing the need for methods that reduce this performance gap while alleviating privacy concerns. In this paper, we present a simple yet effective method that harnesses LLMs' medical proficiency to boost SLM performance in medical tasks under $privacy-restricted$ scenarios. Specifically, we mitigate patient privacy issues by extracting keywords from medical data and prompting the LLM to generate a medical knowledge-intensive context by simulating clinicians' thought processes. This context serves as additional input for SLMs, augmenting their decision-making capabilities. Our method significantly enhances performance in both few-shot and full training settings across three medical knowledge-intensive tasks, achieving up to a 22.57% increase in absolute accuracy compared to SLM fine-tuning without context, and sets new state-of-the-art results in two medical tasks within privacy-restricted scenarios. Further out-of-domain testing and experiments in two general domain datasets showcase its generalizability and broad applicability.
****

Senmao Li,Joost van de Weijer,taihang Hu,Fahad Khan,Qibin Hou,Yaxing Wang,jian Yang

Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Ima
ge Diffusion Models
The success of recent text-to-image diffusion models is largely due to their cap
acity to be guided by a complex text prompt, which enables users to precisely de
scribe the desired content. However, these models struggle to effectively suppre
ss the generation of undesired content, which is explicitly requested to be omit
ted from the generated image in the prompt. In this paper, we analyze how to man
ipulate the text embeddings and remove unwanted content from them. We introduce
two contributions, which we refer to as soft-weighted regularization and inferen
ce-time text embedding optimization. The first regularizes the text embedding ma
trix and effectively suppresses the undesired content. The second method aims to
 further suppress the unwanted content generation of the prompt, and encourages
the generation of desired content. We evaluate our method quantitatively and qua
litatively on extensive experiments, validating its effectiveness. Furthermore,
our method is generalizability to both the pixel-space diffusion models (i.e. De
epFloyd-IF) and the latent-space diffusion models (i.e. Stable Diffusion).
****
Shuai Fu,Xiequn Wang,Qiushi Huang,Yu Zhang
Nemesis: Normalizing the Soft-prompt Vectors of Vision-Language Models
With the prevalence of large-scale pretrained vision-language models (VLMs), suc
h as CLIP, soft-prompt tuning has become a popular method for adapting these mod
els to various downstream tasks. However, few works delve into the inherent prop
erties of learnable soft-prompt vectors, specifically the impact of their norms
to the performance of VLMs. This motivates us to pose an unexplored research que
stion: ``Do we need to normalize the soft prompts in VLMs?'' To fill this resear
ch gap, we first uncover a phenomenon, called the $\textbf{Low-Norm Effect}$ by
performing extensive corruption experiments, suggesting that reducing the norms
of certain learned prompts occasionally enhances the performance of VLMs, while
increasing them often degrades it. To harness this effect, we propose a novel me
thod named $\textbf{N}$ormalizing th$\textbf{e}$ soft-pro$\textbf{m}$pt v$\textb
f{e}$ctors of vi$\textbf{si}$on-language model$\textbf{s}$ ($\textbf{Nemesis}$)
to normalize soft-prompt vectors in VLMs. To the best of our knowledge, our work
 is the first to systematically investigate the role of norms of soft-prompt vec
tor in VLMs, offering valuable insights for future research in soft-prompt tunin
g.
****
Lorenz Vaitl,Ludwig Winkler,Lorenz Richter,Pan Kessel
Fast and unified path gradient estimators for normalizing flows
Recent work shows that path gradient estimators for normalizing flows have lower
 variance compared to standard estimators, resulting in improved training. Howev
er, they are often prohibitively more expensive from a computational point of vi
ew and cannot be applied to maximum likelihood training in a scalable manner, wh
ich severely hinders their widespread adoption. In this work, we overcome these
crucial limitations. Specifically, we propose a fast path gradient estimator whi
ch works for all normalizing flow architectures of practical relevance for sampl
ing from an unnormalized target distribution. We then show that this estimator c
an also be applied to maximum likelihood training and empirically establish its
superior performance for several natural sciences applications.
****
Young-Jae Park,Minseok Seo,Doyi Kim,Hyeri Kim,Sanghoon Choi,Beomkyu Choi,Jeongwo
n Ryu,Sohee Son,Hae-Gon Jeon,Yeji Choi
Long-Term Typhoon Trajectory Prediction: A Physics-Conditioned Approach Without
Reanalysis Data
In the face of escalating climate changes, typhoon intensities and their ensuing
 damage have surged. Accurate trajectory prediction is crucial for effective dam
age control. Traditional physics-based models, while comprehensive, are computat
ionally intensive and rely heavily on the expertise of forecasters. Contemporary
 data-driven methods often rely on reanalysis data, which can be considered to b
e the closest to the true representation of weather conditions. However, reanaly
sis data is not produced in real-time and requires time for adjustment since pre

diction models are calibrated with observational data. This reanalysis data, suc
h as ERA5, falls short in challenging real-world situations. Optimal preparednes
s necessitates predictions at least 72 hours in advance, beyond the capabilities
 of standard physics models. In response to these constraints, we present an app
roach that harnesses real-time Unified Model (UM) data, sidestepping the limitat
ions of reanalysis data. Our model provides predictions at 6-hour intervals for
up to 72 hours in advance and outperforms both state-of-the-art data-driven meth
ods and numerical weather prediction models. In line with our efforts to mitigat
e adversities inflicted by \rthree{typhoons}, we release our preprocessed \texti
t{PHYSICS TRACK} dataset, which includes ERA5 reanalysis data, typhoon best-trac
k, and UM forecast data.
****
Ziyang Yu,Wenbing Huang,Yang Liu
Rigid Protein-Protein Docking via Equivariant Elliptic-Paraboloid Interface Pred
iction
The study of rigid protein-protein docking plays an essential role in a variety
of tasks such as drug design and protein engineering. Recently, several learning
-based methods have been proposed for the task, exhibiting much faster docking s
peed than those computational methods. In this paper, we propose a novel learnin
g-based method called ElliDock, which predicts an elliptic paraboloid to represe
nt the protein-protein docking interface. To be specific, our model estimates el
liptic paraboloid interfaces for the two input proteins respectively, and obtain
s the roto-translation transformation for docking by making two interfaces coinc
ide. By its design, ElliDock is independently equivariant with respect to arbitr
ary rotations/translations of the proteins, which is an indispensable property t
o ensure the generalization of the docking process. Experimental evaluations sho
w that ElliDock achieves the fastest inference time among all compared methods,
and outperforms state-of-the-art learning-based methods, like DiffDock-PP and Al
phafold-Multimer, for particularly antibody-antigen docking.
****
Yiding Jiang,Christina Baek,J Zico Kolter
On the Joint Interaction of Models, Data, and Features
Learning features from data is one of the defining characteristics of deep learn
ing,
but our theoretical understanding of the role features play in deep learning is
still
rudimentary. To address this gap, we introduce a new tool, the interaction tenso
r,
for empirically analyzing the interaction between data and model through feature
s.
With the interaction tensor, we make several key observations about how features
are distributed in data and how models with different random seeds learn differe
nt
features. Based on these observations, we propose a conceptual framework for fea
-
ture learning. Under this framework, the expected accuracy for a single hypothes
is
and agreement for a pair of hypotheses can both be derived in closed-form. We
demonstrate that the proposed framework can explain empirically observed phenome
na, including the recently discovered Generalization Disagreement Equality
(GDE) that allows for estimating the generalization error with only unlabeled da
ta.
Further, our theory also provides explicit construction of natural data distribu
tions
that break the GDE. Thus, we believe this work provides valuable new insight int
o
our understanding of feature learning.
****
Panagiotis Theodoropoulos,Guan-Horng Liu,Tianrong Chen,Augustinos D Saravanos,Ev
angelos Theodorou

A ROBUST DIFFERENTIAL NEURAL ODE OPTIMIZER
Neural networks and neural ODEs tend to be vulnerable to adversarial attacks, re
ndering robust optimizers critical to curb the success of such attacks. In this
regard, the key insight of this work is to interpret Neural ODE optimization as
a min-max optimal control problem. More particularly, we present Game Theoretic
Second-Order Neural Optimizer (GTSONO), a robust game theoretic optimizer based
on the principles of min-max Differential Dynamic Programming.
The proposed method exhibits significant computational benefits due to efficient
 matrix decompositions and provides convergence guarantees to local saddle point
s.
Empirically, the robustness of the proposed optimizer is demonstrated through gr
eater robust accuracy compared to benchmark optimizers when trained on clean im
ages. Additionally, its ability to provide a performance increase when adapted t
o an already existing adversarial defense technique is also illustrated.
Finally, the superiority of the proposed update law over its gradient based coun
terpart highlights the potential benefits of incorporating robust optimal contro
l paradigms into adversarial training methods.
****
Tao Yu,Toni J.B. Liu,Albert Tseng,Christopher De Sa
Shadow Cones: A Generalized Framework for Partial Order Embeddings
Hyperbolic space has proven to be well-suited for capturing hierarchical relatio
ns in data, such as trees and directed acyclic graphs. Prior work introduced the
 concept of entailment cones, which uses partial orders defined by nested cones
in the Poincar\'e ball to model hierarchies. Here, we introduce the ``shadow con
es" framework, a physics-inspired entailment cone construction. Specifically, we
 model partial orders as subset relations between shadows formed by a light sour
ce and opaque objects in hyperbolic space. The shadow cones framework generalize
s entailment cones to a broad class of formulations and hyperbolic space models
beyond the Poincar\'e ball. This results in clear advantages over existing const
ructions: for example, shadow cones possess better optimization properties over
constructions limited to the Poincar\'e ball. Our experiments on datasets of var
ious sizes and hierarchical structures show that shadow cones consistently and s
ignificantly outperform existing entailment cone constructions. These results in
dicate that shadow cones are an effective way to model partial orders in hyperbo
lic space, offering physically intuitive and novel insights about the nature of
such structures.
****
Jiahai Feng,Jacob Steinhardt
How do Language Models Bind Entities in Context?
Language models (LMs) can recall facts mentioned in context, as shown by their p
erformance on reading comprehension tasks. When the context describes facts abou
t more than one entity, the LM has to correctly bind attributes to their corresp
onding entity. We show, via causal experiments, that LMs' internal activations r
epresent binding information by exhibiting appropriate binding ID vectors at the
 entity and attribute positions. We further show that binding ID vectors form a
subspace and often transfer across tasks. Our results demonstrate that LMs learn
 interpretable strategies for representing symbolic knowledge in context, and th
at studying context activations is a fruitful direction for understanding LM cog
nition.
****
Muthu Chidambaram,Rong Ge
On the Limitations of Temperature Scaling for Distributions with Overlaps
Despite the impressive generalization capabilities of deep neural networks, they
 have been repeatedly shown to be overconfident when they are wrong. Fixing this
 issue is known as model calibration, and has consequently received much attenti
on in the form of modified training schemes and post-training calibration proced
ures such as temperature scaling. While temperature scaling is frequently used b
ecause of its simplicity, it is often outperformed by modified training schemes.
 In this work, we identify a specific bottleneck for the performance of temperat
ure scaling. We show that for empirical risk minimizers for a general set of di

stributions in which the supports of classes have overlaps, the performance of t
emperature scaling degrades with the amount of overlap between classes, and asym
ptotically becomes no better than random when there are a large number of classe
s. On the other hand, we prove that optimizing a modified form of the empirical
risk induced by the Mixup data augmentation technique can in fact lead to reason
ably good calibration performance, showing that training-time calibration may be
 necessary in some situations. We also verify that our theoretical results refle
ct practice by showing that Mixup significantly outperforms empirical risk minim
ization (with respect to multiple calibration metrics) on image classification b
enchmarks with class overlaps introduced in the form of label noise.
****
Xiaohu Huang,Hao Zhou,Kun Yao,Kai Han
FROSTER: Frozen CLIP is A Strong Teacher for Open-Vocabulary Action Recognition
In this paper, we introduce \textbf{FROSTER}, an effective framework for open-vo
cabulary action recognition. The CLIP model has achieved remarkable success in a
 range of image-based tasks, benefiting from its strong generalization capabilit
y stemming from pretraining on massive image-text pairs. However, applying CLIP d
irectly to the open-vocabulary action recognition task is challenging due to the
 absence of temporal information in CLIP's pretraining. Further, fine-tuning CLI
P on action recognition datasets may lead to overfitting and hinder its generali
zability, resulting in unsatisfactory results when dealing with unseen actions.
To address these issues, FROSTER employs a residual feature distillation approac
h to ensure that CLIP retains its generalization capability while effectively ad
apting to the action recognition task. Specifically, the residual feature distil
lation treats the frozen CLIP model as a teacher to maintain the generalizabilit
y exhibited by the original CLIP and supervises the feature learning for the ext
raction of video-specific features to bridge the gap between images and videos.
Meanwhile, it uses a residual sub-network for feature distillation to reach a ba
lance between the two distinct objectives of learning generalizable and video-sp
ecific features.
We extensively evaluate FROSTER on open-vocabulary action recognition benchmarks
 under both base-to-novel and cross-dataset settings. FROSTER consistently achi
eves state-of-the-art performance on all datasets across the board.
Project page: \url{https://visual-ai.github.io/froster}.
****
Weijia Shi,Anirudh Ajith,Mengzhou Xia,Yangsibo Huang,Daogao Liu,Terra Blevins,Da
nqi Chen,Luke Zettlemoyer
Detecting Pretraining Data from Large Language Models
Although large language models (LLMs) are widely deployed, the data used to trai
n them is rarely disclosed. Given the incredible scale of this data, up to trill
ions of tokens, it is all but certain that it includes potentially problematic t
ext such as copyrighted materials, personally identifiable information, and test
 data for widely reported reference benchmarks. However, we currently have no wa
y to know which data of these types is included or in what proportions. In this
paper, we study the pretraining data detection problem: given a piece of text an
d black-box access to an LLM without knowing the pretraining data, can we determ
ine if the model was trained on the provided text? To facilitate this study, we
introduce a dynamic benchmark WIKIMIA that uses data created before and after mo
del training to support gold truth detection. We also introduce a new detection
method MIN-K PROB based on a simple hypothesis: an unseen example is likely to c
ontain a few outlier words with low probabilities under the LLM, while a seen ex
ample is less likely to have words with such low probabilities. MIN-K PROB can b
e applied without any knowledge about the pretrainig corpus or any additional tr
aining, departing from previous detection methods that require training a refere
nce model on data that is similar to the pretraining data. Moreover, our experim
ents demonstrate that MIN-K PROB achieves a 7.4% improvement on WIKIMIA over the
se previous methods. We apply MIN-K PROB to two real-world scenarios, copyrighte
d book detection and contaminated downstream example detection, and find that it
 to be a consistently effective solution.
****

Yucheng Yang,Tianyi Zhou,Qiang He,Lei Han,Mykola Pechenizkiy,Meng Fang
Task Adaptation from Skills: Information Geometry, Disentanglement, and New Obje
ctives for Unsupervised Reinforcement Learning
Unsupervised reinforcement learning (URL) aims to learn general skills for unsee
n downstream tasks. Mutual Information Skill Learning (MISL) addresses URL by ma
ximizing the mutual information between states and skills but lacks sufficient t
heoretical analysis, e.g., how well its learned skills can initialize a downstre
am task's policy. Our new theoretical analysis shows that the diversity and sepa
ratability of learned skills are fundamentally critical to downstream task adapt
ation but MISL does not necessarily guarantee them. To improve MISL, we propose
a novel disentanglement metric LSEPIN and build an information-geometric connect
ion between LSEPIN and downstream task adaptation cost. For better geometric pro
perties, we investigate a new strategy that replaces the KL divergence in inform
ation geometry with Wasserstein distance. We extend the geometric analysis to it
, which leads to a novel skill-learning objective WSEP. It is theoretically just
ified to be helpful to task adaptation and it is capable of discovering more ini
tial policies for downstream tasks than MISL. We further propose a Wasserstein d
istance-based algorithm PWSEP can theoretically discover all potentially optimal
 initial policies.
****
Adriano Cardace,Pierluigi Zama Ramirez,Francesco Ballerini,Allan Zhou,Samuele Sa
lti,Luigi di Stefano
Neural Processing of Tri-Plane Hybrid Neural Fields
Driven by the appealing properties of neural fields for storing and communicatin
g 3D data, the problem of directly processing them to address tasks such as clas
sification and part segmentation has emerged and has been investigated in recent
 works.
Early approaches employ neural fields parameterized by shared networks trained o
n the whole dataset, achieving good task performance but sacrificing reconstruct
ion quality.
To improve the latter, later methods focus on individual neural fields parameter
ized as large Multi-Layer Perceptrons (MLPs), which are, however, challenging to
 process due to the high dimensionality of the weight space, intrinsic weight sp
ace symmetries, and sensitivity to random initialization. Hence, results turn ou
t significantly inferior to those achieved by processing explicit representation
s, e.g., point clouds or meshes.
In the meantime, hybrid representations, in particular based on tri-planes, have
 emerged as a more effective and efficient alternative to realize neural fields,
 but their direct processing has not been investigated yet.
In this paper, we show that the tri-plane discrete data structure encodes rich i
nformation, which can be effectively processed by standard deep-learning machine
ry. We define an extensive benchmark covering a diverse set of fields such as oc
cupancy, signed/unsigned distance, and, for the first time, radiance fields. Wh
ile processing a field with the same reconstruction quality, we achieve task per
formance far superior to frameworks that process large MLPs and, for the first t
ime, almost on par with architectures handling explicit representations.
****
Tianjian Li,Haoran Xu,Philipp Koehn,Daniel Khashabi,Kenton Murray
Error Norm Truncation: Robust Training in the Presence of Data Noise for Text Ge
neration Models
Text generation models are notoriously vulnerable to errors in the training data
. With the wide-spread availability of massive amounts of web-crawled data becom
ing more commonplace, how can we enhance the robustness of models trained on a m
assive amount of noisy web-crawled text? In our work, we propose Error Norm Trun
cation (ENT), a robust enhancement method to the standard training objective tha
t truncates noisy data. Compared to methods that only uses the negative log-like
lihood loss to estimate data quality, our method provides a more accurate estima
tion by considering the distribution of non-target tokens, which is often overlo
oked by previous work. Through comprehensive experiments across language modelin
g, machine translation, and text summarization, we show that equipping text gene

ration models with ENT improves generation quality over standard training and previous soft and hard truncation methods. Furthermore, we show that our method improves the robustness of models against two of the most detrimental types of noise in machine translation, resulting in an increase of more than 2 BLEU points over the MLE baseline when up to 50\% of noise is added to the data.
****
Seyedmorteza Sadat,Jakob Buhmann,Derek Bradley,Otmar Hilliges,Romann M. Weber
CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling
While conditional diffusion models are known to have good coverage of the data distribution, they still face limitations in output diversity, particularly when sampled with a high classifier-free guidance scale for optimal image quality or when trained on small datasets. We attribute this problem to the role of the conditioning signal in inference and offer an improved sampling strategy for diffusion models that can increase generation diversity, especially at high guidance scales, with minimal loss of sample quality. Our sampling strategy anneals the conditioning signal by adding scheduled, monotonically decreasing Gaussian noise to the conditioning vector during inference to balance diversity and condition alignment. Our Condition-Annealed Diffusion Sampler (CADS) can be used with any pretrained model and sampling algorithm, and we show that it boosts the diversity of diffusion models in various conditional generation tasks. Further, using an existing pretrained diffusion model, CADS achieves a new state-of-the-art FID of 1.70 and 2.31 for class-conditional ImageNet generation at 256$\times$256 and 512$\times$512 respectively.
****
Nathan C. Frey,Dan Berenberg,Karina Zadorozhny,Joseph Kleinhenz,Julien Lafrance-Vanasse,Isidro Hotzel,Yan Wu,Stephen Ra,Richard Bonneau,Kyunghyun Cho,Andreas Loukas,Vladimir Gligorijevic,Saeed Saremi
Protein Discovery with Discrete Walk-Jump Sampling
We resolve difficulties in training and sampling from a discrete generative model by learning a smoothed energy function, sampling from the smoothed data manifold with Langevin Markov chain Monte Carlo (MCMC), and projecting back to the true data manifold with one-step denoising. Our $\textit{Discrete Walk-Jump Sampling}$ formalism combines the contrastive divergence training of an energy-based model and improved sample quality of a score-based model, while simplifying training and sampling by requiring only a single noise level. We evaluate the robustness of our approach on generative modeling of antibody proteins and introduce the $\textit{distributional conformity score}$ to benchmark protein generative models. By optimizing and sampling from our models for the proposed distributional conformity score, 97-100\% of generated samples are successfully expressed and purified and 70\% of functional designs show equal or improved binding affinity compared to known functional antibodies on the first attempt in a single round of laboratory experiments. We also report the first demonstration of long-run fast-mixing MCMC chains where diverse antibody protein classes are visited in a single MCMC chain.
****
Kang Liu
SetCSE: Set Operations using Contrastive Learning of Sentence Embeddings
Taking inspiration from Set Theory, we introduce SetCSE, an innovative information retrieval framework. SetCSE employs sets to represent complex semantics and incorporates well-defined operations for structured information querying under the provided context. Within this framework, we introduce an inter-set contrastive learning objective to enhance comprehension of sentence embedding models concerning the given semantics. Furthermore, we present a suite of operations, including SetCSE intersection, difference, and operation series, that leverage sentence embeddings of the enhanced model for complex sentence retrieval tasks. Throughout this paper, we demonstrate that SetCSE adheres to the conventions of human language expressions regarding compounded semantics, provides a significant enhancement in the discriminatory capability of underlying sentence embedding models, and enables numerous information retrieval tasks involving convoluted and intric

ate prompts which cannot be achieved using existing querying methods.
****
Xiao Liu,Hao Yu,Hanchen Zhang,Yifan Xu,Xuanyu Lei,Hanyu Lai,Yu Gu,Hangliang Ding,Kaiwen Men,Kejuan Yang,Shudan Zhang,Xiang Deng,Aohan Zeng,Zhengxiao Du,Chenhui Zhang,Sheng Shen,Tianjun Zhang,Yu Su,Huan Sun,Minlie Huang,Yuxiao Dong,Jie Tang
AgentBench: Evaluating LLMs as Agents
Large Language Models (LLMs) are becoming increasingly smart and autonomous, targeting real-world pragmatic missions beyond traditional NLP tasks.
As a result, there has been an urgent need to evaluate LLMs as agents on challenging tasks in interactive environments.
We present AgentBench, a multi-dimensional evolving benchmark that currently consists of 8 distinct environments to assess LLM-as-Agent's reasoning and decision-making abilities in a multi-turn open-ended generation setting.
Our extensive test over 27 API-based and open-sourced (OSS) LLMs shows that, while top commercial LLMs present a strong ability of acting as agents in complex environments, there is a significant disparity in performance between them and OSS competitors.
We identify the typical reasons of failures in environments and LLMs, showing that poor long-term reasoning, decision-making, and instruction following abilities are the main obstacles for developing usable LLM agents.
Training on code and high quality multi-turn alignment data could improve agent performance.
Datasets, environments, and an integrated evaluation package for AgentBench are released.
****
Ke Wang,Houxing Ren,Aojun Zhou,Zimu Lu,Sichun Luo,Weikang Shi,Renrui Zhang,Linqi Song,Mingjie Zhan,Hongsheng Li
MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning
The recently released GPT-4 Code Interpreter has demonstrated remarkable proficiency in solving challenging math problems, primarily attributed to its ability to seamlessly reason with natural language, generate code, execute code, and continue reasoning based on the execution output. In this paper, we present a method to fine-tune open-source language models, enabling them to use code for modeling and deriving math equations and, consequently, enhancing their mathematical reasoning abilities. We propose a method of generating novel and high-quality datasets with math problems and their code-based solutions, referred to as MathCodeInstruct. Each solution interleaves $\textit{natural language}$, $\textit{code}$, and $\textit{execution results}$. We also introduce a customized supervised fine-tuning and inference approach. This approach yields the MathCoder models, a family of models capable of generating code-based solutions for solving challenging math problems. Impressively, the MathCoder models achieve state-of-the-art scores among open-source LLMs on the MATH (45.2%) and GSM8K (83.9%) datasets, substantially outperforming other open-source alternatives. Notably, the MathCoder model not only surpasses ChatGPT-3.5 and PaLM-2 on GSM8K and MATH but also outperforms GPT-4 on the competition-level MATH dataset. The proposed dataset and models will be released upon acceptance.
****
Jonathan Vacher,Pascal Mamassian
Perceptual Scales Predicted by Fisher Information Metrics
Perception is often viewed as a process that transforms physical variables, external to an observer, into internal psychological variables. Such a process can be modeled by a function coined *perceptual scale*. The *perceptual scale* can be deduced from psychophysical measurements that consist in comparing the relative differences between stimuli (i.e. difference scaling experiments). However, this approach is often overlooked by the modeling and experimentation communities. Here, we demonstrate the value of measuring the *perceptual scale* of classical (spatial frequency, orientation) and less classical physical variables (interpolation between textures) by embedding it in recent probabilistic modeling of perception. First, we show that the assumption that an observer has an internal representation of univariate parameters such as spatial frequency or orientation whi

le stimuli are high-dimensional does not lead to contradictory predictions when following the theoretical framework. Second, we show that the measured *perceptual scale* corresponds to the transduction function hypothesized in this framework. In particular, we demonstrate that it is related to the Fisher information of the generative model that underlies perception and we test the predictions given by the generative model of different stimuli in a set a of difference scaling experiments. Our main conclusion is that the *perceptual scale* is mostly driven by the stimulus power spectrum. Finally, we propose that this measure of *perceptual scale* is a way to push further the notion of perceptual distances by estimating the perceptual geometry of images i.e. the path between images instead of simply the distance between those.
****
Aravind Gollakota,Adam Klivans,Konstantinos Stavropoulos,Arsen Vasilyan
An Efficient Tester-Learner for Halfspaces
We give the first efficient algorithm for learning halfspaces in the testable learning model recently defined by Rubinfeld and Vasilyan [2022]. In this model, a learner certifies that the accuracy of its output hypothesis is near optimal whenever the training set passes an associated test, and training sets drawn from some target distribution must pass the test. This model is more challenging than distribution-specific agnostic or Massart noise models where the learner is allowed to fail arbitrarily if the distributional assumption does not hold. We consider the setting where the target distribution is the standard Gaussian in d dimensions and the label noise is either Massart or adversarial (agnostic). For Massart noise, our tester-learner runs in polynomial time and outputs a hypothesis with (information-theoretically optimal) error $\mathrm{opt}+\epsilon$ (and extends to any fixed strongly log-concave target distribution). For adversarial noise, our tester-learner obtains error $O(\mathrm{opt})+\epsilon$ in polynomial time. Prior work on testable learning ignores the labels in the training set and checks that the empirical moments of the covariates are close to the moments of the base distribution. Here we develop new tests of independent interest that make critical use of the labels and combine them with the moment-matching approach of Gollakota et al. [2022]. This enables us to implement a testable variant of the algorithm of Diakonikolas et al. [2020a, 2020b] for learning noisy halfspaces using nonconvex SGD.
****
Simin Li,Jun Guo,Jingqiao Xiu,Ruixiao Xu,Xin Yu,Jiakai Wang,Aishan Liu,Yaodong Yang,Xianglong Liu
Byzantine Robust Cooperative Multi-Agent Reinforcement Learning as a Bayesian Game
In this study, we explore the robustness of cooperative multi-agent reinforcement learning (c-MARL) against Byzantine failures, where any agent can enact arbitrary, worst-case actions due to malfunction or adversarial attack. To address the uncertainty that any agent can be adversarial, we propose a Bayesian Adversarial Robust Dec-POMDP (BARDec-POMDP) framework, which views Byzantine adversaries as nature-dictated types, represented by a separate transition. This allows agents to learn policies grounded on their posterior beliefs about the type of other agents, fostering collaboration with identified allies and minimizing vulnerability to adversarial manipulation. We define the optimal solution to the BARDec-POMDP as an ex interim robust Markov perfect Bayesian equilibrium, which we proof to exist and the corresponding policy weakly dominates previous approaches as time goes to infinity. To realize this equilibrium, we put forward a two-timescale actor-critic algorithm with almost sure convergence under specific conditions. Experiments on matrix game, Level-based Foraging and StarCraft II indicate that, our method successfully acquires intricate micromanagement skills and adaptively aligns with allies under worst-case perturbations, showing resilience against non-oblivious adversaries, random allies, observation-based attacks, and transfer-based attacks.
****
Maryam Haghighat,Peyman Moghadam,Shaheer Mohamed,Piotr Koniusz
Pre-training with Random Orthogonal Projection Image Modeling

Masked Image Modeling (MIM) is a powerful self-supervised strategy for visual pre-training without the use of labels. MIM applies random crops to input images, processes them with an encoder, and then recovers the masked inputs with a decoder, which encourages the network to capture and learn structural information about objects and scenes. The intermediate feature representations obtained from MIM are suitable for fine-tuning on downstream tasks. In this paper, we propose an Image Modeling framework based on random orthogonal projection instead of binary masking as in MIM. Our proposed Random Orthogonal Projection Image Modeling (ROPIM) reduces spatially-wise token information under guaranteed bound on the noise variance and can be considered as masking entire spatial image area under locally varying masking degrees. Since ROPIM uses a random subspace for the projection that realizes the masking step, the readily available complement of the subspace can be used during unmasking to promote recovery of removed information.
In this paper, we show that using random orthogonal projection leads to superior performance compared to crop-based masking. We demonstrate state-of-the-art results on several popular benchmarks.

****

Lingbing Guo,Zhuo Chen,Jiaoyan Chen,Yin Fang,Wen Zhang,Huajun Chen

Revisit and Outstrip Entity Alignment: A Perspective of Generative Models

Recent embedding-based methods have achieved great successes in exploiting entity alignment from knowledge graph (KG) embeddings of multiple modalities. In this paper, we study embedding-based entity alignment (EEA) from a perspective of generative models. We show that EEA shares similarities with typical generative models and prove the effectiveness of the recently developed generative adversarial network (GAN)-based EEA methods theoretically. We then reveal that their incomplete objective limits the capacity on both entity alignment and entity synthesis (i.e., generating new entities). We mitigate this problem by introducing a generative EEA (GEEA) framework with the proposed mutual variational autoencoder (M-VAE) as the generative model. M-VAE enables entity conversion between KGs and generation of new entities from random noise vectors. We demonstrate the power of GEEA with theoretical analysis and empirical experiments on both entity alignment and entity synthesis tasks. The source code and datasets are available at github.com/zjukg/GEEA.

****

Matt Barnes,Matthew Abueg,Oliver F. Lange,Matt Deeds,Jason Trader,Denali Molitor,Markus Wulfmeier,Shawn O'Banion

Massively Scalable Inverse Reinforcement Learning in Google Maps

Inverse reinforcement learning (IRL) offers a powerful and general framework for learning humans' latent preferences in route recommendation, yet no approach has successfully addressed planetary-scale problems with hundreds of millions of states and demonstration trajectories. In this paper, we introduce scaling techniques based on graph compression, spatial parallelization, and improved initialization conditions inspired by a connection to eigenvector algorithms. We revisit classic IRL methods in the routing context, and make the key observation that there exists a trade-off between the use of cheap, deterministic planners and expensive yet robust stochastic policies. This insight is leveraged in Receding Horizon Inverse Planning (RHIP), a new generalization of classic IRL algorithms that provides fine-grained control over performance trade-offs via its planning horizon. Our contributions culminate in a policy that achieves a 16-24% improvement in route quality at a global scale, and to the best of our knowledge, represents the largest published study of IRL algorithms in a real-world setting to date. We conclude by conducting an ablation study of key components, presenting negative results from alternative eigenvalue solvers, and identifying opportunities to further improve scalability via IRL-specific batching strategies.

****

Yang Liu,Muzhi Zhu,Hengtao Li,Hao Chen,Xinlong Wang,Chunhua Shen

Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching

Powered by large-scale pre-training, vision foundation models exhibit significant potential in open-world image understanding. However, unlike large language models that excel at directly tackling various language tasks, vision foundation m

odels require a task-specific model structure followed by fine-tuning on specific tasks. In this work, we present $\textbf{Matcher}$, a novel perception paradigm that utilizes off-the-shelf vision foundation models to address various perception tasks. Matcher can segment anything by using an in-context example without training. Additionally, we design three effective components within the Matcher framework to collaborate with these foundation models and unleash their full potential in diverse perception tasks. Matcher demonstrates impressive generalization performance across various segmentation tasks, all without training. For example, it achieves 52.7% mIoU on COCO-20^i with one example, surpassing the state-of-the-art specialist model by 1.6%. In addition, Matcher achieves 33.0% mIoU on the proposed LVIS-92^i for one-shot semantic segmentation, outperforming the state-of-the-art generalist model by 14.4%. Our visualization results further showcase the open-world generality and flexibility of Matcher when applied to images in the wild.
****
Saujas Vaduguru,Daniel Fried,Yewen Pu
Generating Pragmatic Examples to Train Neural Program Synthesizers
Programming-by-example is the task of synthesizing a program that is consistent with a set of user-provided input-output examples.
As examples are often an under-specification of one's intent, a good synthesizer must choose the intended program from the many that are consistent with the given set of examples. Prior work frames program synthesis as a cooperative game between a listener (that synthesizes programs) and a speaker (a user choosing examples), and shows that models of computational pragmatic inference is effective in choosing the user intended programs. However, these models requires counterfactual reasoning over a large set of programs and examples, which is infeasible in realistic program spaces. In this paper, we propose a novel way to amortize this search. We sample pairs of programs and examples via self-play between listener and speaker models, and use pragmatic inference to choose informative training examples from this sample.We then use the informative dataset to train models to improve the synthesizer's ability to disambiguate user-provided examples _without human supervision. We validate our method on the challenging task of synthesizing regular expressions from example strings, finding that our method (1) outperforms models trained without choosing pragmatic examples by 23% (a 51% relative increase) (2) matches the performance of supervised learning on a dataset of pragmatic examples provided by humans, despite using no human data in training.
****
Divyat Mahajan,Ioannis Mitliagkas,Brady Neal,Vasilis Syrgkanis
Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation
We study the problem of model selection in causal inference, specifically for the case of conditional average treatment effect (CATE) estimation under binary treatments. Unlike model selection in machine learning, there is no perfect analogue of cross-validation as we do not observe the counterfactual potential outcome for any data point. Towards this, there have been a variety of proxy metrics proposed in the literature, that depend on auxiliary nuisance models estimated from the observed data (propensity score model, outcome regression model). However, the effectiveness of these metrics has only been studied on synthetic datasets as we can access the counterfactual data for them. We conduct an extensive empirical analysis to judge the performance of these metrics introduced in the literature, and novel ones introduced in this work, where we utilize the latest advances in generative modeling to incorporate multiple realistic datasets. Our analysis suggests novel model selection strategies based on careful hyperparameter tuning of CATE estimators and causal ensembling.
****
Xinran Gu,Kaifeng Lyu,Sanjeev Arora,Jingzhao Zhang,Longbo Huang
A Quadratic Synchronization Rule for Distributed Deep Learning
In distributed deep learning with data parallelism, synchronizing gradients at each training step can cause a huge communication overhead, especially when many nodes work together to train large models.
Local gradient methods, such as Local SGD, address this issue by allowing work

ers to compute locally for H steps without synchronizing with others, hence re ducing communication frequency.
While H has been viewed as a hyperparameter to trade optimization efficiency for communication cost, recent research indicates that setting a proper H va lue can lead to generalization improvement. Yet, selecting a proper H is elusi ve. This work proposes a theory-grounded method for determining H, named the Q uadratic Synchronization Rule (QSR), which recommends dynamically setting H in proportion to $\frac{1}{\eta^2}$ as the learning rate η decays over time.
Extensive ImageNet experiments on ResNet and ViT show that local gradient meth ods with QSR consistently improve the test accuracy over other synchronization s trategies. Compared to the standard data parallel training, QSR enables Local Ad amW to cut the training time on 16 or 64 GPUs down from 26.7 to 20.2 hours or fr om 8.6 to 5.5 hours and, at the same time, achieves 1.16% or 0.84% higher top-1 validation accuracy.
****
Ilyes Batatia,Lars Leon Schaaf,Gabor Csanyi,Christoph Ortner,Felix Andreas Faber
Equivariant Matrix Function Neural Networks
Graph Neural Networks (GNNs), especially message-passing neural networks (MPNNs) , have emerged as powerful architectures for learning on graphs in diverse appli cations. However, MPNNs face challenges when modeling non-local interactions in systems such as large conjugated molecules, metals, or amorphous materials.
Although Spectral GNNs and traditional neural networks such as recurrent neural networks and transformers mitigate these challenges, they often lack extensivity , adaptability, generalizability, computational efficiency, or fail to capture d etailed structural relationships or symmetries in the data. To address these con cerns, we introduce Matrix Function Neural Networks (MFNs), a novel architecture that parameterizes non-local interactions through analytic matrix equivariant f unctions. Employing resolvent expansions offers a straightforward implementation and the potential for linear scaling with system size.
The MFN architecture achieves state-of-the-art performance in standard graph ben chmarks, such as the ZINC and TU datasets, and is able to capture intricate non-local interactions in quantum systems. The code and the datasets will be made pu blic.
****
Aditya Desai,Anshumali Shrivastava
In defense of parameter sharing for model-compression
When considering a model architecture, there are several ways to reduce its memo ry footprint. Historically, popular approaches included selecting smaller archit ectures and creating sparse networks through pruning. More recently, randomized parameter-sharing (RPS) methods have gained traction for model compression at start of training. In this paper, we comprehensively assess the trade-off betwee n
memory and accuracy across RPS, pruning techniques, and building smaller models. Our findings demonstrate that RPS, which is both data and model-agnostic, consi stently outperforms smaller models and all moderately informed pruning strategie s, such as MAG, SNIP, SYNFLOW, and GRASP, across the entire compression range. T his advantage becomes particularly pronounced in higher compression scenarios. N otably, even when compared to highly informed pruning techniques like Lottery Ti cket Rewinding (LTR), RPS exhibits superior performance in high compression sett ings. This points out inherent capacity advantage that RPS enjoys over sparse mo dels. Theoretically, we establish RPS as a superior
technique in terms of memory-efficient representation when compared to pruning for linear models. This paper argues in favor of paradigm shift towards RPS base d
models. During our rigorous evaluation of RPS, we identified issues in the state -
of-the-art RPS technique ROAST, specifically regarding stability (ROAST’s sensit ivity to initialization hyperparameters, often leading to divergence) and Pareto -continuity (ROAST’s inability to recover the accuracy of the original model at zero

compression). We provably address both of these issues. We refer to the modified RPS, which incorporates our improvements, as STABLE-RPS
****
Cheng Han,James Chenhao Liang,Qifan Wang,MAJID RABBANI,Sohail Dianat,Raghuveer Rao,Ying Nian Wu,Dongfang Liu
Image Translation as Diffusion Visual Programmers
We introduce the novel Diffusion Visual Programmer (DVP), a neuro-symbolic image translation framework. Our proposed DVP seamlessly embeds a condition-flexible diffusion model within the GPT architecture, orchestrating a coherent sequence of visual programs ($i.e.$, computer vision models) for various pro-symbolic steps, which span RoI identification, style transfer, and position manipulation, facilitating transparent and controllable image translation processes. Extensive experiments demonstrate DVP’s remarkable performance, surpassing concurrent arts. This success can be attributed to several key features of DVP: First, DVP achieves condition-flexible translation via instance normalization, enabling the model to eliminate sensitivity caused by the manual guidance and optimally focus on textual descriptions for high-quality content generation. Second, the frame work enhances in-context reasoning by deciphering intricate high-dimensional concepts in feature spaces into more accessible low-dimensional symbols ($e.g.$, [Prompt], [RoI object]), allowing for localized, context-free editing while maintaining overall coherence. Last but not least, DVP improves systemic controllability and explainability by offering explicit symbolic representations at each programming stage, empowering users to intuitively interpret and modify results. Our research marks a substantial step towards harmonizing artificial image translation processes with cognitive intelligence, promising broader applications.
****
Peizhong Ju,Arnob Ghosh,Ness Shroff
Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning
Fairness plays a crucial role in various multi-agent systems (e.g., communication networks, financial markets, etc.). Many multi-agent dynamical interactions can be cast as Markov Decision Processes (MDPs). While existing research has focused on studying fairness in known environments, the exploration of fairness in such systems for unknown environments remains open. In this paper, we propose a Reinforcement Learning (RL) approach to achieve fairness in multi-agent finite-horizon episodic MDPs. Instead of maximizing the sum of individual agents' value functions, we introduce a fairness function that ensures equitable rewards across agents. Since the classical Bellman's equation does not hold when the sum of individual value functions is not maximized, we cannot use traditional approaches. Instead, in order to explore, we maintain a confidence bound of the unknown environment and then propose an online convex optimization based approach to obtain a policy constrained to this confidence region. We show that such an approach achieves sub-linear regret in terms of the number of episodes. Additionally, we provide a probably approximately correct (PAC) guarantee based on the obtained regret bound. We also propose an offline RL algorithm and bound the optimality gap with respect to the optimal fair solution. To mitigate computational complexity, we introduce a policy-gradient type method for the fair objective. Simulation experiments also demonstrate the efficacy of our approach.
****
Dongjun Kim,Chieh-Hsin Lai,Wei-Hsiang Liao,Naoki Murata,Yuhta Takida,Toshimitsu Uesaka,Yutong He,Yuki Mitsufuji,Stefano Ermon
Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion
Consistency Models (CM) (Song et al., 2023) accelerate score-based diffusion model sampling at the cost of sample quality but lack a natural way to trade-off quality for speed. To address this limitation, we propose Consistency Trajectory Model (CTM), a generalization encompassing CM and score-based models as special cases. CTM trains a single neural network that can -- in a single forward pass -- output scores (i.e., gradients of log-density) and enables unrestricted traversal between any initial and final time along the Probability Flow Ordinary Differential Equation (ODE) in a diffusion process. CTM enables the efficient combinat

ion of adversarial training and denoising score matching loss to enhance performance and achieves new state-of-the-art FIDs for single-step diffusion model sampling on CIFAR-10 (FID 1.73) and ImageNet at 64X64 resolution (FID 1.92). CTM also enables a new family of sampling schemes, both deterministic and stochastic, involving long jumps along the ODE solution trajectories. It consistently improves sample quality as computational budgets increase, avoiding the degradation seen in CM. Furthermore, unlike CM, CTM's access to the score function can streamline the adoption of established controllable/conditional generation methods from the diffusion community. This access also enables the computation of likelihood. The code is available at https://github.com/sony/ctm.
****
Sen Pei
Image Background Serves as Good Proxy for Out-of-distribution Data
Out-of-distribution (OOD) detection empowers the model trained on the closed image set to identify unknown data in the open world. Though many prior techniques have yielded considerable improvements in this research direction, two crucial obstacles still remain. Firstly, a unified perspective has yet to be presented to view the developed arts with individual designs, which is vital for providing insights into future work. Secondly, we expect sufficient natural OOD supervision to promote the generation of compact boundaries between the in-distribution (ID) and OOD data without collecting explicit OOD samples. To tackle these issues, we propose a general probabilistic framework to interpret many existing methods and an OOD-data-free model, namely $\textbf{S}$elf-supervised $\textbf{S}$ampling for $\textbf{O}$OD $\textbf{D}$etection (SSOD). SSOD efficiently exploits natural OOD signals from the ID data based on the local property of convolution. With these supervisions, it jointly optimizes the OOD detection and conventional ID classification in an end-to-end manner. Extensive experiments reveal that SSOD establishes competitive state-of-the-art performance on many large-scale benchmarks, outperforming the best previous method by a large margin, e.g., reporting $\textbf{-6.28}$% FPR95 and $\textbf{+0.77}$% AUROC on ImageNet, $\textbf{-19.01}$% FPR95 and $\textbf{+3.04}$% AUROC on CIFAR-10, and top-ranked performance on hard OOD datasets, i.e., ImageNet-O and OpenImage-O.
****
Ling Pan,Moksh Jain,Kanika Madan,Yoshua Bengio
Pre-Training and Fine-Tuning Generative Flow Networks
Generative Flow Networks (GFlowNets) are amortized samplers that learn stochastic policies to sequentially generate compositional objects from a given unnormalized reward distribution.
They can generate diverse sets of high-reward objects, which is an important consideration in scientific discovery tasks. However, as they are typically trained from a given extrinsic reward function, it remains an important open challenge about how to leverage the power of pre-training and train GFlowNets in an unsupervised fashion for efficient adaptation to downstream tasks.
Inspired by recent successes of unsupervised pre-training in various domains, we introduce a novel approach for reward-free pre-training of GFlowNets. By framing the training as a self-supervised problem, we propose an outcome-conditioned GFlowNet (OC-GFN) that learns to explore the candidate space. Specifically, OC-GFN learns to reach any targeted outcomes, akin to goal-conditioned policies in reinforcement learning.
We show that the pre-trained OC-GFN model can allow for a direct extraction of a policy capable of sampling from any new reward functions in downstream tasks.
Nonetheless, adapting OC-GFN on a downstream task-specific reward involves an intractable marginalization over possible outcomes. We propose a novel way to approximate this marginalization by learning an amortized predictor enabling efficient fine-tuning.
Extensive experimental results validate the efficacy of our approach, demonstrating the effectiveness of pre-training the OC-GFN, and its ability to swiftly adapt to downstream tasks and discover modes more efficiently.
This work may serve as a foundation for further exploration of pre-training strategies in the context of GFlowNets.

****
Zhepei Wei,Chuanhao Li,Tianze Ren,Haifeng Xu,Hongning Wang
Incentivized Truthful Communication for Federated Bandits
To enhance the efficiency and practicality of federated bandit learning, recent advances have introduced incentives to motivate communication among clients, where a client participates only when the incentive offered by the server outweighs its participation cost. However, existing incentive mechanisms naively assume the clients are truthful: they all report their true cost and thus the higher cost one participating client claims, the more the server has to pay. Therefore, such mechanisms are vulnerable to strategic clients aiming to optimize their own utility by misreporting. To address this issue, we propose an incentive compatible (i.e., truthful) communication protocol, named Truth-FedBan, where the incentive for each participant is independent of its self-reported cost, and reporting the true cost is the only way to achieve the best utility. More importantly, Truth-FedBan still guarantees the sub-linear regret and communication cost without any overhead. In other words, the core conceptual contribution of this paper is, for the first time, demonstrating the possibility of simultaneously achieving incentive compatibility and nearly optimal regret in federated bandit learning. Extensive numerical studies further validate the effectiveness of our proposed solution.
****
Saeed Saremi,Ji Won Park,Francis Bach
Chain of Log-Concave Markov Chains
We introduce a theoretical framework for sampling from unnormalized densities based on a smoothing scheme that uses an isotropic Gaussian kernel with a single fixed noise scale. We prove one can decompose sampling from a density (minimal assumptions made on the density) into a sequence of sampling from log-concave conditional densities via accumulation of noisy measurements with equal noise levels. Our construction is unique in that it keeps track of a history of samples, making it non-Markovian as a whole, but it is lightweight algorithmically as the history only shows up in the form of a running empirical mean of samples. Our sampling algorithm generalizes walk-jump sampling (Saremi & Hyvärinen, 2019). The "walk" phase becomes a (non-Markovian) chain of (log-concave) Markov chains. The "jump" from the accumulated measurements is obtained by empirical Bayes. We study our sampling algorithm quantitatively using the 2-Wasserstein metric and compare it with various Langevin MCMC algorithms. We also report a remarkable capacity of our algorithm to "tunnel" between modes of a distribution.
****
Zihao Zhu,Mingda Zhang,Shaokui Wei,Bingzhe Wu,Baoyuan Wu
VDC: Versatile Data Cleanser based on Visual-Linguistic Inconsistency by Multimodal Large Language Models
The role of data in building AI systems has recently been emphasized by the emerging concept of data-centric AI. Unfortunately, in the real-world, datasets may contain dirty samples, such as poisoned samples from backdoor attack, noisy labels in crowdsourcing, and even hybrids of them. The presence of such dirty samples makes the DNNs vunerable and unreliable.
Hence, it is critical to detect dirty samples to improve the quality and realiability of dataset.
Existing detectors only focus on detecting poisoned samples or noisy labels, that are often prone to weak generalization when dealing with dirty samples from other fields.
In this paper, we find a commonality of various dirty samples is visual-linguistic inconsistency between images and associated labels.
To capture the semantic inconsistency between modalities, we propose versatile data cleanser (VDC) leveraging the surpassing capabilities of multimodal large language models (MLLM) in cross-modal alignment and reasoning.
It consists of three consecutive modules: the visual question generation module to generate insightful questions about the image; the visual question answering module to acquire the semantics of the visual content by answering the questions with MLLM; followed by the visual answer evaluation module to evaluate the inco

nsistency.
Extensive experiments demonstrate its superior performance and generalization to various categories and types of dirty samples.
The code is available at https://github.com/zihao-ai/vdc.
****
Xiong Zhou,Xianming Liu,Hao Yu,Jialiang Wang,Zeke Xie,Junjun Jiang,Xiangyang Ji
Variance-enlarged Poisson Learning for Graph-based Semi-Supervised Learning with Extremely Sparse Labeled Data
Graph-based semi-supervised learning, particularly in the context of extremely sparse labeled data, often suffers from degenerate solutions where label functions tend to be nearly constant across unlabeled data. In this paper, we introduce Variance-enlarged Poisson Learning (VPL), a simple yet powerful framework tailored to alleviate the issues arising from the presence of degenerate solutions. VPL incorporates a variance-enlarged regularization term, which induces a Poisson equation specifically for unlabeled data. This intuitive approach increases the dispersion of labels from their average mean, effectively reducing the likelihood of degenerate solutions characterized by nearly constant label functions. We subsequently introduce two streamlined algorithms, V-Laplace and V-Poisson, each intricately designed to enhance Laplace and Poisson learning, respectively. Furthermore, we broaden the scope of VPL to encompass graph neural networks, introducing Variance-enlarged Graph Poisson Networks (V-GPN) to facilitate improved label propagation. To achieve a deeper understanding of VPL's behavior, we conduct a comprehensive theoretical exploration in both discrete and variational cases. Our findings elucidate that VPL inherently amplifies the importance of connections within the same class while concurrently tempering those between different classes. We support our claims with extensive experiments, demonstrating the effectiveness of VPL and showcasing its superiority over existing methods. The code is available at https://github.com/hitcszx/VPL.
****
Jiawei Yang,Boris Ivanovic,Or Litany,Xinshuo Weng,Seung Wook Kim,Boyi Li,Tong Che,Danfei Xu,Sanja Fidler,Marco Pavone,Yue Wang
EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision
We present EmerNeRF, a simple yet powerful approach for learning spatial-temporal representations of dynamic driving scenes. Grounded in neural fields, EmerNeRF simultaneously captures scene geometry, appearance, motion, and semantics via self-bootstrapping. EmerNeRF hinges upon two core components: First, it stratifies scenes into static and dynamic fields. This decomposition emerges purely from self-supervision, enabling our model to learn from general, in-the-wild data sources. Second, EmerNeRF parameterizes an induced flow field from the dynamic field and uses this flow field to further aggregate multi-frame features, amplifying the rendering precision of dynamic objects. Coupling these three fields (static, dynamic, and flow) enables EmerNeRF to represent highly-dynamic scenes self-sufficiently, without relying on ground truth object annotations or pre-trained models for dynamic object segmentation or optical flow estimation. Our method achieves state-of-the-art performance in sensor simulation, significantly outperforming previous methods when reconstructing static (+2.39 PSNR) and dynamic (+3.25 PSNR) scenes. In addition, to bolster EmerNeRF's semantic generalization, we lift 2D visual foundation model features into 4D space-time and address a general positional bias in modern Transformers, significantly boosting 3D perception performance (e.g., 78.5% relative improvement in occupancy prediction accuracy). Finally, we construct a diverse and challenging 120-sequence dataset to benchmark neural fields under extreme and highly-dynamic settings. Visualizations, code, and data will be anonymously available at https://anonymous.4open.science/r/EmerNeRF_review-003B/
****
Haruka Kiyohara,Ren Kishimoto,Kosuke Kawakami,Ken Kobayashi,Kazuhide Nakata,Yuta Saito
Towards Assessing and Benchmarking Risk-Return Tradeoff of Off-Policy Evaluation
Off-Policy Evaluation (OPE) aims to assess the effectiveness of counterfactu

al policies using offline logged data and is frequently utilized to identify the top-k promising policies for deployment in online A/B tests. Existing evaluation metrics for OPE estimators primarily focus on the "accuracy" of OPE or that of downstream policy selection, neglecting risk-return tradeoff and *efficiency* in subsequent online policy deployment. To address this issue, we draw inspiration from portfolio evaluation in finance and develop a new metric, called **SharpeRatio@k**, which measures the risk-return tradeoff and efficiency of policy portfolios formed by an OPE estimator under varying online evaluation budgets (k). We first demonstrate, in two example scenarios, that our proposed metric can clearly distinguish between conservative and high-stakes OPE estimators and reliably identify the most *efficient* estimator capable of forming superior portfolios of candidate policies that maximize return with minimal risk during online deployment, while existing evaluation metrics produce only degenerate results. To facilitate a quick, accurate, and consistent evaluation of OPE via SharpeRatio@k, we have also implemented the proposed metric in an open-source software. Using SharpeRatio@k and the software, we conduct a benchmark experiment of various OPE estimators regarding their risk-return tradeoff, presenting several future directions for OPE research.
****
Riku Togashi,Tatsushi Oka,Naoto Ohsaka,Tetsuro Morimura
Safe Collaborative Filtering
Excellent tail performance is crucial for modern machine learning tasks, such as algorithmic fairness, class imbalance, and risk-sensitive decision making, as it ensures the effective handling of challenging samples within a dataset. Tail performance is also a vital determinant of success for personalized recommender systems to reduce the risk of losing users with low satisfaction. This study introduces a "safe" collaborative filtering method that prioritizes recommendation quality for less-satisfied users rather than focusing on the average performance. Our approach minimizes the conditional value at risk (CVaR), which represents the average risk over the tails of users' loss. To overcome computational challenges for web-scale recommender systems, we develop a robust yet practical algorithm that extends the most scalable method, implicit alternating least squares (iALS). Empirical evaluation on real-world datasets demonstrates the excellent tail performance of our approach while maintaining competitive computational efficiency.
****
Zihao Yin,Chen Gan,Kelei He,Yang Gao,Junfeng Zhang
Hybrid Sharing for Multi-Label Image Classification
Existing multi-label classification methods have long suffered from label heterogeneity, where learning a label obscures another. By modeling multi-label classification as a multi-task problem, this issue can be regarded as a negative transfer, which indicates challenges to achieve simultaneously satisfied performance across multiple tasks. In this work, we propose the Hybrid Sharing Query (HSQ), a transformer-based model that introduces the mixture-of-experts architecture to image multi-label classification. HSQ is designed to leverage label correlations while mitigating heterogeneity effectively. To this end, HSQ is incorporated with a fusion expert framework that enables it to optimally combine the strengths of task-specialized experts with shared experts, ultimately enhancing multi-label classification performance across most labels. Extensive experiments are conducted on two benchmark datasets, with the results demonstrating that the proposed method achieves state-of-the-art performance and yields simultaneous improvements across most labels. The code is available at https://github.com/zihao-yin/HSQ
****
Linfeng Ye,Shayan Mohajer Hamidi,Renhao Tan,EN-HUI YANG
Bayes Conditional Distribution Estimation for Knowledge Distillation Based on Conditional Mutual Information
It is believed that in knowledge distillation (KD), the role of the teacher is to provide an estimate for the unknown Bayes conditional probability distribution (BCPD) to be used in the student training process. Conventionally, this estimat

e is obtained by training the teacher using maximum log-likelihood (MLL) method. To improve this estimate for KD, in this paper we introduce the concept of conditional mutual information (CMI) into the estimation of BCPD and propose a novel estimator called the maximum CMI (MCMI) method. Specifically, in MCMI estimation, both the log-likelihood and CMI of the teacher are simultaneously maximized when the teacher is trained. In fact, maximizing the teacher's CMI value ensures that the teacher can effectively capture the contextual information within the images, and for visualizing this information, we deploy Eigen-CAM. Via conducting a thorough set of experiments, we show that by employing a teacher trained via MCMI estimation rather than one trained via MLL estimation in various state-of-the-art KD frameworks, the student's classification accuracy consistently increases, with the gain of up to 3.32\%. This suggests that the teacher's BCPD estimate provided by MCMI method is more accurate than that provided by MLL method. In addition, we show that such improvements in the student's accuracy are more drastic in zero-shot and few-shot settings. Notably, the student's accuracy increases with the gain of up to 5.72\% when 5\% of the training samples are available to student (few-shot), and increases from 0\% to as high as 84\% for an omitted class (zero-shot).
****
Kim-Celine Kahl,Carsten T. Lüth,Maximilian Zenk,Klaus Maier-Hein,Paul F Jaeger
ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation
Uncertainty estimation is an essential and heavily-studied component for the reliable application of semantic segmentation methods. While various studies exist claiming methodological advances on the one hand, and successful application on the other hand, the field is currently hampered by a gap between theory and practice leaving fundamental questions unanswered: Can data-related and model-related uncertainty really be separated in practice? Which components of an uncertainty method are essential for real-world performance? Which uncertainty method works well for which application? In this work, we link this research gap to a lack of systematic and comprehensive evaluation of uncertainty methods. Specifically, we identify three key pitfalls in current literature and present an evaluation framework that bridges the research gap by providing 1) a controlled environment for studying data ambiguities as well as distribution shifts, 2) systematic ablations of relevant method components, and 3) test-beds for the five predominant uncertainty applications: OoD-detection, active learning, failure detection, calibration, and ambiguity modeling. Empirical results on simulated as well as real-world data demonstrate how the proposed framework is able to answer the predominant questions in the field revealing for instance that 1) separation of uncertainty types works on simulated data but does not necessarily translate to real-world data, 2) aggregation of scores is a crucial but currently neglected component of uncertainty methods, 3) While ensembles are performing most robustly across the different downstream tasks and settings, test-time augmentation often constitutes a light-weight alternative. Code is at: https://github.com/IML-DKFZ/values
****
Wenhao Zhan,Masatoshi Uehara,Wen Sun,Jason D. Lee
Provable Reward-Agnostic Preference-Based Reinforcement Learning
Preference-based Reinforcement Learning (PbRL) is a paradigm in which an RL agent learns to optimize a task using pair-wise preference-based feedback over trajectories, rather than explicit reward signals. While PbRL has demonstrated practical success in fine-tuning language models, existing theoretical work focuses on regret minimization and fails to capture most of the practical frameworks. In this study, we fill in such a gap between theoretical PbRL and practical algorithms by proposing a theoretical reward-agnostic PbRL framework where exploratory trajectories that enable accurate learning of hidden reward functions are acquired before collecting any human feedback. Theoretical analysis demonstrates that our algorithm requires less human feedback for learning the optimal policy under preference-based models with linear parameterization and unknown transitions, compared to the existing theoretical literature. Specifically, our framework can i

ncorporate linear and low-rank MDPs with efficient sample complexity. Additionally, we investigate reward-agnostic RL with action-based comparison feedback and introduce an efficient querying algorithm tailored to this scenario.
****
Xinyu Zhao,Xuxi Chen,Yu Cheng,Tianlong Chen
Sparse MoE with Language Guided Routing for Multilingual Machine Translation
Sparse Mixture-of-Experts (SMoE) has gained increasing popularity as a promising framework for scaling up multilingual machine translation (MMT) models with negligible extra computational overheads. However, current SMoE solutions neglect the intrinsic structures of the MMT problem: (a) $\textit{Linguistics Hierarchy.}$ Languages are naturally grouped according to their lingual properties like genetic families, phonological characteristics, etc; (b) $\textit{Language Complexity.}$ The learning difficulties are varied for diverse languages due to their grammar complexity, available resources, etc. Therefore, routing a fixed number of experts (e.g., 1 or 2 experts in usual) only at the word level leads to inferior performance. To fill in the missing puzzle, we propose $\textbf{\texttt{Lingual-SMoE}}$ by equipping the SMoE with adaptive and linguistic-guided routing policies. Specifically, it (1) extracts language representations to incorporate linguistic knowledge and uses them to allocate experts into different groups; (2) determines the number of activated experts for each target language in an adaptive and automatic manner, according to their translation difficulties, which aims to mitigate the potential over-/under-fitting issues of learning simple/challenges translations. Sufficient experimental studies on MMT benchmarks with {16, 50, 100} language pairs and various network architectures, consistently validate the superior performance of our proposals. For instance, $\texttt{Lingual-SMoE}$ outperforms its dense counterpart by over 5% BLEU scores on $\texttt{OPUS-100}$ dataset. Codes are included in the supplement.
****
Shengchao Liu,Jiongxiao Wang,Yijin Yang,Chengpeng Wang,Ling Liu,Hongyu Guo,Chaowei Xiao
Conversational Drug Editing Using Retrieval and Domain Feedback
Recent advancements in conversational large language models (LLMs), such as ChatGPT, have demonstrated remarkable promise in various domains, including drug discovery. However, existing works mainly focus on investigating the capabilities of conversational LLMs on chemical reactions and retrosynthesis. While drug editing, a critical task in the drug discovery pipeline, remains largely unexplored. To bridge this gap, we propose ChatDrug, a framework to facilitate the systematic investigation of drug editing using LLMs. ChatDrug jointly leverages a prompt module, a retrieval and domain feedback module, and a conversation module to streamline effective drug editing. We empirically show that ChatDrug reaches the best performance on all 39 drug editing tasks, encompassing small molecules, peptides, and proteins. We further demonstrate, through 10 case studies, that ChatDrug can successfully identify the key substructures for manipulation, generating diverse and valid suggestions for drug editing. Promisingly, we also show that ChatDrug can offer insightful explanations from a domain-specific perspective, enhancing interpretability and enabling informed decision-making.
****
Tserendorj Adiya,Jae Shin Yoon,JUNGEUN LEE,Sanghun Kim,Hwasup Lim
Bidirectional Temporal Diffusion Model for Temporally Consistent Human Animation
We introduce a method to generate temporally coherent human animation from a single image, a video, or a random noise.
This problem has been formulated as modeling of an auto-regressive generation, i.e., to regress past frames to decode future frames.
However, such unidirectional generation is highly prone to motion drifting over time, generating unrealistic human animation with significant artifacts such as appearance distortion.
We claim that bidirectional temporal modeling enforces temporal coherence on a generative network by largely suppressing the appearance ambiguity.
To prove our claim, we design a novel human animation framework using a denoising diffusion model:

a neural network learns to generate the image of a person by denoising temporal Gaussian noises whose intermediate results are cross-conditioned bidirectionally between consecutive frames.
In the experiments, our method demonstrates strong performance compared to existing unidirectional approaches with realistic temporal coherence.
****
Licong Lin,Yu Bai,Song Mei
Transformers as Decision Makers: Provable In-Context Reinforcement Learning via Supervised Pretraining
Large transformer models pretrained on offline reinforcement learning datasets have demonstrated remarkable in-context reinforcement learning (ICRL) capabilities, where they can make good decisions when prompted with interaction trajectories from unseen environments. However, when and how transformers can be trained to perform ICRL have not been theoretically well-understood. In particular, it is unclear which reinforcement-learning algorithms transformers can perform in context, and how distribution mismatch in offline training data affects the learned algorithms.

This paper provides a theoretical framework that analyzes supervised pretraining for ICRL. This includes two recently proposed training methods --- algorithm distillation and decision-pretrained transformers. First, assuming model realizability, we prove the supervised-pretrained transformer will imitate the conditional expectation of the expert algorithm given the observed trajectory. The generalization error will scale with model capacity and a distribution divergence factor between the expert and offline algorithms. Second, we show transformers with ReLU attention can efficiently approximate near-optimal online reinforcement learning algorithms like LinUCB and Thompson sampling for stochastic linear bandits, and UCB-VI for tabular Markov decision processes. This provides the first quantitative analysis of the ICRL capabilities of transformers pretrained from offline trajectories.
****
Xiang Yue,Xingwei Qu,Ge Zhang,Yao Fu,Wenhao Huang,Huan Sun,Yu Su,Wenhu Chen
MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning
We introduce MAmmoTH, a series of open-source large language models (LLMs) specifically tailored for general math problem-solving. The MAmmoTH models are trained on MathInstruct, our meticulously curated instruction tuning dataset. MathInstruct is compiled from 13 math datasets with intermediate rationales, six of which have rationales newly curated by us. It presents a unique hybrid of chain-of-thought (CoT) and program-of-thought (PoT) rationales, and also ensures extensive coverage of diverse fields in math. The hybrid of CoT and PoT not only unleashes the potential of tool use but also allows different thought processes for different math problems. As a result, the MAmmoTH series substantially outperform existing open-source models on nine mathematical reasoning datasets across all scales with an average accuracy gain between 16% and 32%. Remarkably, our MAmmoTH-7B model reaches 33% on MATH (a competition-level dataset), which exceeds the best open-source 7B model (WizardMath) by 23%, and the MAmmoTH-34B model achieves 44% accuracy on MATH, even surpassing GPT-4’s CoT result. Our work underscores the importance of diverse problem coverage and the use of hybrid rationales in developing superior math generalist models.
****
Andrew William Engel,Zhichao Wang,Natalie Frank,Ioana Dumitriu,Sutanay Choudhury,Anand Sarwate,Tony Chiang
Faithful and Efficient Explanations for Neural Networks via Neural Tangent Kernel Surrogate Models
A recent trend in explainable AI research has focused on surrogate modeling, where neural networks are approximated as simpler ML algorithms such as kernel machines. A second trend has been to utilize kernel functions in various explain-by-example or data attribution tasks. In this work, we combine these two trends to analyze approximate empirical neural tangent kernels (eNTK) for data attribution. Approximation is critical for eNTK analysis due to the high computational cost

 to compute the eNTK. We define new approximate eNTK and perform novel analysis on how well the resulting kernel machine surrogate models correlate with the und erlying neural network. We introduce two new random projection variants of appro ximate eNTK which allow users to tune the time and memory complexity of their ca lculation. We conclude that kernel machines using approximate neural tangent ker nel as the kernel function are effective surrogate models, with the introduced t race NTK the most consistent performer.
****
Shida Wang,Zhong Li,Qianxiao Li
Inverse Approximation Theory for Nonlinear Recurrent Neural Networks
We prove an inverse approximation theorem for the approximation of nonlinear seq uence-to-sequence relationships using recurrent neural networks (RNNs). This is a so-called Bernstein-type result in approximation theory, which deduces propert ies of a target function under the assumption that it can be effectively approxi mated by a hypothesis space. In particular, we show that nonlinear sequence rela tionships that can be stably approximated by nonlinear RNNs must have an exponen tial decaying memory structure - a notion that can be made precise. This extends the previously identified curse of memory in linear RNNs into the general nonli near setting, and quantifies the essential limitations of the RNN architecture f or learning sequential relationships with long-term memory. Based on the analysi s, we propose a principled reparameterization method to overcome the limitations . Our theoretical results are confirmed by numerical experiments.
****
Martino Bernasconi,Matteo Castiglioni,Andrea Celli,Federico Fusco
Bandits with Replenishable Knapsacks: the Best of both Worlds
The bandits with knapsacks (BwK) framework models online decision-making problem s in which an agent makes a sequence of decisions subject to resource consumptio n constraints. The traditional model assumes that each action consumes a non-neg ative amount of resources and the process ends when the initial budgets are full y depleted. We study a natural generalization of the BwK framework which allows non-monotonic resource utilization, i.e., resources can be replenished by a posi tive amount. We propose a best-of-both-worlds primal-dual template that can hand le any online learning problem with replenishment for which a suitable primal re gret minimizer exists. In particular, we provide the first positive results for the case of adversarial inputs by showing that our framework guarantees a consta nt competitive ratio α when $B=\Omega(T)$ or when the possible per-round replenishment is a positive constant. Moreover, under a stochastic input model, our algorithm yields an instance-independent $\tilde{\mathcal{O}}(T^{1/2})$ regr et bound which complements existing instance-dependent bounds for the same setti ng. Finally, we provide applications of our framework to some economic problems of practical relevance.
****
Daixuan Cheng,Shaohan Huang,Furu Wei
Adapting Large Language Models via Reading Comprehension
We explore how continued pre-training on domain-specific corpora influences larg e language models, revealing that training on the raw corpora endows the model w ith domain knowledge, but drastically hurts its prompting ability for question a nswering. Taken inspiration from human learning via reading comprehension--pract ice after reading improves the ability to answer questions based on the learned knowledge--we propose a simple method for transforming raw corpora into reading comprehension texts. Each raw text is enriched with a series of tasks related to its content. Our method, highly scalable and applicable to any pre-training cor pora, consistently enhances performance across various tasks in three different domains: biomedicine, finance, and law. Notably, our 7B language model achieves competitive performance with domain-specific models of much larger scales, such as BloombergGPT-50B. Furthermore, we demonstrate that domain-specific reading co mprehension texts can improve the model's performance even on general benchmarks , showing the potential to develop a general model across even more domains. Our model, code, and data are available at https://github.com/microsoft/LMOps.
****

Jiachun Pan,Jun Hao Liew,Vincent Tan,Jiashi Feng,Hanshu Yan
AdjointDPM: Adjoint Sensitivity Method for Gradient Backpropagation of Diffusion Probabilistic Models
This paper considers a ubiquitous problem underlying several applications of DPMs, i.e.,
optimizing the parameters of DPMs when the objective is a differentiable metric defined on the generated contents.
Since the sampling procedure of DPMs involves recursive calls to the denoising UNet, naive gradient backpropagation requires storing the intermediate states of all iterations, resulting in extremely high memory consumption.
To overcome this issue, we propose a novel method AdjointDPM, which first generates new samples from diffusion models by solving the corresponding probability-flow ODEs. It then uses the adjoint sensitivity method to backpropagate the gradients of the loss to the models' parameters (including conditioning signals, network weights, and initial noises) by solving another augmented ODE.
To reduce numerical errors in both the forward generation and gradient backpropagation processes, we further reparameterize the probability-flow ODE and augmented ODE as simple non-stiff ODEs using exponential integration.
AdjointDPM can effectively compute the gradients of all types of parameters in DPMs, including the network weights, conditioning text prompts, and noisy states.
Finally, we demonstrate the effectiveness of AdjointDPM on several interesting tasks: guided generation via modifying sampling trajectories, finetuning DPM weights for stylization, and converting visual effects into text embeddings.
****
Jingyu Chen,Runlin Lei,Zhewei Wei
PolyGCL: GRAPH CONTRASTIVE LEARNING via Learnable Spectral Polynomial Filters
Recently, Graph Contrastive Learning (GCL) has achieved significantly superior performance in self-supervised graph representation learning.
However, the existing GCL technique has inherent smooth characteristics because of its low-pass GNN encoder and objective based on homophily assumption, which poses a challenge when applying it to heterophilic graphs.
In supervised learning tasks, spectral GNNs with polynomial approximation excel in both homophilic and heterophilic settings by adaptively fitting graph filters of arbitrary shapes.
Yet, their applications in unsupervised learning are rarely explored.
Based on the above analysis, a natural question arises: Can we incorporate the excellent properties of spectral polynomial filters into graph contrastive learning?
In this paper, we address the question by studying the necessity of introducing high-pass information for heterophily from a spectral perspective.
We propose PolyGCL, a GCL pipeline that utilizes polynomial filters to achieve contrastive learning between the low-pass and high-pass views.
Specifically, PolyGCL utilizes polynomials with learnable filter functions to generate different spectral views and an objective that incorporates high-pass information through a linear combination.
We theoretically prove that PolyGCL outperforms previous GCL paradigms when applied to graphs with varying levels of homophily.
We conduct extensive experiments on both synthetic and real-world datasets, which demonstrate the promising performance of PolyGCL on homophilic and heterophilic graphs.
****
Theo X. Olausson,Jeevana Priya Inala,Chenglong Wang,Jianfeng Gao,Armando Solar-Lezama
Is Self-Repair a Silver Bullet for Code Generation?
Large language models have shown remarkable aptitude in code generation, but still struggle to perform complex tasks. Self-repair---in which the model debugs and repairs its own code---has recently become a popular way to boost performance in these settings. However, despite its increasing popularity, existing studies of self-repair have been limited in scope; in many settings, its efficacy thus remains poorly understood. In this paper, we analyze Code Llama, GPT-3.5 and GPT-

4's ability to perform self-repair on problems taken from HumanEval and APPS. We
 find that when the cost of carrying out repair is taken into account, performan
ce gains are often modest, vary a lot between subsets of the data, and are somet
imes not present at all. We hypothesize that this is because self-repair is bott
lenecked by the model's ability to provide feedback on its own code; using a str
onger model to artificially boost the quality of the feedback, we observe substa
ntially larger performance gains. Similarly, a small-scale study in which we pro
vide GPT-4 with feedback from human participants suggests that even for the stro
ngest models, self-repair still lags far behind what can be achieved with human-
level debugging.
****
Runpei Dong,Chunrui Han,Yuang Peng,Zekun Qi,Zheng Ge,Jinrong Yang,Liang Zhao,Jia
njian Sun,Hongyu Zhou,Haoran Wei,Xiangwen Kong,Xiangyu Zhang,Kaisheng Ma,Li Yi
DreamLLM: Synergistic Multimodal Comprehension and Creation
This paper presents DreamLLM, a learning framework that first achieves versatile
 Multimodal Large Language Models (MLLMs) empowered with frequently overlooked s
ynergy between multimodal comprehension and creation. DreamLLM operates on two f
undamental principles. The first focuses on the generative modeling of both lang
uage and image posteriors by direct sampling in the raw multimodal space. This a
pproach circumvents the limitations and information loss inherent to external fe
ature extractors like CLIP, and a more thorough multimodal understanding is obta
ined. Second, DreamLLM fosters the generation of raw, interleaved documents, mod
eling both text and image contents, along with unstructured layouts. This allows
 DreamLLM to learn all conditional, marginal, and joint multimodal distributions
 effectively. As a result, DreamLLM is the first MLLM capable of generating free
-form interleaved content. Comprehensive experiments highlight DreamLLM's superi
or performance as a zero-shot multimodal generalist, reaping from the enhanced l
earning synergy. Project page: https://dreamllm.github.io.
****
Devavrat Tomar,Guillaume Vray,Jean-Philippe Thiran,Behzad Bozorgtabar
Un-Mixing Test-Time Normalization Statistics: Combatting Label Temporal Correlat
ion
Recent test-time adaptation methods heavily rely on nuanced adjustments of batch
 normalization (BN) parameters. However, one critical assumption often goes over
looked: that of independently and identically distributed (i.i.d.) test batches
with respect to unknown labels. This oversight leads to skewed BN statistics an
d undermines the reliability of the model under non-i.i.d. scenarios. To tackle
this challenge, this paper presents a novel method termed '$\textbf{Un-Mix}$ing
$\textbf{T}$est-Time $\textbf{N}$ormalization $\textbf{S}$tatistics' (UnMix-TNS)
. Our method re-calibrates the statistics for each instance within a test batch
by $\textit{mixing}$ it with multiple distinct statistics components, thus inher
ently simulating the i.i.d. scenario. The core of this method hinges on a distin
ctive online $\textit{unmixing}$ procedure that continuously updates these stati
stics components by incorporating the most similar instances from new test batch
es. Remarkably generic in its design, UnMix-TNS seamlessly integrates with a wid
e range of leading test-time adaptation methods and pre-trained architectures eq
uipped with BN layers. Empirical evaluations corroborate the robustness of UnMix
-TNS under varied scenarios—ranging from single to continual and mixed domain sh
ifts, particularly excelling with temporally correlated test data and corrupted
non-i.i.d. real-world streams. This adaptability is maintained even with very sm
all batch sizes or single instances. Our results highlight UnMix-TNS's capacity
to markedly enhance stability and performance across various benchmarks. Our cod
e is publicly available at https://github.com/devavratTomar/unmixtns.
****
Zijian Liu,Zhengyuan Zhou
Revisiting the Last-Iterate Convergence of Stochastic Gradient Methods
In the past several years, the last-iterate convergence of the Stochastic Gradie
nt Descent (SGD) algorithm has triggered people's interest due to its good perfo
rmance in practice but lack of theoretical understanding. For Lipschitz convex f
unctions, different works have established the optimal $O(\log(1/\delta)\log T/\

sqrt{T})$ or $O(\sqrt{\log(1/\delta)/T})$ high-probability convergence rates for the final iterate, where T is the time horizon and δ is the failure probability. However, to prove these bounds, all the existing works are either limited to compact domains or require almost surely bounded noises. It is natural to ask whether the last iterate of SGD can still guarantee the optimal convergence rate but without these two restrictive assumptions. Besides this important question, there are still lots of theoretical problems lacking an answer. For example, compared with the last-iterate convergence of SGD for non-smooth problems, only few results for smooth optimization have yet been developed. Additionally, the existing results are all limited to a non-composite objective and the standard Euclidean norm. It still remains unclear whether the last-iterate convergence can be provably extended to wider composite optimization and non-Euclidean norms. In this work, to address the issues mentioned above, we revisit the last-iterate convergence of stochastic gradient methods and provide the first unified way to prove the convergence rates both in expectation and in high probability to accommodate general domains, composite objectives, non-Euclidean norms, Lipschitz conditions, smoothness, and (strong) convexity simultaneously.
****
Max Zimmer,Christoph Spiegel,Sebastian Pokutta
Sparse Model Soups: A Recipe for Improved Pruning via Model Averaging
Neural networks can be significantly compressed by pruning, yielding sparse models with reduced storage and computational demands while preserving predictive performance. Model soups (Wortsman et al., 2022) enhance generalization and out-of-distribution (OOD) performance by averaging the parameters of multiple models into a single one, without increasing inference time. However, achieving both sparsity and parameter averaging is challenging as averaging arbitrary sparse models reduces the overall sparsity due to differing sparse connectivities. This work addresses these challenges by demonstrating that exploring a single retraining phase of Iterative Magnitude Pruning (IMP) with varied hyperparameter configurations such as batch ordering or weight decay yields models suitable for averaging, sharing identical sparse connectivity by design. Averaging these models significantly enhances generalization and OOD performance over their individual counterparts. Building on this, we introduce Sparse Model Soups (SMS), a novel method for merging sparse models by initiating each prune-retrain cycle with the averaged model from the previous phase. SMS preserves sparsity, exploits sparse network benefits, is modular and fully parallelizable, and substantially improves IMP's performance. We further demonstrate that SMS can be adapted to enhance state-of-the-art pruning-during-training approaches.
****
Sina Khajehabdollahi,Roxana Zeraati,Emmanouil Giannakakis,Tim Jakob Schäfer,Georg Martius,Anna Levina
Emergent mechanisms for long timescales depend on training curriculum and affect performance in memory tasks
Recurrent neural networks (RNNs) in the brain and \emph{in silico} excel at solving tasks with intricate temporal dependencies.
Long timescales required for solving such tasks can arise from properties of individual neurons (single-neuron timescale, τ, e.g., membrane time constant in biological neurons) or recurrent interactions among them (network-mediated timescale, $\tau_\textrm{\small{net}}$).
However, the contribution of each mechanism for optimally solving memory-dependent tasks remains poorly understood. Here, we train RNNs to solve N-parity and N-delayed match-to-sample tasks with increasing memory requirements controlled by N, by simultaneously optimizing recurrent weights and τs. We find that RNNs develop longer timescales with increasing N, but depending on the learning objective, they use different mechanisms. Two distinct curricula define learning objectives: sequential learning of a single-N (single-head) or simultaneous learning of multiple Ns (multi-head). Single-head networks increase their τ with N and can solve large-N tasks, but suffer from catastrophic forgetting. However, multi-head networks, which are explicitly required to hold multiple concurrent memories, keep τ constant and develop longer timescales thro

ugh recurrent connectivity. We show that the multi-head curriculum increases training speed and stability to perturbations, and allows generalization to tasks beyond the training set.
This curriculum also significantly improves training GRUs and LSTMs for large-N tasks.
Our results suggest that adapting timescales to task requirements via recurrent interactions allows learning more complex objectives and improves the RNN's performance.
****
Peng Xu,Wei Ping,Xianchao Wu,Lawrence McAfee,Chen Zhu,Zihan Liu,Sandeep Subramanian,Evelina Bakhturina,Mohammad Shoeybi,Bryan Catanzaro
Retrieval meets Long Context Large Language Models
Extending the context window of large language models (LLMs) is getting popular recently, while the solution of augmenting LLMs with retrieval has existed for years. The natural questions are: i) Retrieval-augmentation versus long context window, which one is better for downstream tasks? ii) Can both methods be combined to get the best of both worlds? In this work, we answer these questions by studying both solutions using two state-of-the-art pretrained LLMs, i.e., a proprietary 43B GPT and Llama2-70B. Perhaps surprisingly, we find that LLM with 4K context window using simple retrieval-augmentation at generation can achieve comparable performance to finetuned LLM with 16K context window via positional interpolation on long context tasks, while taking much less computation. More importantly, we demonstrate that retrieval can significantly improve the performance of LLMs regardless of their extended context window sizes. Our best model, retrieval-augmented Llama2-70B with 32K context window, outperforms GPT-3.5-turbo-16k and Davinci003 in terms of average score on nine long context tasks including question answering, query-based summarization, and in-context few-shot learning tasks. It also outperforms its non-retrieval Llama2-70B-32k baseline by a margin, while being much faster at generation. Our study provides general insights on the choice of retrieval-augmentation versus long context extension of LLM for practitioners.
****
Han Guo,Philip Greengard,Eric Xing,Yoon Kim
LQ-LoRA: Low-rank plus Quantized Matrix Decomposition for Efficient Language Model Finetuning
We propose a simple approach for memory-efficient adaptation of pretrained language models. Our approach uses an iterative algorithm to decompose each pretrained matrix into a high-precision low-rank component and a memory-efficient quantized component. During finetuning, the quantized component remains fixed and only the low-rank component is updated. We present an integer linear programming formulation of the quantization component which enables dynamic configuration of quantization parameters (e.g., bit-width, block size) for each matrix given an overall target memory budget. We further explore a data-aware version of the algorithm which uses an approximation of the Fisher information matrix to weight the reconstruction objective during matrix decomposition. Experiments on finetuning RoBERTa and LLaMA-2 (7B and 70B) demonstrate that our low-rank plus quantized matrix decomposition approach (LQ-LoRA) outperforms strong QLoRA and GPTQ-LoRA baselines and enables aggressive quantization to sub-3 bits with only minor performance degradations. When finetuned on a language modeling calibration dataset, LQ-LoRA can also be used for model compression; in this setting our 2.75-bit LLaMA-2-70B model (which has 2.85 bits on average when including the low-rank components and requires 27GB of GPU memory) performs respectably compared to the 16-bit baseline.
****
Zecheng Hao,Tong Bu,Xinyu Shi,Zihan Huang,Zhaofei Yu,Tiejun Huang
Threaten Spiking Neural Networks through Combining Rate and Temporal Information
Spiking Neural Networks (SNNs) have received widespread attention in academic communities due to their superior spatio-temporal processing capabilities and energy-efficient characteristics. With further in-depth application in various fields, the vulnerability of SNNs under adversarial attack has become a focus of conc

ern.
In this paper, we draw inspiration from two mainstream learning algorithms of SNNs and observe that SNN models reserve both rate and temporal information. To better understand the capabilities of these two types of information, we conduct a quantitative analysis separately for each. In addition, we note that the retention degree of temporal information is related to the parameters and input settings of spiking neurons. Building on these insights, we propose a hybrid adversarial attack based on rate and temporal information (HART), which allows for dynamic adjustment of the rate and temporal attributes. Experimental results demonstrate that compared to previous works, HART attack can achieve significant superiority under different attack scenarios, data types, network architecture, time-steps, and model hyper-parameters. These findings call for further exploration into how both types of information can be effectively utilized to enhance the reliability of SNNs. Code is available at https://github.com/hzc1208/HART_Attack.
****
Yogesh Verma,Markus Heinonen,Vikas Garg
ClimODE: Climate Forecasting With Physics-informed Neural ODEs
Climate prediction traditionally relies on complex numerical simulations of atmospheric physics. Deep learning approaches, such as transformers, have recently challenged the simulation paradigm with complex network forecasts. However, they often act as data-driven black-box models that neglect the underlying physics and lack uncertainty quantification. We address these limitations with ClimODE, a spatiotemporal continuous-time process that implements a key principle of advection from statistical mechanics, namely, weather changes due to a spatial movement of quantities over time. ClimODE models precise weather evolution with value-conserving dynamics, learning global weather transport as a neural flow, which also enables estimating the uncertainty in predictions. Our approach outperforms existing data-driven methods in global and regional forecasting with an order of magnitude smaller parameterization, establishing a new state of the art.
****
Arjun Ashok,Étienne Marcotte,Valentina Zantedeschi,Nicolas Chapados,Alexandre Drouin
TACTiS-2: Better, Faster, Simpler Attentional Copulas for Multivariate Time Series
We introduce a new model for multivariate probabilistic time series prediction, designed to flexibly address a range of tasks including forecasting, interpolation, and their combinations. Building on copula theory, we propose a simplified objective for the recently-introduced transformer-based attentional copulas (TACTiS), wherein the number of distributional parameters now scales linearly with the number of variables instead of factorially. The new objective requires the introduction of a training curriculum, which goes hand-in-hand with necessary changes to the original architecture. We show that the resulting model has significantly better training dynamics and achieves state-of-the-art performance across diverse real-world forecasting tasks, while maintaining the flexibility of prior work, such as seamless handling of unaligned and unevenly-sampled time series. Code is made available at https://github.com/ServiceNow/TACTiS.
****
Harshit Sikchi,Qinqing Zheng,Amy Zhang,Scott Niekum
Dual RL: Unification and New Methods for Reinforcement and Imitation Learning
The goal of reinforcement learning (RL) is to find a policy that maximizes the expected cumulative return. It has been shown that this objective can be represented as an optimization problem of state-action visitation distribution under linear constraints. The dual problem of this formulation, which we refer to as *dual RL*, is unconstrained and easier to optimize. In this work, we first cast several state-of-the-art offline RL and offline imitation learning (IL) algorithms as instances of dual RL approaches with shared structures. Such unification allows us to identify the root cause of the shortcomings of prior methods. For offline IL, our analysis shows that prior methods are based on a restrictive coverage assumption that greatly limits their performance in practice. To fix this limita

tion, we propose a new discriminator-free method ReCOIL that learns to imitate from arbitrary off-policy data to obtain near-expert performance. For offline RL, our analysis frames a recent offline RL method XQL in the dual framework, and we further propose a new method f-DVL that provides alternative choices to the Gumbel regression loss that fixes the known training instability issue of XQL. The performance improvements by both of our proposed methods, ReCOIL and f-DVL, in IL and RL are validated on an extensive suite of simulated robot locomotion and manipulation tasks.
****

Chenhao Li,Elijah Stanger-Jones,Steve Heim,Sang bae Kim
FLD: Fourier Latent Dynamics for Structured Motion Representation and Learning
Motion trajectories offer reliable references for physics-based motion learning but suffer from sparsity, particularly in regions that lack sufficient data coverage. To address this challenge, we introduce a self-supervised, structured representation and generation method that extracts spatial-temporal relationships in periodic or quasi-periodic motions. The motion dynamics in a continuously parameterized latent space enable our method to enhance the interpolation and generalization capabilities of motion learning algorithms. The motion learning controller, informed by the motion parameterization, operates online tracking of a wide range of motions, including targets unseen during training. With a fallback mechanism, the controller dynamically adapts its tracking strategy and automatically resorts to safe action execution when a potentially risky target is proposed. By leveraging the identified spatial-temporal structure, our work opens new possibilities for future advancements in general motion representation and learning algorithms.
****

Joseph Early,Gavin Cheung,Kurt Cutajar,Hanting Xie,Jas Kandola,Niall Twomey
Inherently Interpretable Time Series Classification via Multiple Instance Learning
Conventional Time Series Classification (TSC) methods are often black boxes that obscure inherent interpretation of their decision-making processes. In this work, we leverage Multiple Instance Learning (MIL) to overcome this issue, and propose a new framework called MILLET: Multiple Instance Learning for Locally Explainable Time series classification. We apply MILLET to existing deep learning TSC models and show how they become inherently interpretable without compromising (and in some cases, even improving) predictive performance. We evaluate MILLET on 85 UCR TSC datasets and also present a novel synthetic dataset that is specially designed to facilitate interpretability evaluation. On these datasets, we show MILLET produces sparse explanations quickly that are of higher quality than other well-known interpretability methods. To the best of our knowledge, our work with MILLET is the first to develop general MIL methods for TSC and apply them to an extensive variety of domains.
****

Erik Jones,Hamid Palangi,Clarisse Simões Ribeiro,Varun Chandrasekaran,Subhabrata Mukherjee,Arindam Mitra,Ahmed Hassan Awadallah,Ece Kamar
Teaching Language Models to Hallucinate Less with Synthetic Tasks
Large language models (LLMs) frequently hallucinate on abstractive summarization tasks such as document-based question-answering, meeting summarization, and clinical report generation, even though all necessary information is included in context. However, optimizing to make LLMs hallucinate less is challenging, as hallucination is hard to efficiently, cheaply, and reliably evaluate at each optimization step. In this work, we show that reducing hallucination on a _synthetic task_ can also reduce hallucination on real-world downstream tasks. Our method, SynTra, first designs a synthetic task where hallucinations are easy to elicit and measure. It next optimizes the LLM's system message via prefix tuning on the synthetic task, then uses the system message on realistic, hard-to-optimize tasks. Across three realistic abstractive summarization tasks, we reduce hallucination for two 13B-parameter LLMs using supervision signal from only a synthetic retrieval task. We also find that optimizing the system message rather than the model weights can be critical; fine-tuning the entire model on the synthetic task can

counterintuitively _increase_ hallucination. Overall, SynTra demonstrates that the extra flexibility of working with synthetic data can help mitigate undesired behaviors in practice.
****
Vincent Grari,Thibault Laugel,Tatsunori Hashimoto,sylvain lamprier,Marcin Detyniecki
On the Fairness ROAD: Robust Optimization for Adversarial Debiasing
In the field of algorithmic fairness, significant attention has been put on group fairness criteria, such as Demographic Parity and Equalized Odds. Nevertheless, these objectives, measured as global averages, have raised concerns about persistent local disparities between sensitive groups. In this work, we address the problem of local fairness, which ensures that the predictor is unbiased not only in terms of expectations over the whole population, but also within any subregion of the feature space, unknown at training time. To enforce this objective, we introduce ROAD, a novel approach that leverages the Distributionally Robust Optimization (DRO) framework within a fair adversarial learning objective, where an adversary tries to infer the sensitive attribute from the predictions. Using an instance-level re-weighting strategy, ROAD is designed to prioritize inputs that are likely to be locally unfair, i.e. where the adversary faces the least difficulty in reconstructing the sensitive attribute. Numerical experiments demonstrate the effectiveness of our method: it achieves Pareto dominance with respect to local fairness and accuracy for a given global fairness level across three standard datasets, and also enhances fairness generalization under distribution shift.
****
xingbin liu,Jinghao Zhou,Tao Kong,Xianming Lin,Rongrong Ji
Exploring Target Representations for Masked Autoencoders
Masked autoencoders have become popular training paradigms for self-supervised visual representation learning. These models randomly mask a portion of the input and reconstruct the masked portion according to assigned target representations. In this paper, we show that a careful choice of the target representation is unnecessary for learning good visual representation since different targets tend to derive similarly behaved models. Driven by this observation, we propose a multi-stage masked distillation pipeline and use a randomly initialized model as the teacher, enabling us to effectively train high-capacity models without any effort to carefully design the target representation.
On various downstream tasks, the proposed method to perform masked knowledge distillation with bootstrapped teachers (dbot) outperforms previous self-supervised methods by nontrivial margins. We hope our findings, as well as the proposed method, could motivate people to rethink the roles of target representations in pre-training masked autoencoders.
****
Zhiwei Li,Guodong Long,Tianyi Zhou
Federated Recommendation with Additive Personalization
Building recommendation systems via federated learning (FL) is a new emerging challenge for next-generation Internet service. Existing FL models share item embedding across clients while keeping the user embedding private and local on the client side. However, identical item embedding cannot capture users' individual differences in perceiving the same item and may lead to poor personalization. Moreover, dense item embedding in FL results in expensive communication costs and latency. To address these challenges, we propose Federated Recommendation withAdditive Personalization (FedRAP), which learns a global view of items via FL and a personalized view locally on each user. FedRAP encourages a sparse global view to save FL's communication cost and enforces the two views to be complementary via two regularizers. We propose an effective curriculum to learn the local and global views progressively with increasing regularization weights. To produce recommendations for a user, FedRAP adds the two views together to obtain a personalized item embedding. FedRAP achieves the best performance in FL setting on multiple benchmarks. It outperforms recent federated recommendation methods and several ablation study baselines. Our code is available at https://github.com/mtics/F

edRAP.
****
Saba Ghaffari,Ehsan Saleh,Alex Schwing,Yu-Xiong Wang,Martin D. Burke,Saurabh Sinha
Robust Model-Based Optimization for Challenging Fitness Landscapes
Protein design, a grand challenge of the day, involves optimization on a fitness landscape, and leading methods adopt a model-based approach where a model is trained on a training set (protein sequences and fitness) and proposes candidates to explore next. These methods are challenged by sparsity of high-fitness samples in the training set, a problem that has been in the literature. A less recognized but equally important problem stems from the distribution of training samples in the design space: leading methods are not designed for scenarios where the desired optimum is in a region that is not only poorly represented in training data, but also relatively far from the highly represented low-fitness regions. We show that this problem of “separation” in the design space is a significant bottleneck in existing model-based optimization tools and propose a new approach that uses a novel VAE as its search model to overcome the problem. We demonstrate its advantage over prior methods in robustly finding improved samples, regardless of the imbalance and separation between low- and high-fitness samples. Our comprehensive benchmark on real and semi-synthetic protein datasets as well as solution design for physics-informed neural networks, showcases the generality of our approach in discrete and continuous design spaces. Our implementation is available at https://github.com/sabagh1994/PGVAE.
****
Alexandru Meterez,Amir Joudaki,Francesco Orabona,Alexander Immer,Gunnar Ratsch,Hadi Daneshmand
Towards Training Without Depth Limits: Batch Normalization Without Gradient Explosion
Normalization layers are one of the key building blocks for deep neural networks. Several theoretical studies have shown that batch normalization improves the signal propagation, by avoiding the representations from becoming collinear across the layers. However, results on mean-field theory of batch normalization also conclude that this benefit comes at the expense of exploding gradients in depth. Motivated by these two aspects of batch normalization, in this study we pose the following question:
Can a batch-normalized network keep the optimal signal propagation properties, but avoid exploding gradients? We answer this question in the affirmative by giving a particular construction of an *MLP with linear activations* and batch-normalization that provably has *bounded gradients* at any depth. Based on Weingarten calculus, we develop a rigorous and non-asymptotic theory for this constructed MLP that gives a precise characterization of forward signal propagation, while proving that gradients remain bounded for linearly independent input samples, which holds in most practical settings. Inspired by our theory, we also design an activation shaping scheme that empirically achieves the same properties for non-linear activations.
****
Anton Bushuiev,Roman Bushuiev,Petr Kouba,Anatolii Filkin,Marketa Gabrielova,Michal Gabriel,Jiri Sedlar,Tomas Pluskal,Jiri Damborsky,Stanislav Mazurenko,Josef Sivic
Learning to design protein-protein interactions with enhanced generalization
Discovering mutations enhancing protein-protein interactions (PPIs) is critical for advancing biomedical research and developing improved therapeutics. While machine learning approaches have substantially advanced the field, they often struggle to generalize beyond training data in practical scenarios. The contributions of this work are three-fold. First, we construct PPIRef, the largest and non-redundant dataset of 3D protein-protein interactions, enabling effective large-scale learning. Second, we leverage the PPIRef dataset to pre-train PPIformer, a new SE(3)-equivariant model generalizing across diverse protein-binder variants. We fine-tune PPIformer to predict effects of mutations on protein-protein interactions via a thermodynamically motivated adjustment of the pre-training loss fun

ction. Finally, we demonstrate the enhanced generalization of our new PPIformer approach by outperforming other state-of-the-art methods on new, non-leaking splits of standard labeled PPI mutational data and independent case studies optimizing a human antibody against SARS-CoV-2 and increasing the thrombolytic activity of staphylokinase.
****
Tianqi Liu,Yao Zhao,Rishabh Joshi,Misha Khalman,Mohammad Saleh,Peter J Liu,Jialu Liu
Statistical Rejection Sampling Improves Preference Optimization
Improving the alignment of language models with human preferences remains an active research challenge. Previous approaches have primarily utilized online Reinforcement Learning from Human Feedback (RLHF). Recently, offline methods such as Sequence Likelihood Calibration (SLiC) and Direct Preference Optimization (DPO) have emerged as attractive alternatives, offering improvements in stability and scalability while maintaining competitive performance. SLiC refines its loss function using sequence pairs sampled from a supervised fine-tuned (SFT) policy, while DPO directly optimizes language models based on preference data, foregoing the need for a separate reward model. However, the maximum likelihood estimator (MLE) of the target optimal policy requires labeled preference pairs sampled from that policy. The absence of a reward model in DPO constrains its ability to sample preference pairs from the optimal policy. Meanwhile, SLiC can only sample preference pairs from the SFT policy. To address these limitations, we introduce a novel approach called Statistical Rejection Sampling Optimization (RSO) designed to source preference data from the target optimal policy using rejection sampling, enabling a more accurate estimation of the optimal policy. We also propose a unified framework that enhances the loss functions used in both SLiC and DPO from a preference modeling standpoint. Through extensive experiments across diverse tasks, we demonstrate that RSO consistently outperforms both SLiC and DPO as evaluated by both Large Language Models (LLMs) and human raters.
****
Qingru Zhang,Chandan Singh,Liyuan Liu,Xiaodong Liu,Bin Yu,Jianfeng Gao,Tuo Zhao
Tell Your Model Where to Attend: Post-hoc Attention Steering for LLMs
In human-written articles, we often leverage the subtleties of text style, such as bold and italics, to guide the attention of readers. These textual emphases are vital for the readers to grasp the conveyed information. When interacting with large language models (LLMs), we have a similar need -- steering the model to pay closer attention to user-specified information, e.g., an instruction. Existing methods, however, are constrained to process plain text and do not support such a mechanism. This motivates us to introduce PASTA -- Post-hoc Attention STeering Approach, a method that allows LLMs to read text with user-specified emphasis marks. To this end, PASTA identifies a small subset of attention heads and applies precise attention reweighting on them, directing the model attention to user-specified parts. Like prompting, PASTA is applied at inference time and does not require changing any model parameters. Experiments demonstrate that PASTA can substantially enhance an LLM's ability to follow user instructions or integrate new knowledge from user inputs, leading to a significant performance improvement on a variety of tasks, e.g., an average accuracy improvement of 22\% for LLAMA-7B. Our code is publicly available at https://github.com/QingruZhang/PASTA .
****
Christopher A. Choquette-Choo,Arun Ganesh,Thomas Steinke,Abhradeep Guha Thakurta
Privacy Amplification for Matrix Mechanisms
Privacy amplification exploits randomness in data selection to provide tighter differential privacy (DP) guarantees. This analysis is key to DP-SGD's success in machine learning (ML), but, is not readily applicable to the newer state-of-the-art (SOTA) algorithms. This is because these algorithms, known as DP-FTRL, use the matrix mechanism to add correlated noise instead of independent noise as in DP-SGD.

In this paper, we propose "MMCC'' (matrix mechanism conditional composition), the first algorithm to analyze privacy amplification via sampling for any generic

matrix mechanism. MMCC is nearly tight in that it approaches a lower bound as $\epsilon\to0$.
To analyze correlated outputs in MMCC, we prove that they can be analyzed as if they were independent, by conditioning them on prior outputs. Our "conditional composition theorem'' has broad utility: we use it to show that the noise added to binary-tree-DP-FTRL can asymptotically match the noise added to DP-SGD with amplification. Our algorithm also has practical empirical utility. We show that amplification leads to significant improvement in the privacy/utility trade-offs for DP-FTRL style algorithms for standard benchmark tasks.
****
Xue Jiang,Feng Liu,Zhen Fang,Hong Chen,Tongliang Liu,Feng Zheng,Bo Han
Negative Label Guided OOD Detection with Pretrained Vision-Language Models
Out-of-distribution (OOD) detection aims at identifying samples from unknown classes, playing a crucial role in trustworthy models against errors on unexpected inputs.
Extensive research has been dedicated to exploring OOD detection in the vision modality.
{Vision-language models (VLMs) can leverage both textual and visual information for various multi-modal applications, whereas few OOD detection methods take into account information from the text modality.
In this paper, we propose a novel post hoc OOD detection method, called NegLabel, which takes a vast number of negative labels from extensive corpus databases. We design a novel scheme for the OOD score collaborated with negative labels.
Theoretical analysis helps to understand the mechanism of negative labels. Extensive experiments demonstrate that our method NegLabel achieves state-of-the-art performance on various OOD detection benchmarks and generalizes well on multiple VLM architectures. Furthermore, our method NegLabel exhibits remarkable robustness against diverse domain shifts. The codes are available at https://github.com/tmlr-group/NegLabel.
****
Zhiqing Sun,Yikang Shen,Hongxin Zhang,Qinhong Zhou,Zhenfang Chen,David Daniel Cox,Yiming Yang,Chuang Gan
SALMON: Self-Alignment with Principle-Following Reward Models
Supervised Fine-Tuning (SFT) on human demonstrations combined with Reinforcement Learning from Human Feedback (RLHF) constitutes a powerful alignment paradigm for Large Language Model (LLM) AI-assistant agents. However, a significant limitation of this approach is its substantial dependency on high-quality human annotations, making its broader application to intricate tasks challenging due to difficulties in obtaining consistent response demonstrations and task-specific response preferences. To address this issue, we present a novel alignment paradigm in this paper, termed SALMON (Self-ALignMent with principle-fOllowiNg reward models). This paradigm offers the ability to align base language models with minimal human supervision, using only a select set of human-defined principles, yet achieves superior performance. Central to our approach is a principle-following reward model. Trained on synthetic preference data, this reward model can generate reward scores based on arbitrary human-defined principles. Therefore, during the RL training phase, by merely adjusting these principles, we gain full control over the preferences of the reward model, subsequently influencing the behavior of the RL-trained policy model, and eliminating the traditional reliance on exhaustive online human preference collection. Applying our method to the LLaMA-2-70b base language model, we developed an AI assistant named Dromedary-2. With only 6 exemplars for in-context learning and 31 human-defined principles, Dromedary-2 significantly surpasses the performance of several state-of-the-art AI systems, including LLaMA-2-Chat-70b, on various benchmark datasets. We have open-sourced the code and model weights to encourage further research into aligning LLM-based AI agents with enhanced supervision efficiency, improved controllability, and scalable oversight.
****
Wenhan Cao,Wei Pan
Impact of Computation in Integral Reinforcement Learning for Continuous-Time Con

trol
Integral reinforcement learning (IntRL) demands the precise computation of the utility function's integral at its policy evaluation (PEV) stage. This is achieved through quadrature rules, which are weighted sums of utility functions evaluated from state samples obtained in discrete time. Our research reveals a critical yet underexplored phenomenon: the choice of the computational method -- in this case, the quadrature rule -- can significantly impact control performance. This impact is traced back to the fact that computational errors introduced in the PEV stage can affect the policy iteration's convergence behavior, which in turn affects the learned controller. To elucidate how computation impacts control, we draw a parallel between IntRL's policy iteration and Newton's method applied to the Hamilton-Jacobi-Bellman equation. In this light, computational error in PEV manifests as an extra error term in each iteration of Newton's method, with its upper bound proportional to the computational error. Further, we demonstrate that when the utility function resides in a reproducing kernel Hilbert space (RKHS), the optimal quadrature is achievable by employing Bayesian quadrature with the RKHS-inducing kernel function. We prove that the local convergence rates for IntRL using the trapezoidal rule and Bayesian quadrature with a Matérn kernel to be $O(N^{-2})$ and $O(N^{-b})$, where N is the number of evenly-spaced samples and b is the Matérn kernel's smoothness parameter. These theoretical findings are finally validated by two canonical control tasks.
****
Wei Yao,Chengming Yu,Shangzhi Zeng,Jin Zhang
Constrained Bi-Level Optimization: Proximal Lagrangian Value Function Approach and Hessian-free Algorithm
This paper presents a new approach and algorithm for solving a class of constrained Bi-Level Optimization (BLO) problems in which the lower-level problem involves constraints coupling both upper-level and lower-level variables. Such problems have recently gained significant attention due to their broad applicability in machine learning. However, conventional gradient-based methods unavoidably rely on computationally intensive calculations related to the Hessian matrix. To address this challenge, we devise a smooth proximal Lagrangian value function to handle the constrained lower-level problem. Utilizing this construct, we introduce a single-level reformulation for constrained BLOs that transforms the original BLO problem into an equivalent optimization problem with smooth constraints. Enabled by this reformulation, we develop a Hessian-free gradient-based algorithm—termed proximal Lagrangian Value function-based Hessian-free Bi-level Algorithm (LV-HBA)—that is straightforward to implement in a single loop manner. Consequently, LV-HBA is especially well-suited for machine learning applications. Furthermore, we offer non-asymptotic convergence analysis for LV-HBA, eliminating the need for traditional strong convexity assumptions for the lower-level problem while also being capable of accommodating non-singleton scenarios. Empirical results substantiate the algorithm's superior practical performance.
****
Lorenzo Loconte,Aleksanteri Mikulus Sladek,Stefan Mengel,Martin Trapp,Arno Solin,Nicolas Gillis,Antonio Vergari
Subtractive Mixture Models via Squaring: Representation and Learning
Mixture models are traditionally represented and learned by adding several distributions as components. Allowing mixtures to subtract probability mass or density can drastically reduce the number of components needed to model complex distributions. However, learning such subtractive mixtures while ensuring they still encode a non-negative function is challenging. We investigate how to learn and perform inference on deep subtractive mixtures by squaring them. We do this in the framework of probabilistic circuits, which enable us to represent tensorized mixtures and generalize several other subtractive models. We theoretically prove that the class of squared circuits allowing subtractions can be exponentially more expressive than traditional additive mixtures; and, we empirically show this increased expressiveness on a series of real-world distribution estimation tasks.
****
Sihang Li,Zhiyuan Liu,Yanchen Luo,Xiang Wang,Xiangnan He,Kenji Kawaguchi,Tat-Sen

g Chua,Qi Tian
Towards 3D Molecule-Text Interpretation in Language Models
Language Models (LMs) have greatly influenced diverse domains. However, their in
herent limitation in comprehending 3D molecular structures has considerably cons
trained their potential in the biomolecular domain. To bridge this gap, we focus
 on 3D molecule-text interpretation, and propose 3D-MoLM: 3D-Molecular Language
Modeling. Specifically, 3D-MoLM enables an LM to interpret and analyze 3D molecu
les by equipping the LM with a 3D molecular encoder. This integration is achieve
d by a 3D molecule-text projector, bridging the 3D molecular encoder’s represent
ation space and the LM’s input space. Moreover, to enhance 3D■MoLM’s ability of
cross-modal molecular understanding and instruction following, we meticulously c
urated a 3D molecule-centric instruction tuning dataset – 3D-MoIT. Through 3D mo
lecule-text alignment and 3D molecule-centric instruction tuning, 3D-MoLM establ
ishes an integration of 3D molecular encoder and LM. It significantly surpasses
existing baselines on downstream tasks, including molecule■text retrieval, molec
ule captioning, and more challenging open-text molecular QA tasks, especially fo
cusing on 3D-dependent properties. We will release our codes and datasets at htt
ps://github.com/lsh0520/3D-MoLM.
****
Christopher A. Choquette-Choo,Krishnamurthy Dj Dvijotham,Krishna Pillutla,Arun G
anesh,Thomas Steinke,Abhradeep Guha Thakurta
Correlated Noise Provably Beats Independent Noise for Differentially Private Lea
rning
Differentially private learning algorithms inject noise into the learning proces
s. While the most common private learning algorithm, DP-SGD, adds independent Ga
ussian noise in each iteration, recent work on matrix factorization mechanisms h
as shown empirically that introducing correlations in the noise can greatly impr
ove their utility. We characterize the asymptotic learning utility for any choic
e of the correlation function, giving precise analytical bounds for linear regre
ssion and as the solution to a convex program for general convex functions. We s
how, using these bounds, how correlated noise provably improves upon vanilla DP-
SGD as a function of problem parameters such as the effective dimension and cond
ition number. Moreover, our analytical expression for the near-optimal correlati
on function circumvents the cubic complexity of the semi-definite program used t
o optimize the noise correlation matrix in previous work. We validate these theo
retical results with experiments on private deep learning. Our work matches or o
utperforms prior work while being efficient both in terms of computation and mem
ory.
****
Nuoya Xiong,Lijun Ding,Simon Shaolei Du
How Over-Parameterization Slows Down Gradient Descent in Matrix Sensing: The Cur
ses of Symmetry and Initialization
This paper rigorously shows how over-parameterization dramatically changes the c
onvergence behaviors of gradient descent (GD) for the matrix sensing problem, wh
ere the goal is to recover an unknown low-rank ground-truth matrix from near-iso
tropic linear measurements.
First, we consider the symmetric setting with the symmetric parameterization whe
re $M^* \in \mathbb{R}^{n \times n}$ is a positive semi-definite unknown matrix
of rank $r \ll n$, and one uses a symmetric parameterization $XX^\top$ to learn
M^*. Here $X \in \mathbb{R}^{n \times k}$ with $k > r$ is the factor matrix. W
e give a novel $\Omega\left(1/T^2\right)$ lower bound of randomly initialized GD
 for the over-parameterized case ($k >r$) where T is the number of iterations.
 This is in stark contrast to the exact-parameterization scenario ($k=r$) where
the convergence rate is $\exp\left(-\Omega\left(T\right)\right)$. Next, we study
 asymmetric setting where $M^* \in \mathbb{R}^{n_1 \times n_2}$ is the unknown m
atrix of rank $r \ll \min\{n_1,n_2\}$, and one uses an asymmetric parameterizati
on $FG^\top$ to learn M^* where $F \in \mathbb{R}^{n_1 \times k}$ and $G \in \
mathbb{R}^{n_2 \times k}$. We give the first global exact convergence result of
randomly initialized GD for the exact-parameterization case ($k=r$) with an $\ex
p\left(-\Omega\left(T\right)\right)$ rate. Furthermore, we give the first global

exact convergence result for the over-parameterization case ($k>r$) with an $\exp\left(-\Omega\left(\alpha^2 T\right)\right)$ rate where α is the initialization scale. This linear convergence result in the over-parameterization case is especially significant because one can apply the asymmetric parameterization to the symmetric setting to speed up from $\Omega\left(1/T^2\right)$ to linear convergence. Therefore, we identify a surprising phenomenon: asymmetric parameterization can exponentially speed up convergence. Equally surprising is our analysis that highlights the importance of imbalance between F and G. This is in sharp contrast to prior works which emphasize balance. We further give an example showing the dependency on α in the convergence rate is unavoidable in the worst case. On the other hand, we propose a novel method that only modifies one step of GD and obtains a convergence rate independent of α, recovering the rate in the exact-parameterization case. We provide empirical studies to verify our theoretical findings.
****
Mang Ning,Mingxiao Li,Jianlin Su,Albert Ali Salah,Itir Onal Ertugrul
Elucidating the Exposure Bias in Diffusion Models
Diffusion models have demonstrated impressive generative capabilities, but their exposure bias problem, described as the input mismatch between training and sampling, lacks in-depth exploration. In this paper, we investigate the exposure bias problem in diffusion models by first analytically modelling the sampling distribution, based on which we then attribute the prediction error at each sampling step as the root cause of the exposure bias issue. Furthermore, we discuss potential solutions to this issue and propose an intuitive metric for it. Along with the elucidation of exposure bias, we propose a simple, yet effective, training-free method called Epsilon Scaling to alleviate the exposure bias. We show that Epsilon Scaling explicitly moves the sampling trajectory closer to the vector field learned in the training phase by scaling down the network output, mitigating the input mismatch between training and sampling. Experiments on various diffusion frameworks (ADM, DDIM, EDM, LDM, DiT, PFGM++) verify the effectiveness of our method. Remarkably, our ADM-ES, as a state-of-the-art stochastic sampler, obtains 2.17 FID on CIFAR-10 under 100-step unconditional generation. The code is at https://github.com/forever208/ADM-ES
****
Huayu Chen,Cheng Lu,Zhengyi Wang,Hang Su,Jun Zhu
Score Regularized Policy Optimization through Diffusion Behavior
Recent developments in offline reinforcement learning have uncovered the immense potential of diffusion modeling, which excels at representing heterogeneous behavior policies. However, sampling from diffusion policies is considerably slow because it necessitates tens to hundreds of iterative inference steps for one action. To address this issue, we propose to extract an efficient deterministic inference policy from critic models and pretrained diffusion behavior models, leveraging the latter to directly regularize the policy gradient with the behavior distribution’s score function during optimization. Our method enjoys powerful generative capabilities of diffusion modeling while completely circumventing the computationally intensive and time-consuming diffusion sampling scheme, both during training and evaluation. Extensive results on D4RL tasks show that our method boosts action sampling speed by more than 25 times compared with various leading diffusion-based methods in locomotion tasks, while still maintaining state-of-the-art performance.
****
Xinyu Yang,Weixin Liang,James Zou
Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on HuggingFace
Advances in machine learning are closely tied to the creation of datasets. While data documentation is widely recognized as essential to the reliability, reproducibility, and transparency of ML, we lack a systematic empirical understanding of current dataset documentation practices. To shed light on this question, here we take Hugging Face - one of the largest platforms for sharing and collaborating on ML models and datasets - as a prominent case study. By analyzing all 7,43

3 dataset documentation on Hugging Face, our investigation provides an overview of the Hugging Face dataset ecosystem and insights into dataset documentation practices, yielding 5 main findings: (1) The dataset card completion rate shows marked heterogeneity correlated with dataset popularity: While 86.0\% of the top 100 downloaded dataset cards fill out all sections suggested by Hugging Face community, only 7.9\% of dataset cards with no downloads complete all these sections. (2) A granular examination of each section within the dataset card reveals that the practitioners seem to prioritize Dataset Description and Dataset Structure sections, accounting for 36.2\% and 33.6\% of the total card length, respectively, for the most downloaded datasets. In contrast, the Considerations for Using the Data section receives the lowest proportion of content, accounting for just 2.1\% of the text. (3) By analyzing the subsections within each section and utilizing topic modeling to identify key topics, we uncover what is discussed in each section, and underscore significant themes encompassing both technical and social impacts, as well as limitations within the Considerations for Using the Data section. (4) Our findings also highlight the need for improved accessibility and reproducibility of datasets in the Usage sections. (5) In addition, our human annotation evaluation emphasizes the pivotal role of comprehensive dataset content in shaping individuals' perceptions of a dataset card's overall quality. Overall, our study offers a unique perspective on analyzing dataset documentation through large-scale data science analysis and underlines the need for more thorough dataset documentation in machine learning research.
****
Kai Chen,Enze Xie,Zhe Chen,Yibo Wang,Lanqing HONG,Zhenguo Li,Dit-Yan Yeung
GeoDiffusion: Text-Prompted Geometric Control for Object Detection Data Generation
Diffusion models have attracted significant attention due to the remarkable ability to create content and generate data for tasks like image classification. However, the usage of diffusion models to generate the high-quality object detection data remains an underexplored area, where not only image-level perceptual quality but also geometric conditions such as bounding boxes and camera views are essential. Previous studies have utilized either copy-paste synthesis or layout-to-image (L2I) generation with specifically designed modules to encode the semantic layouts. In this paper, we propose the GeoDiffusion, a simple framework that can flexibly translate various geometric conditions into text prompts and empower pre-trained text-to-image (T2I) diffusion models for high-quality detection data generation. Unlike previous L2I methods, our GeoDiffusion is able to encode not only the bounding boxes but also extra geometric conditions such as camera views in self-driving scenes. Extensive experiments demonstrate GeoDiffusion outperforms previous L2I methods while maintaining 4x training time faster. To the best of our knowledge, this is the first work to adopt diffusion models for layout-to-image generation with geometric conditions and demonstrate that L2I-generated images can be beneficial for improving the performance of object detectors.
****
Yinan Huang,William Lu,Joshua Robinson,Yu Yang,Muhan Zhang,Stefanie Jegelka,Pan Li
On the Stability of Expressive Positional Encodings for Graphs
Designing effective positional encodings for graphs is key to building powerful graph transformers and enhancing message-passing graph neural networks. Although widespread, using Laplacian eigenvectors as positional encodings faces two fundamental challenges: (1) *Non-uniqueness*: there are many different eigendecompositions of the same Laplacian, and (2) *Instability*: small perturbations to the Laplacian could result in completely different eigenspaces, leading to unpredictable changes in positional encoding. Despite many attempts to address non-uniqueness, most methods overlook stability, leading to poor generalization on unseen graph structures. We identify the cause of instability to be the use of "hard partition'' of eigenspaces. Hence, we introduce Stable and Expressive Positional Encodings (SPE), an architecture for processing eigenvectors that uses eigenvalues to ``softly partition'' eigenspaces. SPE is the first architecture that is (1) provably stable, and (2) universally expressive for basis invariant functions

whilst respecting all symmetries of eigenvectors. Besides guaranteed stability, we prove that SPE is at least as expressive as existing methods, and highly capable of counting graph structures. Finally, we evaluate the effectiveness of our method on molecular property prediction, and out-of-distribution generalization tasks, finding improved generalization compared to existing positional encoding methods.
****
yisheng xiao,Juntao Li,Zechen Sun,Zechang Li,Qingrong Xia,Xinyu Duan,Zhefeng Wang,Min Zhang
Are Bert Family Good Instruction Followers? A Study on Their Potential And Limitations
Language modeling at scale has proven very effective and brought unprecedented success to natural language models. Many typical representatives, especially decoder-only models, e.g., BLOOM and LLaMA, and encoder-decoder models, e.g., Flan-T5 and AlexaTM, have exhibited incredible instruction-following capabilities while keeping strong task completion ability. These large language models can achieve superior performance in various tasks and even yield emergent capabilities, e.g., reasoning and universal generalization. Though the above two paradigms are mainstream and well explored, the potential of the BERT family, which are encoder-only based models and have ever been one of the most representative pre-trained models, also deserves attention, at least should be discussed. In this work, we adopt XML-R to explore the effectiveness of the BERT family for instruction following and zero-shot learning. We first design a simple yet effective strategy to utilize the encoder-only models for generation tasks and then conduct multi-task instruction tuning. Experimental results demonstrate that our fine-tuned model, Instruct-XMLR, outperforms Bloomz on all evaluation tasks and achieves comparable performance with mT0 on most tasks. Surprisingly, Instruct-XMLR also possesses strong task and language generalization abilities, indicating that Instruct-XMLR can also serve as a good instruction follower and zero-shot learner. Besides, Instruct-XMLR can accelerate decoding due to its non-autoregressive generation manner, achieving around 3 times speedup compared with current autoregressive large language models. Although we also witnessed several limitations through our experiments, such as the performance decline in long-generation tasks and the shortcoming of length prediction, Instruct-XMLR can still become a good member of the family of current large language models.
****
Yuyang Hu,Mauricio Delbracio,Peyman Milanfar,Ulugbek Kamilov
A Restoration Network as an Implicit Prior
Image denoisers have been shown to be powerful priors for solving inverse problems in imaging. In this work, we introduce a generalization of these methods that allows any image restoration network to be used as an implicit prior. The proposed method uses priors specified by deep neural networks pre-trained as general restoration operators. The method provides a principled approach for adapting state-of-the-art restoration models for other inverse problems. Our theoretical result analyzes its convergence to a stationary point of a global functional associated with the restoration operator. Numerical results show that the method using a super-resolution prior achieves state-of-the-art performance both quantitatively and qualitatively. Overall, this work offers a step forward for solving inverse problems by enabling the use of powerful pre-trained restoration models as priors.
****
Xinyu Huang,Youcai Zhang,Jinyu Ma,Weiwei Tian,Rui Feng,Yuejie Zhang,Yaqian Li,Yandong Guo,Lei Zhang
Tag2Text: Guiding Vision-Language Model via Image Tagging
This paper presents Tag2Text, a vision language pre-training (VLP) framework, which introduces image tagging into vision-language models to guide the learning of visual-linguistic features. In contrast to prior works which utilize object tags either manually labeled or automatically detected with a limited detector, our approach utilizes tags parsed from its paired text to learn an image tagger and meanwhile provides guidance to vision-language models. Given that, Tag2Text ca

n utilize large-scale annotation-free image tags in accordance with image-text pairs, and provides more diverse tag categories beyond objects. Strikingly, Tag2Text showcases the ability of a foundational image tagging model, with superior zero-shot performance even comparable to full supervision manner. Moreover, by leveraging tagging guidance, Tag2Text effectively enhances the performance of vision-language models on both generation-based and alignment-based tasks. Across a wide range of downstream benchmarks, Tag2Text achieves state-of-the-art results with similar model sizes and data scales, demonstrating the efficacy of the proposed tagging guidance.
****
Roman Pogodin,Jonathan Cornford,Arna Ghosh,Gauthier Gidel,Guillaume Lajoie,Blake Aaron Richards
Synaptic Weight Distributions Depend on the Geometry of Plasticity
A growing literature in computational neuroscience leverages gradient descent and learning algorithms that approximate it to study synaptic plasticity in the brain. However, the vast majority of this work ignores a critical underlying assumption: the choice of distance for synaptic changes - i.e. the geometry of synaptic plasticity. Gradient descent assumes that the distance is Euclidean, but many other distances are possible, and there is no reason that biology necessarily uses Euclidean geometry. Here, using the theoretical tools provided by mirror descent, we show that the distribution of synaptic weights will depend on the geometry of synaptic plasticity. We use these results to show that experimentally-observed log-normal weight distributions found in several brain areas are not consistent with standard gradient descent (i.e. a Euclidean geometry), but rather with non-Euclidean distances. Finally, we show that it should be possible to experimentally test for different synaptic geometries by comparing synaptic weight distributions before and after learning. Overall, our work shows that the current paradigm in theoretical work on synaptic plasticity that assumes Euclidean synaptic geometry may be misguided and that it should be possible to experimentally determine the true geometry of synaptic plasticity in the brain.
****
Haopeng Sun,Lumin Xu,Sheng Jin,Ping Luo,Chen Qian,Wentao Liu
PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation
Test-time adaptation (TTA) aims to adapt a pre-trained model from a source domain to a target domain only using online unlabeled target data during testing, without accessing to the source data or modifying the original training process. Among the various TTA methods, pseudo-labeling has gained popularity. However, the presence of incorrect pseudo-labels can hinder the effectiveness of target domain adaptation. To overcome this challenge, we propose a novel TTA method, called PROtotype GRAph Model based pseudo-label learning (PROGRAM). PROGRAM consists of two key components: (1) Prototype Graph Model (PGM) for reliable pseudo-label generation; (2) Robust Self-Training (RST) for test-time adaptation with noisy pseudo-labels. PGM constructs the graph using prototypes and test samples, facilitating effective message passing among them to generate more reliable pseudo-labels. RST combines the advantages of consistency regularization and pseudo-labeling to achieve robust target domain adaptation in the presence of noisy pseudo-labels. Our proposed PROGRAM can be easily integrated into existing baselines, resulting in consistent improvement. Extensive experiments show that our PROGRAM outperforms the existing TTA methods on multiple domain generalization and image corruption benchmarks.
****
Tong Wu,Ashwinee Panda,Jiachen T. Wang,Prateek Mittal
Privacy-Preserving In-Context Learning for Large Language Models
In-context learning (ICL) is an important capability of Large Language Models (LLMs), enabling these models to dynamically adapt based on specific, in-context exemplars, thereby improving accuracy and relevance.
However, LLM's responses may leak the sensitive private information contained in in-context exemplars.
To address this challenge, we propose Differentially Private In-context Learning

(DP-ICL), a general paradigm for privatizing ICL tasks.
The key idea for DP-ICL paradigm is generating differentially private responses through a noisy consensus among an ensemble of LLM's responses based on disjoint exemplar sets.
Based on the general paradigm of DP-ICL, we instantiate several techniques showing how to privatize ICL for text classification and language generation.
We experiment on four text classification benchmarks and two language generation tasks, and our empirical findings suggest that our DP-ICL achieves a strong utility-privacy tradeoff.
****
Jiuxiang Gu,Xiangxi Shi,Jason Kuen,Lu Qi,Ruiyi Zhang,Anqi Liu,Ani Nenkova,Tong Sun
ADOPD: A Large-Scale Document Page Decomposition Dataset
Research in document image understanding is hindered by limited high-quality data. To address this, we introduce ADOPD, a comprehensive dataset for document page decomposition. ADOPD stands out with its innovative data-driven approach for document taxonomy discovery during data collection, complemented by dense annotations. Our approach integrates large-scale pretrained models with a human-in-the-loop process to guarantee diversity and balance in the resulting data collection. Leveraging our data-driven document taxonomy, we collected and densely annotated labels for document images, covering four document image understanding tasks: Doc2Mask, Doc2Box, Doc2Tag, and Doc2Seq. Specifically, for each image, the annotations include human-labeled entity masks, text bounding boxes, as well as automatically generated tags and captions that have been manually cleaned. We conduct comprehensive experimental analyses to validate our data and assess the four tasks using various models. We envision ADOPD as a foundational dataset with the potential to drive future research in document understanding.
****
Bill Yuchen Lin,Abhilasha Ravichander,Ximing Lu,Nouha Dziri,Melanie Sclar,Khyathi Chandu,Chandra Bhagavatula,Yejin Choi
The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning
Alignment tuning has become the de facto standard practice for enabling base large language models (LLMs) to serve as open-domain AI assistants. The alignment tuning process typically involves instruction learning through supervised fine-tuning (SFT) and preference tuning via reinforcement learning from human feedback (RLHF). A recent study, LIMA (Zhou et al., 2023), shows that using merely 1K examples for SFT can achieve significant alignment performance as well, suggesting that the effect of alignment tuning might be "superficial." This raises questions about how exactly the alignment tuning transforms a base LLM.

We analyze the effect of alignment tuning by examining the token distribution shift between base LLMs and their aligned counterparts (e.g., Llama-2 and Llama-2-chat). Our findings reveal that base LLMs and their alignment-tuned versions perform nearly identically in decoding on the majority of token positions (i.e., they share the top-ranked tokens). Most distribution shifts occur with stylistic tokens (e.g., discourse markers, safety disclaimers). This direct evidence strongly supports the hypothesis that alignment tuning primarily learns to adopt the language style of AI assistants, and that the knowledge required for answering user queries predominantly comes from the base LLMs themselves.

Based on these findings, we rethink the alignment of LLMs by posing the research question: how effectively can we align base LLMs without SFT or RLHF? To address this, we introduce a simple, tuning-free alignment method, URIAL (Untuned LLMs with Restyled In-context Alignment). URIAL achieves effective alignment purely through in-context learning (ICL) with base LLMs, requiring as few as three constant stylistic examples and a system prompt. We conduct a fine-grained and interpretable evaluation on a diverse set of examples, named just-eval-instruct. Results demonstrate that base LLMs with URIAL can match or even surpass the performance of LLMs aligned with SFT (Mistral-7b-Instruct) or SFT+RLHF (Llama-2-70b-chat). We show that the gap between tuning-free and tuning-based alignment methods c

an be significantly reduced through strategic prompting and ICL. Our findings on the superficial nature of alignment tuning and results with URIAL suggest that deeper analysis and theoretical understanding of alignment is crucial to future LLM research.
****
Chengyu Dong,Liyuan Liu,Jingbo Shang
Toward Student-oriented Teacher Network Training for Knowledge Distillation
How to conduct teacher training for knowledge distillation is still an open problem. It has been widely observed that a best-performing teacher does not necessarily yield the best-performing student, suggesting a fundamental discrepancy between the current teacher training practice and the ideal teacher training strategy. To fill this gap, we explore the feasibility of training a teacher that is oriented toward student performance with empirical risk minimization (ERM). Our analyses are inspired by the recent findings that the effectiveness of knowledge distillation hinges on the teacher's capability to approximate the true label distribution of training inputs. We theoretically establish that ERM minimizer can approximate the true label distribution of training data as long as the feature extractor of the learner network is Lipschitz continuous and is robust to feature transformations. In light of our theory, we propose a teacher training method SoTeacher which incorporates Lipschitz regularization and consistency regularization into ERM. Experiments on benchmark datasets using various knowledge distillation algorithms and teacher-student pairs confirm that SoTeacher can improve student accuracy consistently.
****
Shuvendu Roy,Ali Etemad
Consistency-guided Prompt Learning for Vision-Language Models
We propose Consistency-guided Prompt learning (CoPrompt), a new fine-tuning method for vision-language models. Our approach improves the generalization of large foundation models when fine-tuned on downstream tasks in a few-shot setting. The basic idea of CoPrompt is to enforce a consistency constraint in the prediction of the trainable and pre-trained models to prevent overfitting on the downstream task. Additionally, we introduce the following two components into our consistency constraint to further boost the performance: enforcing consistency on two perturbed inputs and combining two dominant paradigms of tuning, prompting and adapter. Enforcing consistency on perturbed input serves to further regularize the consistency constraint, thereby improving generalization. Moreover, the integration of adapters and prompts not only enhances performance on downstream tasks but also offers increased tuning flexibility in both input and output spaces. This facilitates more effective adaptation to downstream tasks in a few-shot learning setting. Experiments show that CoPrompt outperforms existing methods on a range of evaluation suites, including base-to-novel generalization, domain generalization, and cross-dataset evaluation. On generalization, CoPrompt improves the state-of-the-art on zero-shot tasks and the overall harmonic mean over 11 datasets. Detailed ablation studies show the effectiveness of each of the components in CoPrompt. We make our code available at https://github.com/ShuvenduRoy/CoPrompt.
****
Yi-Rui Yang,Chang-Wei Shi,Wu-Jun Li
On the Effect of Batch Size in Byzantine-Robust Distributed Learning
Byzantine-robust distributed learning (BRDL), in which computing devices are likely to behave abnormally due to accidental failures or malicious attacks, has recently become a hot research topic. However, even in the independent and identically distributed (i.i.d.) case, existing BRDL methods will suffer a significant drop on model accuracy due to the large variance of stochastic gradients. Increasing batch sizes is a simple yet effective way to reduce the variance. However, when the total number of gradient computation is fixed, a too-large batch size will lead to a too-small iteration number (update number), which may also degrade the model accuracy. In view of this challenge, we mainly study the effect of batch size when the total number of gradient computation is fixed in this work. In particular, we show that when the total number of gradient computation is fixed

, the optimal batch size corresponding to the tightest theoretical upper bound in BRDL increases with the fraction of Byzantine workers. Therefore, compared to the case without attacks, a larger batch size is preferred when under Byzantine attacks. Motivated by the theoretical finding, we propose a novel method called Byzantine-robust stochastic gradient descent with normalized momentum (ByzSGDnm) in order to further increase model accuracy in BRDL. We theoretically prove the convergence of ByzSGDnm for general non-convex cases under Byzantine attacks. Empirical results show that when under Byzantine attacks, compared to the cases of small batch sizes, setting a relatively large batch size can significantly increase the model accuracy, which is consistent with our theoretical results. Moreover, ByzSGDnm can achieve higher model accuracy than existing BRDL methods when under deliberately crafted attacks. In addition, we empirically show that increasing batch sizes has the bonus of training acceleration.
****
Ruqi Bai,Saurabh Bagchi,David I. Inouye
Benchmarking Algorithms for Federated Domain Generalization
While prior federated learning (FL) methods mainly consider client heterogeneity, we focus on the *Federated Domain Generalization (DG)* task, which introduces train-test heterogeneity in the FL context.
Existing evaluations in this field are limited in terms of the scale of the clients and dataset diversity.
Thus, we propose a Federated DG benchmark that aim to test the limits of current methods with high client heterogeneity, large numbers of clients, and diverse datasets.
Towards this objective, we introduce a novel data partitioning method that allows us to distribute any domain dataset among few or many clients while controlling client heterogeneity. We then introduce and apply our methodology to evaluate 13 Federated DG methods, which include centralized DG methods adapted to the FL context, FL methods that handle client heterogeneity, and methods designed specifically for Federated DG on 7 datasets.
Our results suggest that, despite some progress, significant performance gaps remain in Federated DG, especially when evaluating with a large number of clients, high client heterogeneity, or more realistic datasets.
Furthermore, our extendable benchmark code will be publicly released to aid in benchmarking future Federated DG approaches.
****
Tingting Jiang,Qi Xu,Xuming Ran,Jiangrong Shen,Pan Lv,Qiang Zhang,Gang Pan
Adaptive deep spiking neural network with global-local learning via balanced excitatory and inhibitory mechanism
The training method of Spiking Neural Networks (SNNs) is an essential problem, and how to integrate local and global learning is a worthy research interest. However, the current integration methods do not consider the network conditions suitable for local and global learning, and thus fail to balance their advantages. In this paper, we propose an Excitation-Inhibition Mechanism-assisted Hybrid Learning(EIHL) algorithm that adjusts the network connectivity by using the excitation-inhibition mechanism and then switches between local and global learning according to the network connectivity. The experimental results on CIFAR10/100 and DVS-CIFAR10 demonstrate that the EIHL not only has better accuracy performance than other methods but also has excellent sparsity advantage. Especially, the Spiking VGG11 is trained by EIHL, STBP, and STDP on DVS_CIFAR10, respectively. The accuracy of the Spiking VGG11 model on EIHL is 62.45%, which is 4.35% higher than STBP and 11.40% higher than STDP, and the sparsity is 18.74%, which is 18.74% higher than the other two methods. Moreover, the excitation-inhibition mechanism used in our method also offers a new perspective on the field of SNN learning.
****
Francisco Andrade,Gabriel Peyré,Clarice Poon
Sparsistency for inverse optimal transport
Optimal Transport is a useful metric to compare probability distributions and to compute a pairing given a ground cost. Its entropic regularization variant (eOT) is crucial to have fast algorithms and reflect fuzzy/noisy matchings. This wo

rk focuses on Inverse Optimal Transport (iOT), the problem of inferring the grou
nd cost from samples drawn from a coupling that solves an eOT problem. It is a r
elevant problem that can be used to infer unobserved/missing links, and to obtai
n meaningful information about the structure of the ground cost yielding the pai
ring. On one side, iOT benefits from convexity, but on the other side, being ill
-posed, it requires regularization to handle the sampling noise. This work prese
nts an in-depth theoretical study of the ℓ_1 regularization to model for in
stance Euclidean costs with sparse interactions between features. Specifically,
 we derive a sufficient condition for the robust recovery of the sparsity of the
 ground cost that can be seen as a far reaching generalization of the Lasso's ce
lebrated ``Irrepresentability Condition''. To provide additional insight into th
is condition (consequently on the types of recoverable costs) we work out in det
ail the Gaussian case. Surprisingly, varying the entropic regularizer provides e
vidence that the Gaussian iOT interpolates between a graphical Lasso and a class
ical Lasso, thereby establishing a connection between iOT and graph estimation,
an important problem in ML.
****
Tailin Wu,Takashi Maruyama,Long Wei,Tao Zhang,Yilun Du,Gianluca Iaccarino,Jure L
eskovec
Compositional Generative Inverse Design
Inverse design, where we seek to design input variables in order to optimize an
underlying objective function, is an important problem that arises across fields
 such as mechanical engineering to aerospace engineering. Inverse design is typi
cally formulated as an optimization problem, with recent works leveraging optimi
zation across learned dynamics models. However, as models are optimized they ten
d to fall into adversarial modes, preventing effective sampling. We illustrate t
hat by instead optimizing over the learned energy function captured by the diffu
sion model, we can avoid such adversarial examples and significantly improve des
ign performance. We further illustrate how such a design system is compositional
, enabling us to combine multiple different diffusion models representing subcom
ponents of our desired system to design systems with every specified component.
In an N-body interaction task and a challenging 2D multi-airfoil design task, we
 demonstrate that by composing the learned diffusion model at test time, our met
hod allows us to design initial states and boundary shapes that are more complex
 than those in the training data. Our method generalizes to more objects for N-b
ody dataset and discovers formation flying to minimize drag in the multi-airfoil
 design task. Project website and code can be found at https://github.com/AI4Sci
ence-WestlakeU/cindm.
****
Sherry Yang,KwangHwan Cho,Amil Merchant,Pieter Abbeel,Dale Schuurmans,Igor Morda
tch,Ekin Dogus Cubuk
Scalable Diffusion for Materials Generation
■■■■Generative models trained on internet-scale data are capable of generating n
ovel and realistic texts, images, and videos. A natural next question is whether
 these models can advance science, for example by generating novel stable materi
als. Traditionally, models with explicit structures (e.g., graphs) have been use
d in modeling structural relationships in scientific data (e.g., atoms and bonds
 in crystals), but generating structures can be difficult to scale to large and
complex systems. Another challenge in generating materials is the mismatch betwe
en standard generative modeling metrics and downstream applications. For instanc
e, common metrics such as the reconstruction error do not correlate well with th
e downstream goal of discovering novel stable materials. In this work, we tackle
 the scalability challenge by developing a unified crystal representation that c
an represent any crystal structure (UniMat), followed by training a diffusion pr
obabilistic model on these UniMat representations. Our empirical results suggest
 that despite the lack of explicit structure modeling, UniMat can generate high
fidelity crystal structures from larger and more complex chemical systems, outpe
rforming previous graph-based approaches under various generative modeling metri
cs. To better connect the generation quality of materials to downstream applicat
ions, such as discovering novel stable materials, we propose additional metrics

for evaluating generative models of materials, including per-composition formation energy and stability with respect to convex hulls through decomposition energy from Density Function Theory (DFT). Lastly, we show that conditional generation with UniMat can scale to previously established crystal datasets with up to millions of crystals structures, outperforming random structure search (the current leading method for structure discovery) in discovering new stable materials.
****

Lingfeng Liu,Dong Ni,Hangjie Yuan
LUM-ViT: Learnable Under-sampling Mask Vision Transformer for Bandwidth Limited Optical Signal Acquisition
Bandwidth constraints during signal acquisition frequently impede real-time detection applications. Hyperspectral data is a notable example, whose vast volume compromises real-time hyperspectral detection. To tackle this hurdle, we introduce a novel approach leveraging pre-acquisition modulation to reduce the acquisition volume. This modulation process is governed by a deep learning model, utilizing prior information. Central to our approach is LUM-ViT, a Vision Transformer variant. Uniquely, LUM-ViT incorporates a learnable under-sampling mask tailored for pre-acquisition modulation. To further optimize for optical calculations, we propose a kernel-level weight binarization technique and a three-stage fine-tuning strategy. Our evaluations reveal that, by sampling a mere 10\% of the original image pixels, LUM-ViT maintains the accuracy loss within 1.8\% on the ImageNet classification task. The method sustains near-original accuracy when implemented on real-world optical hardware, demonstrating its practicality. Code will be available at https://github.com/MaxLLF/LUM-ViT.
****

Yining Jiao,Carlton Jude ZDANSKI,Julia S Kimbell,Andrew Prince,Cameron P Worden,Samuel Kirse,Christopher Rutter,Benjamin Shields,William Alexander Dunn,Jisan Mahmud,Marc Niethammer
$\texttt{NAISR}$: A 3D Neural Additive Model for Interpretable Shape Representation
Deep implicit functions (DIFs) have emerged as a powerful paradigm for many computer vision tasks such as 3D shape reconstruction, generation, registration, completion, editing, and understanding. However, given a set of 3D shapes with associated covariates there is at present no shape representation method which allows to precisely represent the shapes while capturing the individual dependencies on each covariate. Such a method would be of high utility to researchers to discover knowledge hidden in a population of shapes. For scientific shape discovery purpose, we propose a 3D Neural Additive Model for Interpretable Shape Representation ($\texttt{NAISR}$) which describes individual shapes by deforming a shape atlas in accordance to the effect of disentangled covariates. Our approach captures shape population trends and allows for patient-specific predictions through shape transfer. $\texttt{NAISR}$ is the first approach to combine the benefits of deep implicit shape representations with an atlas deforming according to specified covariates. We evaluate $\texttt{NAISR}$ with respect to shape reconstruction, shape disentanglement, shape evolution, and shape transfer on three datasets, i.e. 1) $\textit{Starman}$, a simulated 2D shape dataset; 2) ADNI hippocampus 3D shape dataset; 3) pediatric airway 3D shape dataset. Our experiments demonstrate that $\texttt{NAISR}$ achieves competitive shape reconstruction performance while retaining interpretability. Our code is available at https://github.com/uncbiag/NAISR.
****

SHIH-YING YEH,Yu-Guan Hsieh,Zhidong Gao,Bernard B W Yang,Giyeong Oh,Yanmin Gong
Navigating Text-To-Image Customization: From LyCORIS Fine-Tuning to Model Evaluation
Text-to-image generative models have garnered immense attention for their ability to produce high-fidelity images from text prompts. Among these, Stable Diffusion distinguishes itself as a leading open-source model in this fast-growing field. However, the intricacies of fine-tuning these models pose multiple challenges from new methodology integration to systematic evaluation. Addressing these

issues, this paper introduces LyCORIS (Lora beYond Conventional methods, Other Rank adaptation Implementations for Stable diffusion), an open-source library that offers a wide selection of fine-tuning methodologies for Stable Diffusion. Furthermore, we present a thorough framework for the systematic assessment of varied fine-tuning techniques. This framework employs a diverse suite of metrics and delves into multiple facets of fine-tuning, including hyperparameter adjustments and the evaluation with different prompt types across various concept categories. Through this comprehensive approach, our work provides essential insights into the nuanced effects of fine-tuning parameters, bridging the gap between state-of-the-art research and practical application.
****
Erik Englesson,Hossein Azizpour
Robust Classification via Regression for Learning with Noisy Labels
Deep neural networks and large-scale datasets have revolutionized the field of machine learning. However, these large networks are susceptible to overfitting to label noise, resulting in reduced generalization. To address this challenge, two promising approaches have emerged: i) loss reweighting, which reduces the influence of noisy examples on the training loss, and ii) label correction that replaces noisy labels with estimated true labels. These directions have been pursued separately or combined as independent methods, lacking a unified approach. In this work, we present a unified method that seamlessly combines loss reweighting and label correction to enhance robustness against label noise in classification tasks. Specifically, by leveraging ideas from compositional data analysis in statistics, we frame the problem as a regression task, where loss reweighting and label correction can naturally be achieved with a shifted Gaussian label noise model. Our unified approach achieves strong performance compared to recent baselines on several noisy labelled datasets. We believe this work is a promising step towards robust deep learning in the presence of label noise. Our code is available at: https://github.com/ErikEnglesson/SGN.
****
Qiyu Kang,Kai Zhao,Qinxu Ding,Feng Ji,Xuhao Li,Wenfei Liang,Yang Song,Wee Peng Tay
Unleashing the Potential of Fractional Calculus in Graph Neural Networks with FROND
We introduce the FRactional-Order graph Neural Dynamical network (FROND), a learning framework that extends traditional graph neural ordinary differential equation (ODE) models by incorporating the time-fractional Caputo derivative. Due to its non-local nature, fractional calculus allows our framework to capture long-term memories in the feature updating process, in contrast to the Markovian nature of updates in traditional graph neural ODE models. This can lead to improved graph representation learning.
We offer an interpretation of the feature updating process on graphs from a non-Markovian random walk perspective when the feature updating is governed by a diffusion process. We demonstrate analytically that over-smoothing can be mitigated in this setting.
To experimentally demonstrate the versatility of the FROND framework, we evaluate the fractional counterparts of various established graph ODE models. Their consistently superior performance, compared to their original counterparts, highlights the potential of the FROND framework as an effective extension to boost the efficacy of various graph neural ODE models.
****
Junsheng Zhou,Jinsheng Wang,Baorui Ma,Yu-Shen Liu,Tiejun Huang,Xinlong Wang
Uni3D: Exploring Unified 3D Representation at Scale
Scaling up representations for images or text has been extensively investigated in the past few years and has led to revolutions in learning vision and language. However, scalable representation for 3D objects and scenes is relatively unexplored. In this work, we present Uni3D, a 3D foundation model to explore the unified 3D representation at scale. Uni3D uses a 2D initialized ViT end-to-end pretrained to align the 3D point cloud features with the image-text aligned features. Via the simple architecture and pretext task, Uni3D can leverage abundant 2D pr

etrained models as initialization and image-text aligned models as the target, u
nlocking the great potential of 2D model zoos and scaling-up strategies to the 3
D world. We efficiently scale up Uni3D to one billion parameters, and set new re
cords on a broad range of 3D tasks, such as zero-shot classification, few-shot c
lassification, open-world understanding and zero-shot part segmentation. We sho
w that the strong Uni3D representation also enables applications such as 3D pain
ting and retrieval in the wild. We believe that Uni3D provides a new direction f
or exploring both scaling up and efficiency of the representation in 3D domain.
****
Diego Martinez-Taboada,Edward Kennedy
Counterfactual Density Estimation using Kernel Stein Discrepancies
Causal effects are usually studied in terms of the means of counterfactual distr
ibutions, which may be insufficient in many scenarios. Given a class of densitie
s known up to normalizing constants, we propose to model counterfactual distribu
tions by minimizing kernel Stein discrepancies in a doubly robust manner. This e
nables the estimation of counterfactuals over large classes of distributions whi
le exploiting the desired double robustness. We present a theoretical analysis o
f the proposed estimator, providing sufficient conditions for consistency and as
ymptotic normality, as well as an examination of its empirical performance.
****
Yongyuan Liang,Yanchao Sun,Ruijie Zheng,Xiangyu Liu,Benjamin Eysenbach,Tuomas Sa
ndholm,Furong Huang,Stephen Marcus McAleer
Game-Theoretic Robust Reinforcement Learning Handles Temporally-Coupled Perturba
tions
Deploying reinforcement learning (RL) systems requires robustness to uncertainty
 and model misspecification, yet prior robust RL methods typically only study no
ise introduced independently across time. However, practical sources of uncertai
nty are usually coupled across time.
We formally introduce temporally-coupled perturbations, presenting a novel chall
enge for existing robust RL methods. To tackle this challenge, we propose GRAD,
a novel game-theoretic approach that treats the temporally-coupled robust RL pro
blem as a partially-observable two-player zero-sum game. By finding an approxima
te equilibrium within this game, GRAD optimizes for general robustness against t
emporally-coupled perturbations. Experiments on continuous control tasks demonst
rate that, compared with prior methods, our approach achieves a higher degree of
 robustness to various types of attacks on different attack domains, both in set
tings with temporally-coupled perturbations and decoupled perturbations.
****
Hanqing Zeng,Hanjia Lyu,Diyi Hu,Yinglong Xia,Jiebo Luo
Mixture of Weak and Strong Experts on Graphs
Realistic graphs contain both rich self-features and informative neighborhood st
ructures, jointly handled by a GNN in the typical setup.
We propose to decouple the two modalities by mixture of weak and strong experts
(Mowst), where the weak expert is a light-weight Multi-layer Perceptron (MLP) ,
and the strong expert is an off-the-shelf Graph Neural Network (GNN). To adapt t
he experts' collaboration to different target nodes, we propose a "confidence" m
echanism based on the dispersion of the weak expert's prediction logits. The str
ong expert is conditionally activated in the low-confidence region when either t
he node's classification relies on neighborhood information, or the weak expert
has low model quality. We reveal interesting training dynamics by analyzing the
influence of the confidence function on loss: our training algorithm encourages
specialization of each expert by effectively generating a soft splitting of the
graph. In addition, our "confidence" design imposes a desirable bias towards the
 strong expert to benefit from the better generalization capability of GNNs. Mow
st is easy to optimize and achieves strong expressive power, with computation co
st comparable to a single GNN. Empirically, Mowst shows significant accuracy imp
rovement on 6 standard node classification benchmarks (including both homophilou
s and heterophilous graphs).
****
Miao Lu,Beining Wu,Xiaodong Yang,Difan Zou

Benign Oscillation of Stochastic Gradient Descent with Large Learning Rate
In this work, we theoretically investigate the generalization properties of neural networks (NN) trained by stochastic gradient descent (SGD) with large learning rates. Under such a training regime, our finding is that, the oscillation of the NN weights caused by SGD with large learning rates turns out to be beneficial to the generalization of the NN, potentially improving over the same NN trained by SGD with small learning rates that converges more smoothly. In view of this finding, we call such a phenomenon “benign oscillation”. Our theory towards demystifying such a phenomenon builds upon the feature learning perspective of deep learning. Specifically, we consider a feature-noise data generation model that consists of (i) weak features which have a small ℓ_2-norm and appear in each data point; (ii) strong features which have a large ℓ_2-norm but appear only in a certain fraction of all data points; and (iii) noise. We prove that NNs trained by oscillating SGD with a large learning rate can effectively learn the weak features in the presence of those strong features. In contrast, NNs trained by SGD with a small learning rate can only learn the strong features but make little progress in learning the weak features. Consequently, when it comes to the new testing data points that consist of only weak features, the NN trained by oscillating SGD with a large learning rate can still make correct predictions, while the NN trained by SGD with a small learning rate could not. Our theory sheds light on how large learning rate training benefits the generalization of NNs. Experimental results demonstrate our findings on the phenomenon of “benign oscillation”.
****
Guanzheng Chen,Xin Li,Zaiqiao Meng,Shangsong Liang,Lidong Bing
CLEX: Continuous Length Extrapolation for Large Language Models
Transformer-based Large Language Models (LLMs) are pioneering advances in many natural language processing tasks, however, their exceptional capabilities are restricted within the preset context window of Transformer. Position Embedding (PE) scaling methods, while effective in extending the context window to a specific length, demonstrate either notable limitations in their extrapolation abilities or sacrificing partial performance within the context window. Length extrapolation methods, although theoretically capable of extending the context window beyond the training sequence length, often underperform in practical long-context applications. To address these challenges, we propose Continuous Length EXtrapolation (CLEX) for LLMs. We generalise the PE scaling approaches to model the continuous dynamics by ordinary differential equations over the length scaling factor, thereby overcoming the constraints of current PE scaling methods designed for specific lengths. Moreover, by extending the dynamics to desired context lengths beyond the training sequence length, CLEX facilitates the length extrapolation with impressive performance in practical tasks. We demonstrate that CLEX can be seamlessly incorporated into LLMs equipped with Rotary Position Embedding, such as LLaMA and GPT-NeoX, with negligible impact on training and inference latency. Experimental results reveal that CLEX can effectively extend the context window to over 4× or almost 8× training length, with no deterioration in performance. Furthermore, when evaluated on the practical LongBench benchmark, our model trained on a 4k length exhibits competitive performance against state-of-the-art open-source models trained on context lengths up to 32k. Our code is available at https://github.com/DAMO-NLP-SG/CLEX.
****
Chenjie Cao,Xinlin Ren,Yanwei Fu
MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo
Recent advancements in learning-based Multi-View Stereo (MVS) methods have prominently featured transformer-based models with attention mechanisms. However, existing approaches have not thoroughly investigated the profound influence of transformers on different MVS modules, resulting in limited depth estimation capabilities. In this paper, we introduce MVSFormer++, a method that prudently maximizes the inherent characteristics of attention to enhance various components of the MVS pipeline. Formally, our approach involves infusing cross-view information into the pre-trained DINOv2 model to facilitate MVS learning. Furthermore, we emp

loy different attention mechanisms for the feature encoder and cost volume regularization, focusing on feature and spatial aggregations respectively. Additionally, we uncover that some design details would substantially impact the performance of transformer modules in MVS, including normalized 3D positional encoding, adaptive attention scaling, and the position of layer normalization. Comprehensive experiments on DTU, Tanks-and-Temples, BlendedMVS, and ETH3D validate the effectiveness of the proposed method. Notably, MVSFormer++ achieves state-of-the-art performance on the challenging DTU and Tanks-and-Temples benchmarks. Codes and models are available at https://github.com/maybeLx/MVSFormerPlusPlus.
****
Hsi-Ai Tsao,Lei Hsiung,Pin-Yu Chen,Sijia Liu,Tsung-Yi Ho
AutoVP: An Automated Visual Prompting Framework and Benchmark
Visual prompting (VP) is an emerging parameter-efficient fine-tuning approach to adapting pre-trained vision models to solve various downstream image-classification tasks. However, there has hitherto been little systematic study of the design space of VP and no clear benchmark for evaluating its performance. To bridge this gap, we propose AutoVP, an end-to-end expandable framework for automating VP design choices, along with 12 downstream image-classification tasks that can serve as a holistic VP-performance benchmark. Our design space covers 1) the joint optimization of the prompts; 2) the selection of pre-trained models, including image classifiers and text-image encoders; and 3) model output mapping strategies, including nonparametric and trainable label mapping. Our extensive experimental results show that AutoVP outperforms the best-known current VP methods by a substantial margin, having up to 6.7% improvement in accuracy; and attains a maximum performance increase of 27.5% compared to linear-probing (LP) baseline. AutoVP thus makes a two-fold contribution: serving both as an efficient tool for hyperparameter tuning on VP design choices, and as a comprehensive benchmark that can reasonably be expected to accelerate VP’s development. The source code is available at https://github.com/IBM/AutoVP.
****
Joar Max Viktor Skalse,Lucy Farnik,Sumeet Ramesh Motwani,Erik Jenner,Adam Gleave,Alessandro Abate
STARC: A General Framework For Quantifying Differences Between Reward Functions
In order to solve a task using reinforcement learning, it is necessary to first formalise the goal of that task as a *reward function*. However, for many real-world tasks, it is very difficult to manually specify a reward function that never incentivises undesirable behaviour. As a result, it is increasingly popular to use *reward learning algorithms*, which attempt to *learn* a reward function from data. However, the theoretical foundations of reward learning are not yet well-developed. In particular, it is typically not known when a given reward learning algorithm with high probability will learn a reward function that is safe to optimise. This means that reward learning algorithms generally must be evaluated empirically, which is expensive, and that their failure modes are difficult to anticipate in advance. One of the roadblocks to deriving better theoretical guarantees is the lack of good methods for *quantifying* the difference between reward functions. In this paper we provide a solution to this problem, in the form of a class of pseudometrics on the space of all reward functions that we call STARC (STAndardised Reward Comparison) metrics. We show that STARC metrics induce both an upper and a lower bound on worst-case regret, which implies that our metrics are tight, and that any metric with the same properties must be bilipschitz equivalent to ours. Moreover, we also identify a number of issues with reward metrics proposed by earlier works. Finally, we evaluate our metrics empirically, to demonstrate their practical efficacy. STARC metrics can be used to make both theoretical and empirical analysis of reward learning algorithms both easier and more principled.
****
Rebekka Burkholz
Batch normalization is sufficient for universal function approximation in CNNs
Normalization techniques, for which Batch Normalization (BN) is a popular choice, is an integral part of many deep learning architectures and contributes signif

icantly to the learning success. We provide a partial explanation for this phenomenon by proving that training normalization parameters alone is already sufficient for universal function approximation if the number of available, potentially random features matches or exceeds the weight parameters of the target networks that can be expressed. Our bound on the number of required features does not only improve on a recent result for fully-connected feed-forward architectures but also applies to CNNs with and without residual connections and almost arbitrary activation functions (which include ReLUs). Our explicit construction of a given target network solves a depth-width trade-off that is driven by architectural constraints and can explain why switching off entire neurons can have representational benefits, as has been observed empirically. To validate our theory, we explicitly match target networks that outperform experimentally obtained networks with trained BN parameters by utilizing a sufficient number of random features.
****

Diyang Li,Charles Ling,zhiqiang xu,Huan Xiong,Bin Gu
Learning No-Regret Sparse Generalized Linear Models with Varying Observation(s)
Generalized Linear Models (GLMs) encompass a wide array of regression and classification models, where prediction is a function of a linear combination of the input variables. Often in real-world scenarios, a number of observations would be added into or removed from the existing training dataset, necessitating the development of learning systems that can efficiently train optimal models with varying observations in an online (sequential) manner instead of retraining from scratch. Despite the significance of data-varying scenarios, most existing approaches to sparse GLMs concentrate on offline batch updates, leaving online solutions largely underexplored. In this work, we present the first algorithm without compromising accuracy for GLMs regularized by sparsity-enforcing penalties trained on varying observations. Our methodology is capable of handling the addition and deletion of observations simultaneously, while adaptively updating data-dependent regularization parameters to ensure the best statistical performance. Specifically, we recast sparse GLMs as a bilevel optimization objective upon varying observations and characterize it as an explicit gradient flow in the underlying space for the inner and outer subproblems we are optimizing over, respectively. We further derive a set of rules to ensure a proper transition at regions of non-smoothness, and establish the guarantees of theoretical consistency and finite convergence. Encouraging results are exhibited on real-world benchmarks.
****

Yaoming Wang,Jin Li,XIAOPENG ZHANG,Bowen Shi,Chenglin Li,Wenrui Dai,Hongkai Xiong,Qi Tian
BarLeRIa: An Efficient Tuning Framework for Referring Image Segmentation
Pre-training followed by full fine-tuning has gradually been substituted by Parameter-Efficient Tuning (PET) in the field of computer vision. PET has gained popularity, especially in the context of large-scale models, due to its ability to reduce transfer learning costs and conserve hardware resources. However, existing PET approaches primarily focus on recognition tasks and typically support uni-modal optimization, while neglecting dense prediction tasks and vision language interactions. To address this limitation, we propose a novel PET framework called **B**i-direction**a**l Inte**r**twined Vision **L**anguage Effici**e**nt Tuning for **R**eferring **I**mage Segment**a**tion (**BarLeRIa**), which leverages bi-directional intertwined vision language adapters to fully exploit the frozen pre-trained models' potential in cross-modal dense prediction tasks. In BarLeRIa, two different tuning modules are employed for efficient attention, one for global, and the other for local, along with an intertwined vision language tuning module for efficient modal fusion.
Extensive experiments conducted on RIS benchmarks demonstrate the superiority of BarLeRIa over prior PET methods with a significant margin, i.e., achieving an average improvement of 5.6\%. Remarkably, without requiring additional training datasets, BarLeRIa even surpasses SOTA full fine-tuning approaches. The code is available at https://github.com/NastrondAd/BarLeRIa.
****

Bowen Peng,Jeffrey Quesnelle,Honglu Fan,Enrico Shippole

YaRN: Efficient Context Window Extension of Large Language Models

Rotary Position Embeddings (RoPE) have been shown to effectively encode positional information in transformer-based language models. However, these models fail to generalize past the sequence length they were trained on. We present YaRN (Yet another RoPE extensioN method), a compute-efficient method to extend the context window of such models, requiring 10x less tokens and 2.5x less training steps than previous methods. Using YaRN, we show that LLaMA models can effectively utilize and extrapolate to context lengths much longer than their original pre-training would allow, while also surpassing previous the state-of-the-art at context window extension. In addition, we demonstrate that YaRN exhibits the capability to extrapolate beyond the limited context of a fine-tuning dataset. The models fine-tuned using YaRN has been made available and reproduced online up to 128k context length.

****

Frank Cole,Yulong Lu

Score-based generative models break the curse of dimensionality in learning a family of sub-Gaussian distributions

While score-based generative models (SGMs) have achieved remarkable successes in enormous image generation tasks, their mathematical foundations are still limited. In this paper, we analyze the approximation and generalization of SGMs in learning a family of sub-Gaussian probability distributions. We introduce a measure of complexity for probability distributions in terms of their relative density with respect to the standard Gaussian measure. We prove that if the log-relative density can be locally approximated by a neural network whose parameters can be suitably bounded, then the distribution generated by empirical score matching approximates the target distribution in total variation with a dimension-independent rate. We illustrate our theory through examples, which include certain mixtures of Gaussians. An essential ingredient of our proof is to derive a dimension-free deep network approximation rate for the true score function associated to the forward process, which is interesting in its own right.

****

Ziqi Xu,Debo Cheng,Jiuyong Li,Jixue Liu,Lin Liu,Kui Yu

Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder

An essential and challenging problem in causal inference is causal effect estimation from observational data. The problem becomes more difficult with the presence of unobserved confounding variables. The front-door adjustment is an approach for dealing with unobserved confounding variables. However, the restriction for the standard front-door adjustment is difficult to satisfy in practice. In this paper, we relax some of the restrictions by proposing the concept of conditional front-door (CFD) adjustment and develop the theorem that guarantees the causal effect identifiability of CFD adjustment. By leveraging the ability of deep generative models, we propose CFDiVAE to learn the representation of the CFD adjustment variable directly from data with the identifiable Variational AutoEncoder and formally prove the model identifiability. Extensive experiments on synthetic datasets validate the effectiveness of CFDiVAE and its superiority over existing methods. The experiments also show that the performance of CFDiVAE is less sensitive to the causal strength of unobserved confounding variables. We further apply CFDiVAE to a real-world dataset to demonstrate its potential application.

****

Ziyi Chen,Yi Zhou,Heng Huang

On the Hardness of Constrained Cooperative Multi-Agent Reinforcement Learning

Constrained cooperative multi-agent reinforcement learning (MARL) is an emerging learning framework that has been widely applied to manage multi-agent systems, and many primal-dual type algorithms have been developed for it. However, the convergence of primal-dual algorithms crucially relies on strong duality -- a condition that has not been formally proved in constrained cooperative MARL. In this work, we prove that strong duality fails to hold in constrained cooperative MARL, by revealing a nonconvex quadratic type constraint on the occupation measure induced by the product policy. Consequently, our reanalysis of the primal-dual a

lgorithm shows that its convergence rate is hindered by the nonzero duality gap. Then, we propose a decentralized primal approach for constrained cooperative MARL to avoid the duality gap, and our analysis shows that its convergence is hindered by another gap induced by the advantage functions. Moreover, we compare these two types of algorithms via concrete examples, and show that neither of them always outperforms the other one. Our study reveals that constrained cooperative MARL is generally a challenging and highly nonconvex problem, and its fundamental structure is very different from that of single-agent constrained RL.
****
Utkarsh Mall,Cheng Perng Phoo,Meilin Kelsey Liu,Carl Vondrick,Bharath Hariharan,Kavita Bala
Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment
We introduce a method to train vision-language models for remote-sensing images without using any textual annotations. Our key insight is to use co-located internet imagery taken on the ground as an intermediary for connecting remote-sensing images and language. Specifically, we train an image encoder for remote sensing images to align with the image encoder of CLIP using a large amount of paired internet and satellite images. Our unsupervised approach enables the training of a first-of-its-kind large scale VLM for remote sensing images at two different resolutions. We show that these VLMs enable zero-shot, open-vocabulary image classification, retrieval, segmentation and visual question answering for satellite images. On each of these tasks, our VLM trained without textual annotations outperforms existing VLMs trained with supervision, with gains of up to 20\% for classification and 80\% for segmentation.
****
Evan Hernandez,Arnab Sen Sharma,Tal Haklay,Kevin Meng,Martin Wattenberg,Jacob Andreas,Yonatan Belinkov,David Bau
Linearity of Relation Decoding in Transformer Language Models
Much of the knowledge encoded in transformer language models (LMs) may be expressed in terms of relations: relations between words and their synonyms, entities and their attributes, etc. We show that, for a subset of relations, this computation is well-approximated by a single linear transformation on the subject representation. Linear relation representations may be obtained by constructing a first-order approximation to the LM from a single prompt, and they exist for a variety of factual, commonsense, and linguistic relations. However, we also identify many cases in which LM predictions capture relational knowledge accurately, but this knowledge is not linearly encoded in their representations. Our results thus reveal a simple, interpretable, but heterogeneously deployed knowledge representation strategy in transformer LMs.
****
Junoh Lee,Hyunjun Jung,Jin-Hwi Park,Inhwan Bae,Hae-Gon Jeon
Geometry-Aware Projective Mapping for Unbounded Neural Radiance Fields
Estimating neural radiance fields (NeRFs) is able to generate novel views of a scene from known imagery. Recent approaches have afforded dramatic progress on small bounded regions of the scene. For an unbounded scene where cameras point in any direction and contents exist at any distance, certain mapping functions are used to represent it within a bounded space, yet they either work in object-centric scenes or focus on objects close to the camera. The goal of this paper is to understand how to design a proper mapping function that considers per-scene optimization, which remains unexplored. We first present a geometric understanding of existing mapping functions that express the relation between the bounded and unbounded scenes. Here, we exploit a stereographic projection method to explain failures of the mapping functions, where input ray samples are too sparse to account for scene geometry in unbounded regions. To overcome the failures, we propose a novel mapping function based on a p-norm distance, allowing to adaptively sample the rays by adjusting the p-value according to scene geometry, even in unbounded regions. To take the advantage of our mapping function, we also introduce a new ray parameterization to properly allocate ray samples in the geometry of unbounded regions. Through the incorporation of both the novel mapping funct

ion and the ray parameterization within existing NeRF frameworks, our method achieves state-of-the-art novel view synthesis results on a variety of challenging datasets.
****
Chencheng Cai,Xu Zhang,Edoardo Airoldi
Independent-Set Design of Experiments for Estimating Treatment and Spillover Effects under Network Interference
Interference is ubiquitous when conducting causal experiments over networks. Except for certain network structures, causal inference on the network in the presence of interference is difficult due to the entanglement between the treatment assignments and the interference levels. In this article, we conduct causal inference under interference on an observed, sparse, but connected network, and we propose a novel design of experiments based on an independent set. Compared to conventional designs, the independent-set design focuses on an independent subset of data and controls their interference exposures through the assignments to the rest (auxiliary set). We provide a lower bound on the size of the independent set from a greedy algorithm and justify the theoretical performance of estimators under the proposed design. Our approach is capable of estimating both spillover effects and treatment effects. We justify its superiority over conventional methods and illustrate the empirical performance through simulations.
****
Stone Tao,Arth Shukla,Tse-kai Chan,Hao Su
Reverse Forward Curriculum Learning for Extreme Sample and Demo Efficiency
Reinforcement learning (RL) presents a promising framework to learn policies through environment interaction, but often requires an infeasible amount of interaction data to solve complex tasks from sparse rewards. One direction includes augmenting RL with offline data demonstrating desired tasks, but past work often require a lot of high-quality demonstration data that is difficult to obtain, especially for domains such as robotics. Our approach consists of a reverse curriculum followed by a forward curriculum. Unique to our approach compared to past work is the ability to efficiently leverage more than one demonstration via a per-demonstration reverse curriculum generated via state resets. The result of our reverse curriculum is an initial policy that performs well on a narrow initial state distribution and helps overcome difficult exploration problems. A forward curriculum is then used to accelerate the training of the initial policy to perform well on the full initial state distribution of the task and improve demonstration and sample efficiency. We show how the combination of a reverse curriculum and forward curriculum in our method, RFCL, enables significant improvements in demonstration and sample efficiency compared against various state-of-the-art learning-from-demonstration baselines, even solving previously unsolvable tasks that require high precision and control. Website with code and visualizations are here: https://reverseforward-cl.github.io/
****
Yeming Wen,Swarat Chaudhuri
Batched Low-Rank Adaptation of Foundation Models
Low-Rank Adaptation (LoRA) has recently gained attention for fine-tuning foundation models by incorporating trainable low-rank matrices, thereby reducing the number of trainable parameters. While \lora/ offers numerous advantages, its applicability for real-time serving to a diverse and global user base
is constrained by its incapability to handle multiple task-specific adapters efficiently. This imposes a performance bottleneck in scenarios requiring personalized, task-specific adaptations for each incoming request.

To address this, we introduce FLoRA (Fast LoRA), a framework in which each input example in a minibatch can be associated with its unique low-rank adaptation weights, allowing for efficient batching of heterogeneous requests. We empirically demonstrate that \flora/ retains the performance merits of \lora/, showcasing competitive results on the MultiPL-E code generation benchmark spanning over 8 languages and a multilingual speech recognition task across 6 languages.
****

Jaehyung Kim,Jaehyun Nam,Sangwoo Mo,Jongjin Park,Sang-Woo Lee,Minjoon Seo,Jung-Woo Ha,Jinwoo Shin
SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs
Large language models (LLMs) have made significant advancements in various natural language processing tasks, including question answering (QA) tasks. While incorporating new information with the retrieval of relevant passages is a promising way to improve QA with LLMs, the existing methods often require additional fine-tuning which becomes infeasible with recent LLMs. Augmenting retrieved passages via prompting has the potential to address this limitation, but this direction has been limitedly explored. To this end, we design a simple yet effective framework to enhance open-domain QA (ODQA) with LLMs, based on the summarized retrieval (SuRe). SuRe helps LLMs predict more accurate answers for a given question, which are well-supported by the summarized retrieval that could be viewed as an explicit rationale extracted from the retrieved passages. Specifically, SuRe first constructs summaries of the retrieved passages for each of the multiple answer candidates. Then, SuRe confirms the most plausible answer from the candidate set by evaluating the validity and ranking of the generated summaries. Experimental results on diverse ODQA benchmarks demonstrate the superiority of SuRe, with improvements of up to 4.6\% in exact match (EM) and 4.0\% in F1 score over standard prompting approaches. SuRe also can be integrated with a broad range of retrieval methods and LLMs. Finally, the generated summaries from SuRe show additional advantages to measure the importance of retrieved passages and serve as more preferred rationales by models and humans.
****
Yizhi LI,Ruibin Yuan,Ge Zhang,Yinghao Ma,Xingran Chen,Hanzhi Yin,Chenghao Xiao,Chenghua Lin,Anton Ragni,Emmanouil Benetos,Norbert Gyenge,Roger Dannenberg,Ruibo Liu,Wenhu Chen,Gus Xia,Yemin Shi,Wenhao Huang,Zili Wang,Yike Guo,Jie Fu
MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training
Self-supervised learning (SSL) has recently emerged as a promising paradigm for training generalisable models on large-scale data in the fields of vision, text, and speech.
Although SSL has been proven effective in speech and audio, its application to music audio has yet to be thoroughly explored. This is partially due to the distinctive challenges associated with modelling musical knowledge, particularly tonal and pitched characteristics of music.
To address this research gap, we propose an acoustic **M**usic und**ER**standing model with large-scale self-supervised **T**raining (**MERT**), which incorporates teacher models to provide pseudo labels in the masked language modelling (MLM) style acoustic pre-training.
In our exploration, we identified an effective combination of teacher models, which outperforms conventional speech and audio approaches in terms of performance.
This combination includes an acoustic teacher based on Residual Vector Quantization - Variational AutoEncoder (RVQ-VAE) and a musical teacher based on the Constant-Q Transform (CQT).
Furthermore, we explore a wide range of settings to overcome the instability in acoustic language model pre-training, which allows our designed paradigm to scale from 95M to 330M parameters.
Experimental results indicate that our model can generalise and perform well on 14 music understanding tasks and attain state-of-the-art (SOTA) overall scores.
****
Catarina G Belém,Preethi Seshadri,Yasaman Razeghi,Sameer Singh
Are Models Biased on Text without Gender-related Language?
In the large language models era, it is imperative to measure and understand how gender biases present in the training data influence model behavior.
Previous works construct benchmarks around known stereotypes (e.g., occupations) and demonstrate high levels of gender bias in large language models, raising serious concerns about models exhibiting undesirable behaviors.
We expand on existing literature by asking the question: \textit{Do large langua

ge models still favor one gender over the other in non-stereotypical settings?}
To tackle this question, we restrict language model evaluation to a \textit{neutral} subset, in which sentences are free of pronounced word-gender associations.

After characterizing these associations in terms of pretraining data statistics, we use them to (1) create a new benchmark with low gender-word associations, and (2) repurpose popular benchmarks in the gendered pronoun setting | WinoBias and \Winogender |, removing pronounced gender-correlated words.
Surprisingly, when testing $20+$ models (e.g., Llama-2, Pythia, and OPT) in the proposed benchmarks, we still detect critically high gender bias across all tested models.
For instance, after adjusting for strong word-gender associations, we find that all models still exhibit clear gender preferences in about 60%-95% of the sentences, representing a small change (up to 5%) from the original \textit{stereotypical} setting.
By demonstrating that measured bias is not necessarily due to the presence of highly gender-associated words, our work highlights important questions about bias evaluation as well as potentially underlying model biases.
****
Zifan Wu,Bo Tang,Qian Lin,Chao Yu,Shangqin Mao,Qianlong Xie,Xingxing Wang,Dong Wang
Off-Policy Primal-Dual Safe Reinforcement Learning
Primal-dual safe RL methods commonly perform iterations between the primal update of the policy and the dual update of the Lagrange Multiplier. Such a training paradigm is highly susceptible to the error in cumulative cost estimation since this estimation serves as the key bond connecting the primal and dual update processes. We show that this problem causes significant underestimation of cost when using off-policy methods, leading to the failure to satisfy the safety constraint. To address this issue, we propose conservative policy optimization, which learns a policy in a constraint-satisfying area by considering the uncertainty in cost estimation. This improves constraint satisfaction but also potentially hinders reward maximization. We then introduce local policy convexification to help eliminate such suboptimality by gradually reducing the estimation uncertainty. We provide theoretical interpretations of the joint coupling effect of these two ingredients and further verify them by extensive experiments. Results on benchmark tasks show that our method not only achieves an asymptotic performance comparable to state-of-the-art on-policy methods while using much fewer samples, but also significantly reduces constraint violation during training. Our code is available at https://github.com/ZifanWu/CAL.
****
Dingyuan Shi,Yongxin Tong,Zimu Zhou,Ke Xu,Zheng Wang,Jieping Ye
GRAPH-CONSTRAINED DIFFUSION FOR END-TO-END PATH PLANNING
Path planning underpins various applications such as transportation, logistics, and robotics.
Conventionally, path planning is formulated with explicit optimization objectives such as distance or time.
However, real-world data reveals that user intentions are hard-to-model, suggesting a need for data-driven path planning that implicitly incorporates the complex user intentions.
In this paper, we propose GDP, a diffusion-based model for end-to-end data-driven path planning.
It effectively learns path patterns via a novel diffusion process that incorporates constraints from road networks, and plans paths as conditional path generation given the origin and destination as prior evidence.
GDP is the first solution that bypasses the traditional search-based frameworks, a long-standing performance bottleneck in path planning.
We validate the efficacy of GDP on two real-world datasets.
Our GDP beats strong baselines by 14.2% ~ 43.5% and achieves state-of-the-art performances.
****

Florian Grötschla,Joël Mathys,Robert Veres,Roger Wattenhofer
CoRe-GD: A Hierarchical Framework for Scalable Graph Visualization with GNNs
Graph Visualization, also known as Graph Drawing, aims to find geometric embeddings of graphs that optimize certain criteria. Stress is a widely used metric; stress is minimized when every pair of nodes is positioned at their shortest path distance. However, stress optimization presents computational challenges due to its inherent complexity and is usually solved using heuristics in practice. We introduce a scalable Graph Neural Network (GNN) based Graph Drawing framework with sub-quadratic runtime that can learn to optimize stress. Inspired by classical stress optimization techniques and force-directed layout algorithms, we create a coarsening hierarchy for the input graph. Beginning at the coarsest level, we iteratively refine and un-coarsen the layout, until we generate an embedding for the original graph. To enhance information propagation within the network, we propose a novel positional rewiring technique based on intermediate node positions. Our empirical evaluation demonstrates that the framework achieves state-of-the-art performance while remaining scalable.
****
Tanishq Kumar,Blake Bordelon,Samuel J. Gershman,Cengiz Pehlevan
Grokking as the transition from lazy to rich training dynamics
We propose that the grokking phenomenon, where the train loss of a neural network decreases much earlier than its test loss, can arise due to a neural network transitioning from lazy training dynamics to a rich, feature learning regime. To illustrate this mechanism, we study the simple setting of vanilla gradient descent on a polynomial regression problem with a two layer neural network which exhibits grokking without regularization in a way that cannot be explained by existing theories. We identify sufficient statistics for the test loss of such a network, and tracking these over training reveals that grokking arises in this setting when the network first attempts to fit a kernel regression solution with its initial features, followed by late-time feature learning where a generalizing solution is identified after train loss is already low. We find that the key determinants of grokking are the rate of feature learning---which can be controlled precisely by parameters that scale the network output---and the alignment of the initial features with the target function $y(x)$. We argue this delayed generalization arises when (1) the top eigenvectors of the initial neural tangent kernel and the task labels $y(x)$ are misaligned, but (2) the dataset size is large enough so that it is possible for the network to generalize eventually, but not so large that train loss perfectly tracks test loss at all epochs, and (3) the network begins training in the lazy regime so does not learn features immediately. We conclude with evidence that this transition from lazy (linear model) to rich training (feature learning) can control grokking in more general settings, like on MNIST, one-layer Transformers, and student-teacher networks.
****
Zhirui Chen,P. N. Karthik,Yeow Meng Chee,Vincent Tan
Fixed-Budget Differentially Private Best Arm Identification
We study best arm identification (BAI) in linear bandits in the fixed-budget regime under differential privacy constraints, when the arm rewards are supported on the unit interval.
Given a finite budget T and a privacy parameter $\varepsilon>0$, the goal is to minimise the error probability in finding the arm with the largest mean after T sampling rounds, subject to the constraint that the policy of the decision maker satisfies a certain {\em ε-differential privacy} (ε-DP) constraint. We construct a policy satisfying the ε-DP constraint (called {\sc DP-BAI}), based on the principle of {\em maximum absolute determinants}, and derive an upper bound on its error probability. Furthermore, we derive a minimax lower bound on the error probability, and demonstrate that the lower and the upper bounds decay exponentially in T, with exponents in the two bounds matching order-wise in (a) the sub-optimality gaps of the arms, (b) ε, and (c) the problem complexity that is expressible as the sum of two terms, one characterising the complexity of standard fixed-budget BAI (without privacy constraints), and the other accounting for the ε-DP constraint.

Additionally, we present some auxiliary results that contribute to the derivation of the lower bound on the error probability. These results, we posit, may be of independent interest and could prove instrumental in proving lower bounds on error probabilities in several other bandit problems.
Whereas prior works provide results for BAI in the fixed-budget regime without privacy constraints or in the fixed-confidence regime with privacy constraints, our work fills the gap in the literature by providing the results for BAI in the fixed-budget regime under the ε-DP constraint.
****
Yehui Tang,Hao Xiong,Nianzu Yang,Tailong Xiao,Junchi Yan
Towards LLM4QPE: Unsupervised Pretraining of Quantum Property Estimation and A Benchmark
Estimating the properties of quantum systems such as quantum phase has been critical in addressing the essential quantum many-body problems in physics and chemistry. Deep learning models have been recently introduced to property estimation, surpassing conventional statistical approaches. However, these methods are tailored to the specific task and quantum data at hand. It remains an open and attractive question for devising a more universal task-agnostic pretraining model for quantum property estimation. In this paper, we propose LLM4QPE, a large language model style quantum task-agnostic pretraining and finetuning paradigm that 1) performs unsupervised pretraining on diverse quantum systems with different physical conditions; 2) uses the pretrained model for supervised finetuning and delivers high performance with limited training data, on downstream tasks. It mitigates the cost for quantum data collection and speeds up convergence. Extensive experiments show the promising efficacy of LLM4QPE in various tasks including classifying quantum phases of matter on Rydberg atom model and predicting two-body correlation function on anisotropic Heisenberg model.
****
Zhijing Jin,Jiarui Liu,Zhiheng LYU,Spencer Poff,Mrinmaya Sachan,Rada Mihalcea,Mona T. Diab,Bernhard Schölkopf
Can Large Language Models Infer Causation from Correlation?
Causal inference is one of the hallmarks of human intelligence. While the field of CausalNLP has attracted much interest in the recent years, existing causal inference datasets in NLP primarily rely on discovering causality from empirical knowledge (e.g., commonsense knowledge). In this work, we propose the first benchmark dataset to test the pure causal inference skills of large language models (LLMs). Specifically, we formulate a novel task CORR2CAUSE, which takes a set of correlational statements and determines the causal relationship between the variables. We curate a large-scale dataset of more than 200K samples, on which we evaluate 17 existing LLMs. Through our experiments, we identify a key shortcoming of LLMs in terms of their causal inference skills, and show that these models achieve almost close to random performance on the task. This shortcoming is somewhat mitigated when we try to re-purpose LLMs for this skill via finetuning, but we find that these models still fail to generalize – they can only perform causal inference in in-distribution settings when variable names and textual expressions used in the queries are similar to those in the training set, but fail in out-of-distribution settings generated by perturbing these queries. CORR2CAUSE is a challenging task for LLMs, and would be helpful in guiding future research on improving LLMs’ pure reasoning skills and generalizability. Our data is at https://huggingface.co/datasets/causalnlp/corr2cause. Our code is at https://github.com/causalNLP/corr2cause.
****
Prajjwal Bhargava,Rohan Chitnis,Alborz Geramifard,Shagun Sodhani,Amy Zhang
When should we prefer Decision Transformers for Offline Reinforcement Learning?
Offline reinforcement learning (RL) allows agents to learn effective, return-maximizing policies from a static dataset. Three popular algorithms for offline RL are Conservative Q-Learning (CQL), Behavior Cloning (BC), and Decision Transformer (DT), from the class of Q-Learning, Imitation Learning, and Sequence Modeling respectively. A key open question is: which algorithm is preferred under what conditions? We study this question empirically by exploring the performance of th

ese algorithms across the commonly used D4RL and Robomimic benchmarks. We design targeted experiments to understand their behavior concerning data suboptimality, task complexity, and stochasticity. Our key findings are: (1) DT requires more data than CQL to learn competitive policies but is more robust; (2) DT is a substantially better choice than both CQL and BC in sparse-reward and low-quality data settings; (3) DT and BC are preferable as task horizon increases, or when data is obtained from human demonstrators; and (4) CQL excels in situations characterized by the combination of high stochasticity and low data quality. We also investigate architectural choices and scaling trends for DT on \textsc{atari} and D4RL and make design/scaling recommendations. We find that scaling the amount of data for DT by 5x gives a 2.5x average score improvement on Atari.
****

Luo donghao,wang xue
ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis
Recently, Transformer-based and MLP-based models have emerged rapidly and won dominance in time series analysis. In contrast, convolution is losing steam in time series tasks nowadays for inferior performance. This paper studies the open question of how to better use convolution in time series analysis and makes efforts to bring convolution back to the arena of time series analysis. To this end,
we modernize the traditional TCN and conduct time series related modifications to make it more suitable for time series tasks. As the outcome, we propose ModernTCN and successfully solve this open question through a seldom-explored way in time series community. As a pure convolution structure, ModernTCN still achieves the consistent state-of-the-art performance on five mainstream time series
analysis tasks while maintaining the efficiency advantage of convolution-based models, therefore providing a better balance of efficiency and performance than state-of-the-art Transformer-based and MLP-based models. Our study further reveals that, compared with previous convolution-based models, our ModernTCN has much larger effective receptive fields (ERFs), therefore can better unleash the
potential of convolution in time series analysis. Code is available at this repository:
https://github.com/luodhhh/ModernTCN.
****

Irene Cannistraci,Luca Moschella,Marco Fumero,Valentino Maiorca,Emanuele Rodolà
From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication
It has been observed that representations learned by distinct neural networks conceal structural similarities when the models are trained under similar inductive biases. From a geometric perspective, identifying the classes of transformations and the related invariances that connect these representations is fundamental to unlocking applications, such as merging, stitching, and reusing different neural modules. However, estimating task-specific transformations a priori can be challenging and expensive due to several factors (e.g., weights initialization, training hyperparameters, or data modality). To this end, we introduce a versatile method to directly incorporate a set of invariances into the representations, constructing a product space of invariant components on top of the latent representations without requiring prior knowledge about the optimal invariance to infuse. We validate our solution on classification and reconstruction tasks, observing consistent latent similarity and downstream performance improvements in a zero-shot stitching setting. The experimental analysis comprises three modalities (vision, text, and graphs), twelve pretrained foundational models, nine benchmarks, and several architectures trained from scratch.
****

Xiaoxiao Sun,Yue Yao,Shengjin Wang,Hongdong Li,Liang Zheng
Alice Benchmarks: Connecting Real World Re-Identification with the Synthetic
For object re-identification (re-ID), learning from synthetic data has become a promising strategy to cheaply acquire large-scale annotated datasets and effecti

ve models, with few privacy concerns. Many interesting research problems arise f
rom this strategy, e.g., how to reduce the domain gap between synthetic source a
nd real-world target. To facilitate developing more new approaches in learning f
rom synthetic data, we introduce the Alice benchmarks, large-scale datasets prov
iding benchmarks as well as evaluation protocols to the research community. With
in the Alice benchmarks, two object re-ID tasks are offered: person and vehicle
re-ID. We collected and annotated two challenging real-world target datasets: Al
icePerson and AliceVehicle, captured under various illuminations, image resoluti
ons, etc. As an important feature of our real target, the clusterability of its
training set is not manually guaranteed to make it closer to a real domain adapt
ation test scenario. Correspondingly, we reuse existing PersonX and VehicleX as
synthetic source domains. The primary goal is to train models from synthetic dat
a that can work effectively in the real world. In this paper, we detail the sett
ings of Alice benchmarks, provide an analysis of existing commonly-used domain a
daptation methods, and discuss some interesting future directions. An online ser
ver has been set up for the community to evaluate methods conveniently and fairl
y. Datasets and the online server details are available at https://sites.google.
com/view/alice-benchmarks.
****
Isaac Reid,Krzysztof Marcin Choromanski,Eli Berger,Adrian Weller
General Graph Random Features
We propose a novel random walk-based algorithm for unbiased estimation of arbitr
ary functions of a weighted adjacency matrix, coined
general graph random features (g-GRFs). This includes many of the most popular e
xamples of kernels defined on the nodes of a graph. Our algorithm enjoys subquad
ratic time complexity with respect to the number of nodes, overcoming the notori
ously prohibitive cubic scaling of exact graph kernel evaluation. It can also be
 trivially distributed across machines, permitting learning on much larger netwo
rks. At the heart of the algorithm is a modulation function which upweights or d
ownweights the contribution from different random walks depending on their lengt
hs. We show that by parameterising it with a neural network we can obtain g-GRFs
 that give higher-quality kernel estimates or perform efficient, scalable kernel
 learning. We provide robust theoretical analysis and support our findings with
experiments including pointwise estimation of fixed graph kernels, solving non-h
omogeneous graph ordinary differential equations, node clustering and kernel reg
ression on triangular meshes.
****
Dejiao Zhang,Wasi Uddin Ahmad,Ming Tan,Hantian Ding,Ramesh Nallapati,Dan Roth,Xi
aofei Ma,Bing Xiang
CODE REPRESENTATION LEARNING AT SCALE
Recent studies have shown that code language model at scale demonstrate signific
ant performance gains on downstream tasks, i.e., code generation. However, most
of the existing works on code representation learning train models at a hundred
million parameter scale using very limited pretraining corpora. In this work, we
 fuel code representation learning with a vast amount of code data via a two-sta
ge pretraining scheme. We first train the encoders via a mix that leverages both
 randomness in masking language modeling and implicit structure and semantic asp
ects of programming language. We then enhance the representations via contrastiv
e learning with hard negative and hard positive constructed in an unsupervised m
anner. We establish an off-the-shelf encoder model that persistently outperforms
 the existing models on a wide variety of downstream tasks by large margins. To
comprehend the factors contributing to successful code representation learning,
we conduct detailed ablations and share our findings on (i) a customized and eff
ective token-level denoising scheme for source code; (ii) the importance of hard
 negatives and hard positives; (iii) how the proposed bimodal contrastive learni
ng boost the cross-lingual semantic search performance; and (iv) how the pretrai
ning schemes decide the downstream task performance scales with the model size.
****
Yaofo Chen,Shuaicheng Niu,Shoukai Xu,Hengjie Song,Yaowei Wang,Mingkui Tan
Towards Robust and Efficient Cloud-Edge Elastic Model Adaptation via Selective E

ntropy Distillation
The conventional deep learning paradigm often involves training a deep model on a server and then deploying the model or its distilled ones to resource-limited edge devices. Usually, the models shall remain fixed once deployed (at least for some period) due to the potential high cost of model adaptation for both the server and edge sides. However, in many real-world scenarios, the test environments may change dynamically (known as distribution shifts), which often results in degraded performance. Thus, one has to adapt the edge models promptly to attain promising performance. Moreover, with the increasing data collected at the edge, this paradigm also fails to further adapt the cloud model for better performance. To address these, we encounter two primary challenges: 1) the edge model has limited computation power and may only support forward propagation; 2) the data transmission budget between cloud and edge devices is limited in latency-sensitive scenarios. In this paper, we establish a Cloud-Edge Elastic Model Adaptation (CEMA) paradigm in which the edge models only need to perform forward propagation and the edge models can be adapted online. In our CEMA, to reduce the communication burden, we devise two criteria to exclude unnecessary samples from uploading to the cloud, i.e., dynamic unreliable and low-informative sample exclusion. Based on the uploaded samples, we update and distribute the affine parameters of normalization layers by distilling from the stronger foundation model to the edge model with a sample replay strategy. Extensive experimental results on ImageNet-C and ImageNet-R verify the effectiveness of our CEMA.
****
Yusuke Sekikawa,Shingo Yashima
SAS: Structured Activation Sparsification
Wide networks usually yield better accuracy than their narrower counterpart at the expense of the massive $\texttt{mult}$ cost.
To break this tradeoff, we advocate a novel concept of $\textit{Structured Activation Sparsification}$, dubbed SAS, which boosts accuracy without increasing computation by utilizing the projected sparsity in activation maps with a specific structure.
Concretely, the projected sparse activation is allowed to have N nonzero value among M consecutive activations.
Owing to the local structure in sparsity, the wide $\texttt{matmul}$ between a dense weight and the sparse activation is executed as an equivalent narrow $\texttt{matmul}$ between a dense weight and dense activation, which is compatible with NVIDIA's $\textit{SparseTensorCore}$ developed for the N:M structured sparse weight.
In extensive experiments, we demonstrate that increasing sparsity monotonically improves accuracy (up to 7% on CIFAR10) without increasing the $\texttt{mult}$ count.
Furthermore, we show that structured sparsification of $\textit{activation}$ scales better than that of $\textit{weight}$ given the same computational budget.
****
Yat Long Lo,Biswa Sengupta,Jakob Nicolaus Foerster,Michael Noukhovitch
Learning Multi-Agent Communication with Contrastive Learning
Communication is a powerful tool for coordination in multi-agent RL. But inducing an effective, common language is a difficult challenge, particularly in the decentralized setting. In this work, we introduce an alternative perspective where communicative messages sent between agents are considered as different incomplete views of the environment state. By examining the relationship between messages sent and received, we propose to learn to communicate using contrastive learning to maximize the mutual information between messages of a given trajectory. In communication-essential environments, our method outperforms previous work in both performance and learning speed. Using qualitative metrics and representation probing, we show that our method induces more symmetric communication and captures global state information from the environment. Overall, we show the power of contrastive learning and the importance of leveraging messages as encodings for effective communication.
****

Nguyen Hung-Quang,Yingjie Lao,Tung Pham,Kok-Seng Wong,Khoa D Doan
Understanding the Robustness of Randomized Feature Defense Against Query-Based A
dversarial Attacks
Recent works have shown that deep neural networks are vulnerable to adversarial
examples that find samples close to the original image but can make the model mi
sclassify. Even with access only to the model's output, an attacker can employ b
lack-box attacks to generate such adversarial examples. In this work, we propose
 a simple and lightweight defense against black-box attacks by adding random noi
se to hidden features at intermediate layers of the model at inference time. Our
 theoretical analysis confirms that this method effectively enhances the model's
 resilience against both score-based and decision-based black-box attacks. Impor
tantly, our defense does not necessitate adversarial training and has minimal im
pact on accuracy, rendering it applicable to any pre-trained model. Our analysis
 also reveals the significance of selectively adding noise to different parts of
 the model based on the gradient of the adversarial objective function, which ca
n be varied during the attack. We demonstrate the robustness of our defense agai
nst multiple black-box attacks through extensive empirical experiments involving
 diverse models with various architectures.
****
Hung Le,Hailin Chen,Amrita Saha,Akash Gokul,Doyen Sahoo,Shafiq Joty
CodeChain: Towards Modular Code Generation Through Chain of Self-revisions with
Representative Sub-modules
Large Language Models (LLMs) have already become quite proficient at solving sim
pler programming tasks like those in HumanEval or MBPP benchmarks. However, solv
ing more complex and competitive programming tasks is still quite challenging fo
r these models - possibly due to their tendency to generate solutions as monolit
hic code blocks instead of decomposing them into logical sub-tasks and sub-modul
es. On the other hand, experienced programmers instinctively write modularized c
ode with abstraction for solving complex tasks, often reusing previously develop
ed modules. To address this gap, we propose CodeChain, a novel framework for inf
erence that elicits modularized code generation through a chain of self-revision
s, each being guided by some representative sub-modules generated in previous it
erations. Concretely, CodeChain first instructs the LLM to generate modularized
codes through chain-of-thought prompting. Then it applies a chain of self-revisi
ons by iterating the two steps: 1) extracting and clustering the generated sub-m
odules and selecting the cluster representatives as the more generic and re-usab
le implementations, and 2) augmenting the original chain-of-thought prompt with
these selected module-implementations and instructing the LLM to re-generate new
 modularized solutions. We find that by naturally encouraging the LLM to reuse t
he previously developed and verified sub-modules, CodeChain can significantly bo
ost both modularity as well as correctness of the generated solutions, achieving
 relative pass@1 improvements of 35\% on APPS and 76\% on CodeContests. It is sh
own to be effective on both OpenAI LLMs as well as open-sourced LLMs like Wizard
Coder. We also conduct comprehensive ablation studies with different methods of
prompting, number of clusters, model sizes, program qualities, etc., to provide
useful insights that underpin CodeChain's success.
****
Hubert Siuzdak
Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for
 high-quality audio synthesis
Recent advancements in neural vocoding are predominantly driven by Generative Ad
versarial Networks (GANs) operating in the time-domain. While effective, this ap
proach neglects the inductive bias offered by time-frequency representations, re
sulting in reduntant and computionally-intensive upsampling operations. Fourier-
based time-frequency representation is an appealing alternative, aligning more a
ccurately with human auditory perception, and benefitting from well-established
fast algorithms for its computation. Nevertheless, direct reconstruction of comp
lex-valued spectrograms has been historically problematic, primarily due to phas
e recovery issues. This study seeks to close this gap by presenting Vocos, a new
 model that directly generates Fourier spectral coefficients. Vocos not only mat

ches the state-of-the-art in audio quality, as demonstrated in our evaluations, but it also substantially improves computational efficiency, achieving an order of magnitude increase in speed compared to prevailing time-domain neural vocodin g approaches. The source code and model weights have been open-sourced.
****
Saleh Ashkboos,Maximilian L. Croci,Marcelo Gennari do Nascimento,Torsten Hoefler ,James Hensman
SliceGPT: Compress Large Language Models by Deleting Rows and Columns
Large language models have become the cornerstone of natural language processing , but their use comes with substantial costs in terms of compute and memory reso urces. Sparsification provides a solution to alleviate these resource constraint s, and recent works have shown that trained models can be sparsified post-hoc. E xisting sparsification techniques face challenges as they need additional data s tructures and offer constrained speedup with current hardware. In this paper we present SliceGPT, a new post-training sparsification scheme which replaces each weight matrix with a smaller (dense) matrix, reducing the embedding dimension of the network. Through extensive experimentation we show that SliceGPT can remove up to 25% of the model parameters (including embeddings) for LLAMA-2 70B, OPT 6 6B and Phi-2 models while maintaining 99%, 99% and 90% zero-shot task performanc e of the dense model respectively. Our sliced models run on fewer GPUs and run f aster without any additional code optimization: on 24GB consumer GPUs we reduce the total compute for inference on LLAMA-2 70B to 64% of that of the dense model ; on 40GB A100 GPUs we reduce it to 66%. We offer a new insight, computational i nvariance in transformer networks, which enables SliceGPT and we hope it will in spire and enable future avenues to reduce memory and computation demands for pre -trained models.
****
Jiale Zhang,Yulun Zhang,Jinjin Gu,Jiahua Dong,Linghe Kong,Xiaokang Yang
Xformer: Hybrid X-Shaped Transformer for Image Denoising
In this paper, we present a hybrid X-shaped vision Transformer, named Xformer, w hich performs notably on image denoising tasks. We explore strengthening the glo bal representation of tokens from different scopes. In detail, we adopt two type s of Transformer blocks. The spatial-wise Transformer block performs fine-graine d local patches interactions across tokens defined by spatial dimension. The cha nnel-wise Transformer block performs direct global context interactions across t okens defined by channel dimension. Based on the concurrent network structure, w e design two branches to conduct these two interaction fashions. Within each bra nch, we employ an encoder-decoder architecture to capture multi-scale features. Besides, we propose the Bidirectional Connection Unit (BCU) to couple the learne d representations from these two branches while providing enhanced information f usion. The joint designs make our Xformer powerful to conduct global information modeling in both spatial and channel dimensions. Extensive experiments show tha t Xformer, under the comparable model complexity, achieves state-of-the-art perf ormance on the synthetic and real-world image denoising tasks. We also provide c ode and models at https://github.com/gladzhang/Xformer.
****
Yu Chen,Yihan Du,Pihe Hu,Siwei Wang,Desheng Wu,Longbo Huang
Provably Efficient Iterated CVaR Reinforcement Learning with Function Approximat ion and Human Feedback
Risk-sensitive reinforcement learning (RL) aims to optimize policies that balanc e the expected reward and risk. In this paper, we present a novel risk-sensitive RL framework that employs an Iterated Conditional Value-at-Risk (CVaR) objectiv e under both linear and general function approximations, enriched by human feedb ack. These new formulations provide a principled way to guarantee safety in each decision making step throughout the control process. Moreover, integrating huma n feedback into risk-sensitive RL framework bridges the gap between algorithmic decision-making and human participation, allowing us to also guarantee safety fo r human-in-the-loop systems. We propose provably sample-efficient algorithms for this Iterated CVaR RL and provide rigorous theoretical analysis. Furthermore, w e establish a matching lower bound to corroborate the optimality of our algorith

ms in a linear context.
****
Hyungi Lee,Giung Nam,Edwin Fong,Juho Lee
Enhancing Transfer Learning with Flexible Nonparametric Posterior Sampling
Transfer learning has recently shown significant performance across various tasks involving deep neural networks. In these transfer learning scenarios, the prior distribution for downstream data becomes crucial in Bayesian model averaging (BMA). While previous works proposed the prior over the neural network parameters centered around the pre-trained solution, such strategies have limitations when dealing with distribution shifts between upstream and downstream data. This paper introduces nonparametric transfer learning (NPTL), a flexible posterior sampling method to address the distribution shift issue within the context of nonparametric learning. The nonparametric learning (NPL) method is a recent approach that employs a nonparametric prior for posterior sampling, efficiently accounting for model misspecification scenarios, which is suitable for transfer learning scenarios that may involve the distribution shift between upstream and downstream tasks. Through extensive empirical validations, we demonstrate that our approach surpasses other baselines in BMA performance.
****
Jingfeng Wu,Difan Zou,Zixiang Chen,Vladimir Braverman,Quanquan Gu,Peter Bartlett
How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression?
Transformers pretrained on diverse tasks exhibit remarkable in-context learning (ICL) capabilities, enabling them to solve unseen tasks solely based on input contexts without adjusting model parameters. In this paper, we study ICL in one of its simplest setups: pretraining a single-layer linear attention model for linear regression with a Gaussian prior. We establish a statistical task complexity bound for the attention model pretraining, showing that effective pretraining only requires a small number of independent tasks. Furthermore, we prove that the pretrained model closely matches the Bayes optimal algorithm, i.e., optimally tuned ridge regression, by achieving nearly Bayes optimal risk on unseen tasks under a fixed context length. These theoretical findings complement prior experimental research and shed light on the statistical foundations of ICL.
****
Ziqiang Li,Hong Sun,Pengfei Xia,Heng Li,Beihao Xia,Yi Wu,Bin Li
Efficient Backdoor Attacks for Deep Neural Networks in Real-world Scenarios
Recent deep neural networks (DNNs) have came to rely on vast amounts of training data, providing an opportunity for malicious attackers to exploit and contaminate the data to carry out backdoor attacks. However, existing backdoor attack methods make unrealistic assumptions, assuming that all training data comes from a single source and that attackers have full access to the training data. In this paper, we introduce a more realistic attack scenario where victims collect data from multiple sources, and attackers cannot access the complete training data. We refer to this scenario as $\textbf{data-constrained backdoor attacks}$. In such cases, previous attack methods suffer from severe efficiency degradation due to the $\textbf{entanglement}$ between benign and poisoning features during the backdoor injection process. To tackle this problem, we introduce three CLIP-based technologies from two distinct streams: $\textit{Clean Feature Suppression}$ and $\textit{Poisoning Feature Augmentation}$. The results demonstrate remarkable improvements, with some settings achieving over $\textbf{100}$% improvement compared to existing attacks in data-constrained scenarios.
****
Na Li,Yuchen Jiao,Hangguan Shan,Shefeng Yan
Provable Memory Efficient Self-Play Algorithm for Model-free Reinforcement Learning
The thriving field of multi-agent reinforcement learning (MARL) studies how a group of interacting agents make decisions autonomously in a shared dynamic environment. Existing theoretical studies in this area suffer from at least two of the following obstacles: memory inefficiency, the heavy dependence of sample complexity on the long horizon and the large state space, the high computational compl

exity, non-Markov policy, non-Nash policy, and high burn-in cost. In this work,
we take a step towards settling this problem by designing a model-free self-play
 algorithm \emph{Memory-Efficient Nash Q-Learning (ME-Nash-QL)} for two-player z
ero-sum Markov games, which is a specific setting of MARL. We prove that ME-Nash
-QL can output an ε-approximate Nash policy with remarkable space co
mplexity $O(SABH)$, sample complexity $\widetilde{O}(H^4SAB/\varepsilon^2)$, and
 computational complexity $O(T\mathrm{poly}(AB))$, where S is the number of st
ates, $\{A, B\}$ is the number of actions for the two players, H is the horizo
n length, and T is the number of samples. Notably, our approach outperforms in
 terms of space complexity compared to existing algorithms for tabular cases. It
 achieves the lowest computational complexity while preserving Markov policies,
setting a new standard. Furthermore, our algorithm outputs a Nash policy and ach
ieves the best sample complexity compared with the existing guarantee for long h
orizons, i.e. when $\min \\{ A, B \\} \ll H^2$. Our algorithm also achieves the
best burn-in cost $O(SAB\,\mathrm{poly}(H))$, whereas previous algorithms need a
t least $O(S^3 AB\,\mathrm{poly}(H))$ to attain the same level of sample complex
ity with ours.
****
Nate Gruver,Anuroop Sriram,Andrea Madotto,Andrew Gordon Wilson,C. Lawrence Zitni
ck,Zachary Ward Ulissi
Fine-Tuned Language Models Generate Stable Inorganic Materials as Text
We propose fine-tuning large language models for generation of stable materials.
 While unorthodox, fine-tuning large language models on text-encoded atomistic d
ata is simple to implement yet reliable, with around 90\% of sampled structures
obeying physical constraints on atom positions and charges. Using energy above h
ull calculations from both learned ML potentials and gold-standard DFT calculati
ons, we show that our strongest model (fine-tuned LLaMA-2 70B) can generate mat
erials predicted to be metastable at about twice the rate (49\% vs 28\%) of CDVA
E, a competing diffusion model. Because of text prompting's inherent flexibility
, our models can simultaneously be used for unconditional generation of stable m
aterial, infilling of partial structures and text-conditional generation. Finall
y, we show that language models' ability to capture key symmetries of crystal st
ructures improves with model scale, suggesting that the biases of pretrained LLM
s are surprisingly well-suited for atomistic data.
****
Jiacheng Chen,Zeyuan Ma,Hongshu Guo,Yining Ma,Jie Zhang,Yue-Jiao Gong
SYMBOL: Generating Flexible Black-Box Optimizers through Symbolic Equation Learn
ing
Recent Meta-learning for Black-Box Optimization (MetaBBO) methods harness neural
 networks to meta-learn configurations of traditional black-box optimizers. Desp
ite their success, they are inevitably restricted by the limitations of predefin
ed hand-crafted optimizers. In this paper, we present SYMBOL, a novel framework
that promotes the automated discovery of black-box optimizers through symbolic e
quation learning. Specifically, we propose a Symbolic Equation Generator (SEG) t
hat allows closed-form optimization rules to be dynamically generated for specif
ic tasks and optimization steps. Within SYMBOL, we then develop three distinct s
trategies based on reinforcement learning, so as to meta-learn the SEG efficient
ly. Extensive experiments reveal that the optimizers generated by SYMBOL not onl
y surpass the state-of-the-art BBO and MetaBBO baselines, but also exhibit excep
tional zero-shot generalization abilities across entirely unseen tasks with diff
erent problem dimensions, population sizes, and optimization horizons. Furthermo
re, we conduct in-depth analyses of our SYMBOL framework and the optimization ru
les that it generates, underscoring its desirable flexibility and interpretabili
ty.
****
Xiaosen Zheng,Tianyu Pang,Chao Du,Jing Jiang,Min Lin
Intriguing Properties of Data Attribution on Diffusion Models
Data attribution seeks to trace model outputs back to training data. With the re
cent development of diffusion models, data attribution has become a desired modu
le to properly assign valuations for high-quality or copyrighted training sample

s, ensuring that data contributors are fairly compensated or credited. Several t
heoretically motivated methods have been proposed to implement data attribution,
 in an effort to improve the trade-off between computational scalability and eff
ectiveness. In this work, we conduct extensive experiments and ablation studies
on attributing diffusion models, specifically focusing on DDPMs trained on CIFAR
-10 and CelebA, as well as a Stable Diffusion model LoRA-finetuned on ArtBench.
Intriguingly, we report counter-intuitive observations that theoretically unjust
ified design choices for attribution empirically outperform previous baselines b
y a large margin, in terms of both linear datamodeling score and counterfactual
evaluation. Our work presents a significantly more efficient approach for attrib
uting diffusion models, while the unexpected findings suggest that at least in n
on-convex settings, constructions guided by theoretical assumptions may lead to
inferior attribution performance. The code is available at https://github.com/sa
il-sg/D-TRAK.
****
Shashanka Venkataramanan,Amir Ghodrati,Yuki M Asano,Fatih Porikli,Amir Habibian
Skip-Attention: Improving Vision Transformers by Paying Less Attention
This work aims to improve the efficiency of vision transformers (ViTs). While V
iTs use computationally expensive self-attention operations in every layer, we i
dentify that these operations are highly correlated across layers -- a key redu
ndancy that causes unnecessary computations. Based on this observation, we propo
se SkipAT a method to reuse self-attention computation from preceding layers to
approximate attention at one or more subsequent layers. To ensure that reusing s
elf-attention blocks across layers does not degrade the performance, we introduc
e a simple parametric function, which outperforms the baseline transformer's per
formance while running computationally faster. We show that SkipAT is agnostic t
o transformer architecture and is effective in image classification, semantic s
egmentation on ADE20K, image denoising on SIDD, and video denoising on DAVIS. We
 achieve improved throughput at the same-or-higher accuracy levels in all these
tasks.
****
Seamus Somerstep,Yuekai Sun,Yaacov Ritov
Learning in reverse causal strategic environments with ramifications on two side
d markets
Motivated by equilibrium models of labor markets, we develop a formulation of ca
usal strategic classification in which strategic agents can directly manipulate
their outcomes. As an application, we consider employers that seek to anticipate
 the strategic response of a labor force when developing a hiring policy. We sho
w theoretically that employers with performatively optimal hiring policies impro
ve employer reward, labor force skill level, and labor force equity (compared to
 employers that do not anticipate the strategic labor force response) in the cla
ssic Coate-Loury labor market model. Empirically, we show that these desirable p
roperties of performative hiring policies do generalize to our own formulation o
f a general equilibrium labor market. On the other hand, we also observe that th
e benefits of performatively optimal hiring policies are brittle in some aspects
. We demonstrate that in our formulation a performative employer both harms work
ers by reducing their aggregate welfare and fails to prevent discrimination when
 more sophisticated wage and cost structures are introduced.
****
Yue Deng,Wenxuan Zhang,Sinno Jialin Pan,Lidong Bing
Multilingual Jailbreak Challenges in Large Language Models
While large language models (LLMs) exhibit remarkable capabilities across a wide
 range of tasks, they pose potential safety concerns, such as the ``jailbreak''
problem, wherein malicious instructions can manipulate LLMs to exhibit undesirab
le behavior. Although several preventive measures have been developed to mitigat
e the potential risks associated with LLMs, they have primarily focused on Engli
sh. In this study, we reveal the presence of multilingual jailbreak challenges w
ithin LLMs and consider two potential risky scenarios: unintentional and intenti
onal. The unintentional scenario involves users querying LLMs using non-English
prompts and inadvertently bypassing the safety mechanisms, while the intentional

scenario concerns malicious users combining malicious instructions with multilingual prompts to deliberately attack LLMs. The experimental results reveal that in the unintentional scenario, the rate of unsafe content increases as the availability of languages decreases. Specifically, low-resource languages exhibit about three times the likelihood of encountering harmful content compared to high-resource languages, with both ChatGPT and GPT-4. In the intentional scenario, multilingual prompts can exacerbate the negative impact of malicious instructions, with astonishingly high rates of unsafe output: 80.92\% for ChatGPT and 40.71\% for GPT-4. To handle such a challenge in the multilingual context, we propose a novel \textsc{Self-Defense} framework that automatically generates multilingual training data for safety fine-tuning. Experimental results show that ChatGPT fine-tuned with such data can achieve a substantial reduction in unsafe content generation. Data is available at \url{https://github.com/DAMO-NLP-SG/multilingual-safety-for-LLMs}.
****
Qin ZHANG,Linghan Xu,Jun Fang,Qingming Tang,Ying Nian Wu,Joseph Tighe,Yifan Xing
Threshold-Consistent Margin Loss for Open-World Deep Metric Learning
Existing losses used in deep metric learning (DML) for image retrieval often lead to highly non-uniform intra-class and inter-class representation structures across test classes and data distributions. When combined with the common practice of using a fixed threshold to declare a match, this gives rise to significant performance variations in terms of false accept rate (FAR) and false reject rate (FRR) across test classes and data distributions. We define this issue in DML as threshold inconsistency. In real-world applications, such inconsistency often complicates the threshold selection process when deploying large-scale image retrieval systems. To measure this inconsistency, we propose a novel variance-based metric called Operating-Point-Inconsistency-Score (OPIS) that quantifies the variance in the operating characteristics across classes. Using the OPIS metric, we find that achieving high accuracy levels in a DML model does not automatically guarantee threshold consistency. In fact, our investigation reveals a Pareto frontier in the high-accuracy regime, where existing methods to improve accuracy often lead to degradation in threshold consistency. To address this trade-off, we introduce the Threshold-Consistent Margin (TCM) loss, a simple yet effective regularization technique that promotes uniformity in representation structures across classes by selectively penalizing hard sample pairs. Large-scale experiments demonstrate TCM's effectiveness in enhancing threshold consistency while preserving accuracy, simplifying the threshold selection process in practical DML settings.
****
Aaron Spieler,Nasim Rahaman,Georg Martius,Bernhard Schölkopf,Anna Levina
The Expressive Leaky Memory Neuron: an Efficient and Expressive Phenomenological Neuron Model Can Solve Long-Horizon Tasks.
Biological cortical neurons are remarkably sophisticated computational devices, temporally integrating their vast synaptic input over an intricate dendritic tree, subject to complex, nonlinearly interacting internal biological processes.
A recent study proposed to characterize this complexity by fitting accurate surrogate models to replicate the input-output relationship of a detailed biophysical cortical pyramidal neuron model and discovered it needed temporal convolutional networks (TCN) with millions of parameters.
Requiring these many parameters, however, could stem from a misalignment between the inductive biases of the TCN and cortical neuron's computations.
In light of this, and to explore the computational implications of leaky memory units and nonlinear dendritic processing, we introduce the Expressive Leaky Memory (ELM) neuron model, a biologically inspired phenomenological model of a cortical neuron.
Remarkably, by exploiting such slowly decaying memory-like hidden states and two-layered nonlinear integration of synaptic input, our ELM neuron can accurately match the aforementioned input-output relationship with under ten thousand trainable parameters.
To further assess the computational ramifications of our neuron design, we evalu

ate it on various tasks with demanding temporal structures, including the Long R
ange Arena (LRA) datasets, as well as a novel neuromorphic dataset based on the
Spiking Heidelberg Digits dataset (SHD-Adding). Leveraging a larger number of me
mory units with sufficiently long timescales, and correspondingly sophisticated
synaptic integration, the ELM neuron displays substantial long-range processing
capabilities, reliably outperforming the classic Transformer or Chrono-LSTM arch
itectures on LRA, and even solving the Pathfinder-X task with over 70\% accuracy
 (16k context length). These findings raise further questions about the computat
ional sophistication of individual cortical neurons and their role in extracting
 complex long-range temporal dependencies.
****
Hansheng Xue,Vijini Mallawaarachchi,Lexing Xie,Vaibhav Rajan
Encoding Unitig-level Assembly Graphs with Heterophilous Constraints for Metagen
omic Contigs Binning
Metagenomics studies genomic material derived from mixed microbial communities i
n diverse environments, holding considerable significance for both human health
and environmental sustainability. Metagenomic binning refers to the clustering o
f genomic subsequences obtained from high-throughput DNA sequencing into distinc
t bins, each representing a constituent organism within the community. Mainstrea
m binning methods primarily rely on sequence features such as composition and ab
undance, making them unable to effectively handle sequences shorter than 1,000 b
p and inherent noise within sequences. Several binning tools have emerged, aimin
g to enhance binning outcomes by using the assembly graph generated by assembler
s, which encodes valuable overlapping information among genomic sequences. Howev
er, existing assembly graph-based binners mainly focus on simplified contig-leve
l assembly graphs that are recreated from assembler’s original graphs, unitig-le
vel assembly graphs. The simplification reduces the resolution of the connectivi
ty information in original graphs. In this paper, we design a novel binning tool
 named UnitigBin, which leverages representation learning on unitig-level assemb
ly graphs while adhering to heterophilious constraints imposed by single-copy ma
rker genes, ensuring that constrained contigs cannot be grouped together. Extens
ive experiments conducted on synthetic and real datasets demonstrate that Unitig
Bin significantly surpasses state-of-the-art binning tools.
****
Xiangyu Zeng,Jie Lin,Piao Hu,Ruizheng Huang,Zhicheng Zhang
A Framework for Inference Inspired by Human Memory Mechanisms
How humans and machines make sense of current inputs for relation reasoning and
question-answering while putting the perceived information into context of our p
ast memories, has been a challenging conundrum in cognitive science and artifici
al intelligence. Inspired by human brain's memory system and cognitive architect
ures, we propose a PMI framework that consists of perception, memory and inferen
ce components. Notably, the memory module comprises working and long-term memory
, with the latter endowed with a higher-order structure to retain extensive and
complex relational knowledge and experience. Through a differentiable competitiv
e write access, current perceptions update working memory, which is later merged
 with long-term memory via outer product associations, reducing information conf
licts and averting memory overflow. In the inference module, relevant informatio
n is retrieved from two separate memory origins and associatively integrated to
attain a more comprehensive and precise interpretation of current perceptions. W
e exploratively apply our PMI to improve prevailing Transformers and CNN models
on question-answering tasks like bAbI-20k and Sort-of-CLEVR datasets, as well as
 detecting equilateral triangles, language modeling and image classification tas
ks, and in each case, our PMI enhancements consistently outshine their original
counterparts significantly. Visualization analyses reveal that relational memory
 consolidation, along with the interaction and integration of information from d
iverse memory sources, substantially contributes to the model effectiveness on i
nference tasks.
****
Zihan Ding,Chi Jin
Consistency Models as a Rich and Efficient Policy Class for Reinforcement Learni

ng
Score-based generative models like the diffusion model have been testified to be
 effective in modeling multi-modal data from image generation to reinforcement l
earning (RL). However, the inference process of diffusion model can be slow, whi
ch hinders its usage in RL with iterative sampling. We propose to apply the cons
istency model as an efficient yet expressive policy representation, namely consi
stency policy, with an actor-critic style algorithm for three typical RL setting
s: offline, offline-to-online and online. For offline RL, we demonstrate the exp
ressiveness of generative models as policies from multi-modal data. For offline-
to-online RL, the consistency policy is shown to be more computational efficient
 than diffusion policy, with a comparable performance. For online RL, the consis
tency policy demonstrates significant speedup and even higher average performanc
es than the diffusion policy.
****
Hunter Lightman,Vineet Kosaraju,Yuri Burda,Harrison Edwards,Bowen Baker,Teddy Le
e,Jan Leike,John Schulman,Ilya Sutskever,Karl Cobbe
Let's Verify Step by Step
In recent years, large language models have greatly improved in their ability to
 perform complex multi-step reasoning. However, even state-of-the-art models sti
ll regularly produce logical mistakes. To train more reliable models, we can tur
n either to outcome supervision, which provides feedback for a final result, or
process supervision, which provides feedback for each intermediate reasoning ste
p. Given the importance of training reliable models, and given the high cost of
human feedback, it is important to carefully compare the both methods. Recent wo
rk has already begun this comparison, but many questions still remain. We conduc
t our own investigation, finding that process supervision significantly outperfo
rms outcome supervision for training models to solve problems from the challengi
ng MATH dataset. Our process-supervised model solves 78% of problems from a repr
esentative subset of the MATH test set. Additionally, we show that active learni
ng significantly improves the efficacy of process supervision. To support relate
d research, we also release PRM800K, the complete dataset of 800,000 step-level
human feedback labels used to train our best reward model.
****
Yubo Zhuang,Xiaohui Chen,Yun Yang,Richard Y. Zhang
Statistically Optimal K-means Clustering via Nonnegative Low-rank Semidefinite
 Programming
K-means clustering is a widely used machine learning method for identifying pa
tterns in large datasets.
    Semidefinite programming (SDP) relaxations have recently been proposed for s
olving the K-means optimization problem that enjoy strong statistical optimal
ity guarantees, but the prohibitive cost of implementing an SDP solver renders t
hese guarantees inaccessible to practical datasets.
    By contrast, nonnegative matrix factorization (NMF) is a simple clustering a
lgorithm that is widely used by machine learning practitioners, but without a so
lid statistical underpinning nor rigorous guarantees. In this paper, we describe
 an NMF-like algorithm that works by solving a \emph{nonnegative} low-rank restr
iction of the SDP relaxed K-means formulation using a nonconvex Burer--Monteir
o factorization approach. The resulting algorithm is just as simple and scalable
 as state-of-the-art NMF algorithms, while also enjoying the same strong statist
ical optimality guarantees as the SDP.
    In our experiments, we observe that our algorithm achieves substantially sma
ller mis-clustering errors compared to the existing state-of-the-art.
****
Oren Mangoubi,Nisheeth K. Vishnoi
Faster Sampling from Log-Concave Densities over Polytopes via Efficient Linear S
olvers
We consider the problem of sampling from a logconcave distribution $\pi(\theta)
\propto e^{-f(\theta)}$ constrained to a polytope $K:=${$\theta \in \mathbb{R}^d
: A\theta \leq b$}, where $A\in   \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^
m$. The fastest-known algorithm for the setting when f is $O(1)$-Lipschitz or

$O(1)$-smooth runs in roughly $O(md \times md^{\omega -1})$ arithmetic operation
s, where the $md^{\omega -1}$ term arises because each Markov chain step require
s computing a matrix inversion and determinant ($\omega \approx 2.37$ is the mat
rix multiplication constant). We present a nearly-optimal implementation of this
 Markov chain with per-step complexity that is roughly the number of non-zero en
tries of A while the number of Markov chain steps remains the same. The key te
chnical ingredients are 1) to show that the matrices that arise in this Dikin wa
lk change slowly, 2) to deploy efficient linear solvers which can leverage this
slow change to speed up matrix inversion by using information computed in previo
us steps, and 3) to speed up the computation of the determinantal term in the Me
tropolis filter step via a randomized Taylor series-based estimator. This result
 directly improves the runtime for applications that involve sampling from Gibbs
 distributions constrained to polytopes that arise in Bayesian statistics and pr
ivate optimization.
****
Kevin Yang,Dan Klein,Asli Celikyilmaz,Nanyun Peng,Yuandong Tian
RLCD: Reinforcement Learning from Contrastive Distillation for LM Alignment
We propose Reinforcement Learning from Contrastive Distillation (RLCD), a method
 for aligning language models to follow principles expressed in natural language
 (e.g., to be more harmless) without using human feedback. RLCD creates preferen
ce pairs from two contrasting model outputs, one using a positive prompt designe
d to encourage following the given principles, and one using a negative prompt d
esigned to encourage violating them. Using two different prompts causes model ou
tputs to be more differentiated on average, resulting in cleaner preference labe
ls in the absence of human annotations. We then use the preference pairs to trai
n a preference model, which is in turn used to improve a base unaligned language
 model via reinforcement learning. Empirically, RLCD outperforms RLAIF (Bai et a
l., 2022b) and context distillation (Huang et al., 2022) baselines across three
diverse alignment tasks—harmlessness, helpfulness, and story outline generation—
and when using both 7B and 30B model scales for simulating preference data
****
Jonas Belouadi,Anne Lauscher,Steffen Eger
AutomaTikZ: Text-Guided Synthesis of Scientific Vector Graphics with TikZ
Generating bitmap graphics from text has gained considerable attention, yet for
scientific figures, vector graphics are often preferred. Given that vector graph
ics are typically encoded using low-level graphics primitives, generating them d
irectly is difficult. To address this, we propose the use of TikZ, a well-known
abstract graphics language that can be compiled to vector graphics, as an interm
ediate representation of scientific figures. TikZ offers human-oriented, high-le
vel commands, thereby facilitating conditional language modeling with any large
language model. To this end, we introduce DaTikZ the first large-scale TikZ data
set, consisting of 120k TikZ drawings aligned with captions. We fine-tune LLaMA
on DaTikZ, as well as our new model CLiMA, which augments LLaMA with multimodal
CLIP embeddings. In both human and automatic evaluation, CLiMA and LLaMA outperf
orm commercial GPT-4 and Claude 2 in terms of similarity to human-created figure
s, with CLiMA additionally improving text-image alignment. Our detailed analysis
 shows that all models generalize well and are not susceptible to memorization.
GPT-4 and Claude 2, however, tend to generate more simplistic figures compared t
o both humans and our models. We make our framework, AutomaTikZ, along with mode
l weights and datasets, publicly available.
****
Sean Kulinski,Zeyu Zhou,Ruqi Bai,Murat Kocaoglu,David I. Inouye
Towards Characterizing Domain Counterfactuals for Invertible Latent Causal Model
s
Answering counterfactual queries has many important applications such as knowled
ge discovery and explainability, but is challenging when causal variables are un
observed and we only see a projection onto an observation space, for instance, i
mage pixels.
    One approach is to recover the latent Structural Causal Model (SCM), but thi
s typically needs unrealistic assumptions, such as linearity of the causal mecha

nisms.
Another approach is to use naïve ML approximations, such as generative models, to generate counterfactual samples; however, these lack guarantees of accuracy.
In this work, we strive to strike a balance between practicality and theoretical guarantees by focusing on a specific type of causal query called *domain counterfactuals*, which hypothesizes what a sample would have looked like if it had been generated in a different domain (or environment).
Concretely, by only assuming invertibility, sparse domain interventions and access to observational data from different domains, we aim to improve domain counterfactual estimation both theoretically and practically with less restrictive assumptions.
We define *domain counterfactually equivalent* models and prove necessary and sufficient properties for equivalent models that provide a tight characterization of the domain counterfactual equivalence classes.
Building upon this result, we prove that every equivalence class contains a model where all intervened variables are at the end when topologically sorted by the causal DAG, i.e., all non-intervened variables have non-intervened ancestors.
This surprising result suggests that a model design that only allows intervention in the last k latent variables may improve model estimation for counterfactuals.
We then test this model design on extensive simulated and image-based experiments which show the sparse canonical model indeed improves counterfactual estimation over baseline non-sparse models.
****
Josh Alman,Zhao Song
How to Capture Higher-order Correlations? Generalizing Matrix Softmax Attention to Kronecker Computation
In the classical transformer attention scheme, we are given three $n \times d$ size matrices Q, K, V (the query, key, and value tokens), and the goal is to compute a new $n \times d$ size matrix $D^{-1} \exp(QK^\top) V$ where $D = \mathrm{diag}(\exp(QK^\top) {\bf 1}_n)$. Here, $\exp()$ is applied entry-wise and ${\bf 1}_n$ denotes a length-n vector whose entries are all ones.

Intuitively, attention computation captures pairwise information between words in a sentence, but not higher-order information. Indeed, recent work \cite{sht23} has shown that attention units cannot solve simple problems about detecting triples of connected words.

In this work, we study a generalization of attention which captures triple-wise correlations. The generalization is based on computations involving tensors defined by tuples of words. More formally, given five $n \times d$ size matrices Q, K_1, K_2, V_1 and V_2 (generalized query, key, and value tokens), our new goal is to compute an $n \times d$ size matrix $D^{-1} \exp(Q (K_1 \oslash K_2)^\top) (V_1 \oslash V_2) $ where $D = \mathrm{diag}(\exp(Q (K_1 \oslash K_2)^\top) {\bf 1}_{n^2})$ and $K_1 \oslash K_2 \in \mathbb{R}^{n^2 \times d}$ denotes the column-wise Kronecker product of K_1 and K_2. This generalization is indeed able to solve problems about detecting triple-wise connections that were shown to be impossible for transformers.

The potential downside of this generalization is that it appears as though computations are even more difficult, since the straightforward algorithm requires cubic time in n. However, we show that in the bounded-entry setting (which arises in practice, and which is well-studied in both theory and practice), there is actually a near-linear time algorithm. More precisely, we show that bounded entries are both necessary and sufficient for quickly performing generalized computations:

$\bullet$ On the positive side, if all entries of the input matrices are bounded

 above by $o(\sqrt[3]{\log n})$ then we show how to approximate the ``tensor-typ
e'' attention matrix in $n^{1+o(1)}$ time.

$\bullet$ On the negative side, we show that if the entries of the input matrice
s may be as large as $\Omega(\sqrt[3]{\log n})$, then there is no algorithm that
 runs faster than $n^{3-o(1)}$ (assuming the Strong Exponential
Time Hypothesis from fine-grained complexity theory).

We also show that our construction, algorithms, and lower bounds naturally gener
alize to higher-order tensors and correlations. Interestingly, the higher the or
der of the tensors, the lower the bound on the entries needs to be for an effici
ent algorithm. Our results thus yield a natural tradeoff between the boundedness
 of the entries, and order of the tensor one may use for more expressive, effici
ent attention computation.

Our constructions make use of a novel connection with a higher-order variant on
the kernel density estimation problem. They combine a number of technical tools,
 including the polynomial method, algebraic geometry codes, and multiparty Merli
n-Arthur communication protocols.
****
Ruinan Jin,Shuai Li,Baoxiang Wang
On Stationary Point Convergence of PPO-Clip
Proximal policy optimization (PPO) has gained popularity in reinforcement learni
ng (RL). Its PPO-Clip variant is one the most frequently implemented algorithms
and is one of the first-to-try algorithms in RL tasks. This variant uses a clipp
ed surrogate objective function not typically found in other algorithms. Many wo
rks have demonstrated the practical performance of PPO-Clip, but the theoretical
 understanding of it is limited to specific settings. In this work, we provide a
 comprehensive analysis that shows the stationary point convergence of PPO-Clip
and the convergence rate thereof. Our analysis is new and overcomes many challen
ges, including the non-smooth nature of the clip operator, the potentially unbou
nded score function, and the involvement of the ratio of two stochastic policies
. Our results and techniques might share new insights into PPO-Clip.
****
Prateek Chanda,Shrey Modi,Ganesh Ramakrishnan
Bayesian Coreset Optimization for Personalized Federated Learning
In a distributed machine learning setting like Federated Learning where there ar
e multiple clients involved which update their individual weights to a single ce
ntral server, often training on the entire individual client's dataset for each
client becomes cumbersome. To address this issue we propose CORESET-PFEDBAYES: a
 personalized coreset weighted federated learning setup where the training updat
es for each individual clients are forwarded to the central server based on only
 individual client coreset based representative data points instead of the entir
e client data. Through theoretical analysis we present how the average generaliz
ation error is minimax optimal up to logarithm bounds $\mathcal{O}(n_k^{-\frac{2
 \beta}{2 \beta+d}} \log ^{2 \delta^{\prime}}(n_k))$, where n_k denotes the co
reset size and how the approximation error on the data likelihood differs from a
 vanilla Federated Learning setup as a function $G(\boldsymbol{w})$ of the cores
et weights $\boldsymbol{w}$.
Our experiments on different benchmark datasets based on a variety of recent per
sonalized federated learning architectures show significant gains (+4.87\% on MN
IST, +8.61\% on FashionMNIST, +9.71\% on CIFAR in terms of model accuracy acro
ss) as compared to random sampling on the training data followed by federated l
earning, thereby indicating how intelligently selecting such training samples ca
n help in performance. Additionally, through experiments on medical datasets our
 proposed method showcases some gains (e.g. +9.74\% under COVID-19 dataset) as c
ompared to other submodular optimization based approaches used for subset selec
tion on client's data.
****

Konstantin Hess,Valentyn Melnychuk,Dennis Frauen,Stefan Feuerriegel
Bayesian Neural Controlled Differential Equations for Treatment Effect Estimation
Treatment effect estimation in continuous time is crucial for personalized medicine. However, existing methods for this task are limited to point estimates of the potential outcomes, whereas uncertainty estimates have been ignored. Needless to say, uncertainty quantification is crucial for reliable decision-making in medical applications. To fill this gap, we propose a novel Bayesian neural controlled differential equation (BNCDE) for treatment effect estimation in continuous time. In our BNCDE, the time dimension is modeled through a coupled system of neural controlled differential equations and neural stochastic differential equations, where the neural stochastic differential equations allow for tractable variational Bayesian inference. Thereby, for an assigned sequence of treatments, our BNCDE provides meaningful posterior predictive distributions of the potential outcomes. To the best of our knowledge, ours is the first tailored neural method to provide uncertainty estimates of treatment effects in continuous time. As such, our method is of direct practical value for promoting reliable decision-making in medicine.
****
Ilan Price,Nicholas Daultry Ball,Adam Christopher Jones,Samuel Chun Hei Lam,Jared Tanner
DEEP NEURAL NETWORK INITIALIZATION WITH SPARSITY INDUCING ACTIVATIONS
Inducing and leveraging sparse activations during training and inference is a promising avenue for improving the computational efficiency of deep networks, which is increasingly important as network sizes continue to grow and their application becomes more widespread. Here we use the large width Gaussian process limit to analyze the behaviour, at random initialization, of nonlinear activations that induce sparsity in the hidden outputs. A previously unreported form of training instability is proven for arguably two of the most natural candidates for hidden layer sparsification; those being a shifted ReLU ($\phi(x)=\max(0, x-\tau)$ for $\tau\ge 0$) and soft thresholding ($\phi(x)=0$ for $|x|\le\tau$ and $x-\text{sign}(x)\tau$ for $|x|>\tau$). We show that this instability is overcome by clipping the nonlinear activation magnitude, at a level prescribed by the shape of the associated Gaussian process variance map. Numerical experiments verify the theory and show that the proposed magnitude clipped sparsifying activations can be trained with training and test fractional sparsity as high as 85\% while retaining close to full accuracy.
****
Anson Bastos,Kuldeep Singh,Abhishek Nadgeri,Manish Singh,Toyotaro Suzumura
Beyond Spatio-Temporal Representations: Evolving Fourier Transform for Temporal Graphs
We present the Evolving Graph Fourier Transform (EFT), the first invertible spectral transform that captures evolving representations on temporal graphs. We motivate our work by the inadequacy of existing methods for capturing the evolving graph spectra, which are also computationally expensive due to the temporal aspect along with the graph vertex domain. We view the problem as an optimization over the Laplacian of the continuous time dynamic graph. Additionally, we propose pseudo-spectrum relaxations that decompose the transformation process, making it highly computationally efficient. The EFT method adeptly captures the evolving graph's structural and positional properties, making it effective for downstream tasks on evolving graphs. Hence, as a reference implementation, we develop a simple neural model induced with \eft for capturing evolving graph spectra. We empirically validate our theoretical findings on a number of large-scale and standard temporal graph benchmarks and demonstrate that our model achieves state-of-the-art performance.
****
Sheng Jin,Xueying Jiang,Jiaxing Huang,Lewei Lu,Shijian Lu
LLMs Meet VLMs: Boost Open Vocabulary Object Detection with Fine-grained Descriptors
Inspired by the outstanding zero-shot capability of vision language models (VLMs

) in image classification tasks, open-vocabulary object detection has attracted increasing interest by distilling the broad VLM knowledge into detector training . However, most existing open-vocabulary detectors learn by aligning region embeddings with categorical labels (e.g., bicycle) only, disregarding the capability of VLMs on aligning visual embeddings with fine-grained text descriptions of object parts (e.g., pedals and bells). This paper presents DVDet, a Descriptor-Enhanced Open Vocabulary Detector that introduces conditional context prompts and hierarchical textual descriptors that enable precise region-text alignment as well as open-vocabulary detection training in general. Specifically, the conditional context prompt transforms regional embeddings into image-like representations that can be directly integrated into general open vocabulary detection training. In addition, we introduce large language models as an interactive and implicit knowledge repository which enables iterative mining and refining visually oriented textual descriptors for precise region-text alignment. Extensive experiments over multiple large-scale benchmarks show that DVDet outperforms the state-of-the-art consistently by large margins.
****
Junyan Cheng,Peter Chin
Bridging Neural and Symbolic Representations with Transitional Dictionary Learning
This paper introduces a novel Transitional Dictionary Learning (TDL) framework that can implicitly learn symbolic knowledge, such as visual parts and relations, by reconstructing the input as a combination of parts with implicit relations. We propose a game-theoretic diffusion model to decompose the input into visual parts using the dictionaries learned by the Expectation Maximization (EM) algorithm, implemented as the online prototype clustering, based on the decomposition results. Additionally, two metrics, clustering information gain, and heuristic shape score are proposed to evaluate the model. Experiments are conducted on three abstract compositional visual object datasets, which require the model to utilize the compositionality of data instead of simply exploiting visual features. Then, three tasks on symbol grounding to predefined classes of parts and relations, as well as transfer learning to unseen classes, followed by a human evaluation, were carried out on these datasets. The results show that the proposed method discovers compositional patterns, which significantly outperforms the state-of-the-art unsupervised part segmentation methods that rely on visual features from pre-trained backbones. Furthermore, the proposed metrics are consistent with human evaluations.
****
Xu Zheng,Farhad Shirani,Tianchun Wang,Wei Cheng,Zhuomin Chen,Haifeng Chen,Hua Wei,Dongsheng Luo
Towards Robust Fidelity for Evaluating Explainability of Graph Neural Networks
Graph Neural Networks (GNNs) are neural models that leverage the dependency structure in graphical data via message passing among the graph nodes. GNNs have emerged as pivotal architectures in analyzing graph-structured data, and their expansive application in sensitive domains requires a comprehensive understanding of their decision-making processes --- necessitating a framework for GNN explainability. An explanation function for GNNs takes a pre-trained GNN along with a graph as input, to produce a `sufficient statistic' subgraph with respect to the graph label. A main challenge in studying GNN explainability is to provide fidelity measures that evaluate the performance of these explanation functions. This paper studies this foundational challenge, spotlighting the inherent limitations of prevailing fidelity metrics, including Fid_+, Fid_-, and Fid_Δ. Specifically, a formal, information-theoretic definition of explainability is introduced and it is shown that existing metrics often fail to align with this definition across various statistical scenarios. The reason is due to potential distribution shifts when subgraphs are removed in computing these fidelity measures. Subsequently, a robust class of fidelity measures are introduced, and it is shown analytically that they are resilient to distribution shift issues and are applicable in a wide range of scenarios. Extensive empirical analysis on both synthetic and real datasets are provided to illustrate that the proposed metrics are mo

re coherent with gold standard metrics.
****
Shaofei Cai,Bowei Zhang,Zihao Wang,Xiaojian Ma,Anji Liu,Yitao Liang
GROOT: Learning to Follow Instructions by Watching Gameplay Videos
We study the problem of building a controller that can follow open-ended instructions in open-world environments. We propose to follow reference videos as instructions, which offer expressive goal specifications while eliminating the need for expensive text-gameplay annotations. A new learning framework is derived to allow learning such instruction-following controllers from gameplay videos while producing a video instruction encoder that induces a structured goal space. We implement our agent GROOT in a simple yet effective encoder-decoder architecture based on causal transformers. We evaluate GROOT against open-world counterparts and human players on a proposed Minecraft SkillForge benchmark. The Elo ratings clearly show that GROOT is closing the human-machine gap as well as exhibiting a 70% winning rate over the best generalist agent baseline. Qualitative analysis of the induced goal space further demonstrates some interesting emergent properties, including the goal composition and complex gameplay behavior synthesis.
****
Woomin Song,Seunghyuk Oh,Sangwoo Mo,Jaehyung Kim,Sukmin Yun,Jung-Woo Ha,Jinwoo Shin
Hierarchical Context Merging: Better Long Context Understanding for Pre-trained LLMs
Large language models (LLMs) have established new standards in various natural language processing tasks.
However, a primary constraint they face is the context limit, i.e., the maximum number of tokens they can process.
To relax the constraint,
previous works have explored architectural changes and modifications in positional encoding, but they often require expensive training or do not address the computational demands of self-attention.
In this paper, we present Hierarchical cOntext MERging (HOMER), a new training-free scheme designed to overcome the limitations. HOMER harnesses a divide-and-conquer methodology, segmenting extensive inputs into manageable units. The segments are then processed collectively, employing a hierarchical strategy that fuses adjacent chunks at progressive transformer layers. A token reduction technique precedes each fusion, ensuring memory usage efficiency.
We also propose an optimized computational order reducing the memory requirement to logarithmically scale with respect to input length, making it especially favorable for environments with tight memory restrictions.
Our experimental results demonstrate the superior performance and memory efficiency of the proposed method, opening doors for broader applications of LLMs in scenarios with extended context requirements.
Code is available at [this https URL](https://github.com/alinlab/HOMER).
****
Zijie Pan,Jiachen Lu,Xiatian Zhu,Li Zhang
Enhancing High-Resolution 3D Generation through Pixel-wise Gradient Clipping
High-resolution 3D object generation remains a challenging task primarily due to the limited availability of comprehensive annotated training data. Recent advancements have aimed to overcome this constraint by harnessing image generative models, pretrained on extensive curated web datasets, using knowledge transfer techniques like Score Distillation Sampling (SDS).
Efficiently addressing the requirements of high-resolution rendering often necessitates the adoption of latent representation-based models, such as the Latent Diffusion Model (LDM). In this framework, a significant challenge arises:
To compute gradients for individual image pixels, it is necessary to backpropagate gradients from the designated latent space through the frozen components of the image model, such as the VAE encoder used within LDM. However, this gradient propagation pathway has never been optimized, remaining uncontrolled during training.
We find that the unregulated gradients adversely affect the 3D model's capacity

in acquiring texture-related information from the image generative model,
leading to poor quality appearance synthesis.
To address this overarching challenge, we propose an innovative operation termed Pixel-wise Gradient Clipping (PGC) designed for seamless integration into existing 3D generative models, thereby enhancing their synthesis quality. Specifically,
we control the magnitude of stochastic gradients by clipping the pixel-wise gradients efficiently,
while preserving crucial texture-related gradient directions.
Despite this simplicity and minimal extra cost, extensive experiments demonstrate the efficacy of our PGC
in enhancing the performance of existing 3D generative models
for high-resolution object rendering.
****
Rares C Cristian,Georgia Perakis
A Discretization Framework for Robust Contextual Stochastic Optimization
We study contextual stochastic optimization problems. Optimization problems have uncertain parameters stemming from unknown, context-dependent, distributions. Due to the inherent uncertainty in these problems, one is often interested not only in minimizing expected cost, but also to be robust and protect against worst case scenarios. We propose a novel method that combines the learning stage with knowledge of the downstream optimization task. The method prescribes decisions which aim to maximize the likelihood that the cost is below a (user-controlled) threshold. The key idea is (1) to discretize the feasible region into subsets so that the uncertain objective function can be well approximated deterministically within each subset, and (2) devise a secondary optimization problem to prescribe decisions by integrating the individual approximations determined in step (1). We provide theoretical guarantees bounding the underlying regret of decisions proposed by our method. In addition, experimental results demonstrate that our approach is competitive in terms of average regret and yields more robust solutions than other methods proposed in the literature, including up to 20 times lower worst-case cost on a real-world electricity generation problem.
****
Albert Xu,Jhih-Yi Hsieh,Bhaskar Vundurthy,Nithya Kemp,Eliana Cohen,Lu Li,Howie Choset
Mathematical Justification of Hard Negative Mining via Isometric Approximation Theorem
In deep metric learning, the triplet loss has emerged as a popular method to learn many computer vision and natural language processing tasks such as facial recognition, object detection, and visual-semantic embeddings. One issue that plagues the triplet loss is network collapse, an undesirable phenomenon where the network projects the embeddings of all data onto a single point. Researchers predominately solve this problem by using triplet mining strategies. While hard negative mining is the most effective of these strategies, existing formulations lack strong theoretical justification for their empirical success. In this paper, we utilize the mathematical theory of isometric approximation to show an equivalence between the triplet loss sampled by hard negative mining and an optimization problem that minimizes a Hausdorff-like distance between the neural network and its ideal counterpart function. This provides the theoretical justifications for hard negative mining's empirical efficacy. Experiments performed on the Market-1501 and Stanford Online Products datasets with various network architectures corroborate our theoretical findings, indicating that network collapse tends to happen when batch size is too large or embedding dimension is too small. In addition, our novel application of the isometric approximation theorem provides the groundwork for future forms of hard negative mining that avoid network collapse.
****
T Mitchell Roddenberry,Vishwanath Saragadam,Maarten V. de Hoop,Richard Baraniuk
Implicit Neural Representations and the Algebra of Complex Wavelets
Implicit neural representations (INRs) have arisen as useful methods for representing signals on Euclidean domains. By parameterizing an image as a multilayer p

erceptron (MLP) on Euclidean space, INRs effectively couple spatial and spectral features of the represented signal in a way that is not obvious in the usual discrete representation. Although INRs using sinusoidal activation functions have been studied in terms of Fourier theory, recent works have shown the advantage of using wavelets instead of sinusoids as activation functions, due to their ability to simultaneously localize in both frequency and space. In this work, we approach such INRs and demonstrate how they resolve high-frequency features of signals from coarse approximations performed in the first layer of the MLP. This leads to multiple prescriptions for the design of INR architectures, including the use of progressive wavelets, decoupling of low and high-pass approximations, and initialization schemes based on the singularities of the target signal.
****
Lingxuan Wu,Xiao Yang,Yinpeng Dong,Liuwei XIE,Hang Su,Jun Zhu
Embodied Active Defense: Leveraging Recurrent Feedback to Counter Adversarial Patches
The vulnerability of deep neural networks to adversarial patches has motivated numerous defense strategies for boosting model robustness. However, the prevailing defenses depend on single observation or pre-established adversary information to counter adversarial patches, often failing to be confronted with unseen or adaptive adversarial attacks and easily exhibiting unsatisfying performance in dynamic 3D environments. Inspired by active human perception and recurrent feedback mechanisms, we develop Embodied Active Defense (EAD), a proactive defensive strategy that actively contextualizes environmental information to address misaligned adversarial patches in 3D real-world settings. To achieve this, EAD develops two central recurrent sub-modules, i.e., a perception module and a policy module, to implement two critical functions of active vision. These models recurrently process a series of beliefs and observations, facilitating progressive refinement of their comprehension of the target object and enabling the development of strategic actions to counter adversarial patches in 3D environments. To optimize learning efficiency, we incorporate a differentiable approximation of environmental dynamics and deploy patches that are agnostic to the adversary’s strategies. Extensive experiments demonstrate that EAD substantially enhances robustness against a variety of patches within just a few steps through its action policy in safety-critical tasks (e.g., face recognition and object detection), without compromising standard accuracy. Furthermore, due to the attack-agnostic characteristic, EAD facilitates excellent generalization to unseen attacks, diminishing the averaged attack success rate by 95% across a range of unseen adversarial attacks.
****
Xun Wu,Shaohan Huang,Furu Wei
MoLE: Mixture of LoRA Experts
LoRA has gained widespread acceptance in the fine-tuning of large pre-trained models to cater to a diverse array of downstream tasks, showcasing notable effectiveness and efficiency, thereby solidifying its position as one of the most prevalent fine-tuning techniques. Due to the modular nature of LoRA's plug-and-play plugins, researchers have delved into the amalgamation of multiple LoRAs to empower models to excel across various downstream tasks. Nonetheless, extant approaches for LoRA fusion grapple with inherent challenges. Direct arithmetic merging may result in the loss of the original pre-trained model's generative capabilities or the distinct identity of LoRAs, thereby yielding suboptimal outcomes. On the other hand, Reference tuning-based fusion exhibits limitations concerning the requisite flexibility for the effective combination of multiple LoRAs. In response to these challenges, this paper introduces the Mixture of LoRA Experts (MoLE) approach, which harnesses hierarchical control and unfettered branch selection. The MoLE approach not only achieves superior LoRA fusion performance in comparison to direct arithmetic merging but also retains the crucial flexibility for combining LoRAs effectively. Extensive experimental evaluations conducted in both the Natural Language Processing (NLP) and Vision \& Language (V\&L) domains substantiate the efficacy of MoLE.
****

Zhengmian Hu,Lichang Chen,Xidong Wu,Yihan Wu,Hongyang Zhang,Heng Huang
Unbiased Watermark for Large Language Models
The recent advancements in large language models (LLMs) have sparked a growing apprehension regarding the potential misuse. One approach to mitigating this risk is to incorporate watermarking techniques into LLMs, allowing for the tracking and attribution of model outputs. This study examines a crucial aspect of watermarking: how significantly watermarks impact the quality of model-generated outputs. Previous studies have suggested a trade-off between watermark strength and output quality. However, our research demonstrates that it is possible to integrate watermarks without affecting the output probability distribution with appropriate implementation. We refer to this type of watermark as an unbiased watermark. This has significant implications for the use of LLMs, as it becomes impossible for users to discern whether a service provider has incorporated watermarks or not. Furthermore, the presence of watermarks does not compromise the performance of the model in downstream tasks, ensuring that the overall utility of the language model is preserved. Our findings contribute to the ongoing discussion around responsible AI development, suggesting that unbiased watermarks can serve as an effective means of tracking and attributing model outputs without sacrificing output quality.
****
Tao Ge,Hu Jing,Lei Wang,Xun Wang,Si-Qing Chen,Furu Wei
In-context Autoencoder for Context Compression in a Large Language Model
We propose the In-context Autoencoder (ICAE), leveraging the power of a large language models (LLM) to compress a long context into short compact memory slots that can be directly conditioned on by the LLM for various purposes. ICAE is first pretrained using both autoencoding and language modeling objectives on massive text data, enabling it to generate memory slots that accurately and comprehensively represent the original context; Then, it is fine-tuned on instruction data for producing desirable responses to various prompts. Experiments demonstrate that our lightweight ICAE, introducing about 1% additional parameters, effectively achieves $4\times$ context compression based on Llama, offering advantages in both improved latency and GPU memory cost during inference, and showing an interesting insight in memorization as well as potential for scalability. These promising results imply a novel perspective on the connection between working memory in cognitive science and representation learning in LLMs, revealing ICAE's significant implications in addressing the long context problem and suggesting further research in LLM context management. Our data, code and models are available at https://github.com/getao/icae.
****
Beomsu Kim,Gihyun Kwon,Kwanyoung Kim,Jong Chul Ye
Unpaired Image-to-Image Translation via Neural Schrödinger Bridge
Diffusion models are a powerful class of generative models which simulate stochastic differential equations (SDEs) to generate data from noise. While diffusion models have achieved remarkable progress, they have limitations in unpaired image-to-image (I2I) translation tasks due to the Gaussian prior assumption. Schrödinger Bridge (SB), which learns an SDE to translate between two arbitrary distributions, have risen as an attractive solution to this problem. Yet, to our best knowledge, none of SB models so far have been successful at unpaired translation between high-resolution images. In this work, we propose Unpaired Neural Schrödinger Bridge (UNSB), which expresses the SB problem as a sequence of adversarial learning problems. This allows us to incorporate advanced discriminators and regularization to learn a SB between unpaired data. We show that UNSB is scalable and successfully solves various unpaired I2I translation tasks. Code: \url{https://github.com/cyclomon/UNSB}
****
Suyu Ge,Yunan Zhang,Liyuan Liu,Minjia Zhang,Jiawei Han,Jianfeng Gao
Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs
In this study, we introduce adaptive KV cache compression, a plug-and-play method that reduces the memory footprint of generative inference for Large Language Models (LLMs). Different from the conventional KV cache that retains key and valu

e vectors for all context tokens, we conduct targeted profiling to discern the i
ntrinsic structure of attention modules. Based on the recognized structure, we t
hen construct the KV cache in an adaptive manner: evicting long-range contexts o
n attention heads emphasizing local contexts, discarding non-special tokens on a
ttention heads centered on special tokens, and only employing the standard KV ca
che for attention heads that broadly attend to all tokens. Moreover, with the li
ghtweight attention profiling used to guide the construction of the adaptive KV
cache, FastGen can be deployed without resource-intensive fine-tuning or re-trai
ning. In our experiments across various asks, FastGen demonstrates substantial r
eduction on GPU memory consumption with negligible generation quality loss. We w
ill release our code and the compatible CUDA kernel for reproducibility.
****
Haodong Lu,Dong Gong,Shuo Wang,Jason Xue,Lina Yao,Kristen Moore
Learning with Mixture of Prototypes for Out-of-Distribution Detection
Out-of-distribution (OOD) detection aims to detect testing samples far away from
 the in-distribution (ID) training data, which is crucial for the safe deploymen
t of machine learning models in the real world. Distance-based OOD detection met
hods have emerged with enhanced deep representation learning. They identify unse
en OOD samples by measuring their distances from ID class centroids or prototype
s. However, existing approaches learn the representation relying on oversimplifi
ed data assumptions, e.g. modeling ID data of each class with one centroid class
 prototype or using loss functions not designed for OOD detection, which overloo
k the natural diversities within the data. Naively enforcing data samples of eac
h class to be compact around only one prototype leads to inadequate modeling of
realistic data and limited performance. To tackle these issues, we propose Proto
typicAl Learning with a Mixture of prototypes (PALM) that models each class with
 multiple prototypes to capture the sample diversities, which learns more faithf
ul and compact samples embeddings for enhanching OOD detection. Our method autom
atically identifies and dynamically updates prototypes, assigning each sample to
 a subset of prototypes via reciprocal neighbor soft assignment weights. To lear
n embeddings with multiple prototypes, PALM optimizes a maximum likelihood estim
ation (MLE) loss to encourage the sample embeddings to compact around the associ
ated prototypes, as well as a contrastive loss on all prototypes to enhance intr
a-class compactness and inter-class discrimination at the prototype level. Compa
red to previous methods with prototypes, the proposed mixture prototype modeling
 of PALM promotes the representations of each ID class to be more compact and se
parable from others and the unseen OOD samples, resulting in more reliable OOD d
etection. Moreover, the automatic estimation of prototypes enables our approach
to be extended to the challenging OOD detection task with unlabelled ID data. Ex
tensive experiments demonstrate the superiority of PALM over previous methods, a
chieving state-of-the-art average AUROC performance of 93.82 on the challenging
CIFAR-100 benchmark.
****
Bowen Gao,Yinjun Jia,YuanLe Mo,Yuyan Ni,Wei-Ying Ma,Zhi-Ming Ma,Yanyan Lan
Self-supervised Pocket Pretraining via Protein Fragment-Surroundings Alignment
Pocket representations play a vital role in various biomedical applications, suc
h as druggability estimation, ligand affinity prediction, and de novo drug desig
n. While existing geometric features and pretrained representations have demonst
rated promising results, they usually treat pockets independent of ligands, negl
ecting the fundamental interactions between them. However, the limited pocket-li
gand complex structures available in the PDB database (less than 100 thousand no
n-redundant pairs) hampers large-scale pretraining endeavors for interaction mod
eling. To address this constraint, we propose a novel pocket pretraining approac
h that leverages knowledge from high-resolution atomic protein structures, assis
ted by highly effective pretrained small molecule representations. By segmenting
 protein structures into drug-like fragments and their corresponding pockets, we
 obtain a reasonable simulation of ligand-receptor interactions, resulting in th
e generation of over 5 million complexes. Subsequently, the pocket encoder is tr
ained in a contrastive manner to align with the representation of pseudo-ligand
furnished by some pretrained small molecule encoders. Our method, named ProFSA,

achieves state-of-the-art performance across various tasks, including pocket dru ggability prediction, pocket matching, and ligand binding affinity prediction. N otably, ProFSA surpasses other pretraining methods by a substantial margin. More over, our work opens up a new avenue for mitigating the scarcity of protein-liga nd complex data through the utilization of high-quality and diverse protein stru cture databases.
****
Frederikke Isa Marin,Felix Teufel,Marc Horlacher,Dennis Madsen,Dennis Pultz,Ole Winther,Wouter Boomsma
BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks
The genome sequence contains the blueprint for governing cellular processes.
While the availability of genomes has vastly increased over the last decades, experimental annotation of the various functional, non-coding and regulatory ele ments encoded in the DNA sequence remains both expensive and challenging. This h as sparked interest in unsupervised language modeling of genomic DNA, a paradigm that has seen great success for protein sequence data.
Although various DNA language models have been proposed, evaluation tasks ofte n differ between individual works, and might not fully recapitulate the fundamen tal challenges of genome annotation, including the length, scale and sparsity of the data. In this study, we introduce **BEND**, a **BEN**chmark for **D**NA lan guage models, featuring
a collection of realistic and biologically meaningful downstream tasks defined on the human genome.
We find that embeddings from current DNA LMs can approach performance of exper t methods on some tasks, but only capture limited information about long-range f eatures.
BEND is available at https://github.com/frederikkemarin/BEND.
****
Takeru Miyato,Bernhard Jaeger,Max Welling,Andreas Geiger
GTA: A Geometry-Aware Attention Mechanism for Multi-View Transformers
As transformers are equivariant to the permutation of input tokens, encoding the positional information of tokens is necessary for many tasks. However, since ex isting positional encoding schemes have been initially designed for NLP tasks, t heir suitability for vision tasks, which typically exhibit different structural properties in their data, is questionable. We argue that existing positional enc oding schemes are suboptimal for 3D vision tasks, as they do not respect their u nderlying 3D geometric structure. Based on this hypothesis, we propose a geometr y-aware attention mechanism that encodes the geometric structure of tokens as re lative transformation determined by the geometric relationship between queries a nd key-value pairs. By evaluating on multiple novel view synthesis (NVS) dataset s in the sparse wide-baseline multi-view setting, we show that our attention, ca lled Geometric Transform Attention (GTA), improves learning efficiency and perfo rmance of state-of-the-art transformer-based NVS models without any additional l earned parameters and only minor computational overhead.
****
Zhenwen Dai,Federico Tomasi,Sina Ghiassian
In-context Exploration-Exploitation for Reinforcement Learning
In-context learning is a promising approach for online policy learning of offlin e reinforcement learning (RL) methods, which can be achieved at inference time w ithout gradient optimization. However, this method is hindered by significant co mputational costs resulting from the gathering of large training trajectory sets and the need to train large Transformer models. We address this challenge by in troducing an In-context Exploration-Exploitation (ICEE) algorithm, designed to o ptimize the efficiency of in-context policy learning. Unlike existing models, IC EE performs an exploration-exploitation trade-off at inference time within a Tra nsformer model, without the need for explicit Bayesian inference. Consequently, ICEE can solve Bayesian optimization problems as efficiently as Gaussian process biased methods do, but in significantly less time. Through experiments in grid world environments, we demonstrate that ICEE can learn to solve new RL tasks usi ng only tens of episodes, marking a substantial improvement over the hundreds of

episodes needed by the previous in-context learning method.
****
Marco Jiralerspong,Bilun Sun,Danilo Vucetic,Tianyu Zhang,Yoshua Bengio,Gauthier Gidel,Nikolay Malkin
Expected flow networks in stochastic environments and two-player zero-sum games
Generative flow networks (GFlowNets) are sequential sampling models trained to match a given distribution. GFlowNets have been successfully applied to various structured object generation tasks, sampling a diverse set of high-reward objects quickly. We propose expected flow networks (EFlowNets), which extend GFlowNets to stochastic environments. We show that EFlowNets outperform other GFlowNet formulations in stochastic tasks such as protein design. We then extend the concept of EFlowNets to adversarial environments, proposing adversarial flow networks (AFlowNets) for two-player zero-sum games. We show that AFlowNets learn to find above 80% of optimal moves in Connect-4 via self-play and outperform AlphaZero in tournaments.
Code: https://github.com/GFNOrg/AdversarialFlowNetworks.
****
Linan Yue,Qi Liu,Yichao Du,Li Wang,Weibo Gao,Yanqing An
Towards Faithful Explanations: Boosting Rationalization with Shortcuts Discovery
The remarkable success in neural networks provokes the selective rationalization. It explains the prediction results by identifying a small subset of the inputs sufficient to support them. Since existing methods still suffer from adopting the shortcuts in data to compose rationales and limited large-scale annotated rationales by human, in this paper, we propose a Shortcuts-fused Selective Rationalization (SSR) method, which boosts the rationalization by discovering and exploiting potential shortcuts. Specifically, SSR first designs a shortcuts discovery approach to detect several potential shortcuts. Then, by introducing the identified shortcuts, we propose two strategies to mitigate the problem of utilizing shortcuts to compose rationales. Finally, we develop two data augmentations methods to close the gap in the number of annotated rationales. Extensive experimental results on real-world datasets clearly validate the effectiveness of our proposed method.
****
Xinwei Zhang,Zhiqi Bu,Steven Wu,Mingyi Hong
Differentially Private SGD Without Clipping Bias: An Error-Feedback Approach
Differentially Private Stochastic Gradient Descent with gradient clipping (DPSGD-GC) is a powerful tool for training deep learning models using sensitive data, providing both a solid theoretical privacy guarantee and high efficiency. However, existing research has shown that DPSGD-GC only converges when using large clipping thresholds that are dependent on problem-specific parameters that are often unknown in practice. Therefore, DPSGD-GC suffers from degraded performance due to the {\it constant} bias introduced by the clipping. In our work, we propose a new error-feedback (EF) DP algorithm as an alternative to DPSGD-GC, which offers a diminishing utility bound without inducing a constant clipping bias. More importantly, it allows for an arbitrary choice of clipping threshold that is independent of the problem. We establish an algorithm-specific DP analysis for our proposed algorithm, providing privacy guarantees based on R{\'e}nyi DP. And we demonstrate that under mild conditions, our algorithm can achieve nearly the same utility bound as DPSGD without gradient clipping. Our empirical results on standard datasets show that the proposed algorithm achieves higher accuracies than DPSGD while maintaining the same level of DP guarantee.
****
Yuanqing Huang,Yinggui Wang,Le Yang,Lei Wang
Enhanced Face Recognition using Intra-class Incoherence Constraint
The current face recognition (FR) algorithms has achieved a high level of accuracy, making further improvements increasingly challenging. While existing FR algorithms primarily focus on optimizing margins and loss functions, limited attention has been given to exploring the feature representation space. Therefore, this paper endeavors to improve FR performance in the view of feature representation space. Firstly, we consider two FR models that exhibit distinct performance dis

crepancies, where one model exhibits superior recognition accuracy compared to t
he other. We implement orthogonal decomposition on the features from the superio
r model along those from the inferior model and obtain two sub-features. Surpris
ingly, we find the sub-feature perpendicular to the inferior still possesses a c
ertain level of face distinguishability. We adjust the modulus of the sub-featur
es and recombine them through vector addition. Experiments demonstrate this reco
mbination is likely to contribute to an improved facial feature representation,
even better than features from the original superior model. Motivated by this di
scovery, we further consider how to improve FR accuracy when there is only one F
R model available. Inspired by knowledge distillation, we incorporate the intra-
class incoherence constraint (IIC) to solve the problem. Experiments on various
FR benchmarks show the existing state-of-the-art method with IIC can be further
improved, highlighting its potential to further enhance FR performance.
****
Dan Haramati,Tal Daniel,Aviv Tamar
Entity-Centric Reinforcement Learning for Object Manipulation from Pixels
Manipulating objects is a hallmark of human intelligence, and an important task
in domains such as robotics. In principle, Reinforcement Learning (RL) offers a
general approach to learn object manipulation. In practice, however, domains wit
h more than a few objects are difficult for RL agents due to the curse of dimens
ionality, especially when learning from raw image observations. In this work we
propose a structured approach for visual RL that is suitable for representing mu
ltiple objects and their interaction, and use it to learn goal-conditioned manip
ulation of several objects. Key to our method is the ability to handle goals wit
h dependencies between the objects (e.g., moving objects in a certain order). We
further relate our architecture to the generalization capability of the trained
agent, based on a theoretical result for compositional generalization, and demo
nstrate agents that learn with 3 objects but generalize to similar tasks with ov
er 10 objects. Videos and code are available on the project website: https://sit
es.google.com/view/entity-centric-rl
****
Xiang Hu,Qingyang Zhu,Kewei Tu,Wei Wu
Augmenting Transformers with Recursively Composed Multi-grained Representations
We present ReCAT, a recursive composition augmented Transformer that is able to
explicitly model hierarchical syntactic structures of raw texts without relying
on gold trees during both learning and inference.
Existing research along this line restricts data to follow a hierarchical tree s
tructure and thus lacks inter-span communications.
To overcome the problem, we propose a novel contextual inside-outside (CIO) laye
r that learns contextualized representations of spans through bottom-up and top-
down passes, where a bottom-up pass forms representations of high-level spans by
composing low-level spans, while a top-down pass combines information inside an
d outside a span. By stacking several CIO layers between the embedding layer and
the attention layers in Transformer, the ReCAT model can perform both deep intr
a-span and deep inter-span interactions, and thus generate multi-grained represe
ntations fully contextualized with other spans.
Moreover, the CIO layers can be jointly pre-trained with Transformers, making Re
CAT enjoy scaling ability, strong performance, and interpretability at the same
time. We conduct experiments on various sentence-level and span-level tasks. Eva
luation results indicate that ReCAT can significantly outperform vanilla Transfo
rmer models on all span-level tasks and recursive models on natural language inf
erence tasks. More interestingly, the hierarchical structures induced by ReCAT e
xhibit strong consistency with human-annotated syntactic trees, indicating good
interpretability brought by the CIO layers.
****
Guang Lin,Chao Li,Jianhai Zhang,Toshihisa Tanaka,Qibin Zhao
Adversarial Training on Purification (AToP): Advancing Both Robustness and Gener
alization
The deep neural networks are known to be vulnerable to well-designed adversarial
attacks. The most successful defense technique based on adversarial training (A

T) can achieve optimal robustness against particular attacks but cannot generalize well to unseen attacks. Another effective defense technique based on adversarial purification (AP) can enhance generalization but cannot achieve optimal robustness. Meanwhile, both methods share one common limitation on the degraded standard accuracy. To mitigate these issues, we propose a novel pipeline to acquire the robust purifier model, named Adversarial Training on Purification (AToP), which comprises two components: perturbation destruction by random transforms (RT) and purifier model fine-tuned (FT) by adversarial loss. RT is essential to avoid overlearning to known attacks, resulting in the robustness generalization to unseen attacks, and FT is essential for the improvement of robustness.
To evaluate our method in an efficient and scalable way, we conduct extensive experiments on CIFAR-10, CIFAR-100, and ImageNette to demonstrate that our method achieves optimal robustness and exhibits generalization ability against unseen attacks.
****
Andrew Szot,Max Schwarzer,Harsh Agrawal,Bogdan Mazoure,Rin Metcalf,Walter Talbott,Natalie Mackraz,R Devon Hjelm,Alexander T Toshev
Large Language Models as Generalizable Policies for Embodied Tasks
We show that large language models (LLMs) can be adapted to be generalizable policies for embodied visual tasks. Our approach, called Large LAnguage model Reinforcement Learning Policy (LLaRP), adapts a pre-trained frozen LLM to take as input text instructions and visual egocentric observations and output actions directly in the environment. Using reinforcement learning, we train LLaRP to see and act solely through environmental interactions. We show that LLaRP is robust to complex paraphrasings of task instructions and can generalize to new tasks that require novel optimal behavior. In particular, on 1,000 unseen tasks it achieves 42% success rate, 1.7x the success rate of other common learned baselines or zero-shot applications of LLMs. Finally, to aid the community in studying language conditioned, massively multi-task, embodied AI problems we release a novel benchmark, Language Rearrangement, consisting of 150,000 training and 1,000 testing tasks for language-conditioned rearrangement.
****
Yingqing He,Shaoshu Yang,Haoxin Chen,Xiaodong Cun,Menghan Xia,Yong Zhang,Xintao Wang,Ran He,Qifeng Chen,Ying Shan
ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models
In this work, we investigate the capability of generating images from pre-trained diffusion models at much higher resolutions than the training image sizes. In addition, the generated images should have arbitrary image aspect ratios. When generating images directly at a higher resolution, 1024 x 1024, with the pre-trained Stable Diffusion using training images of resolution 512 x 512, we observe persistent problems of object repetition and unreasonable object structures. Existing works for higher-resolution generation, such as attention-based and joint-diffusion approaches, cannot well address these issues. As a new perspective, we examine the structural components of the U-Net in diffusion models and identify the crucial cause as the limited perception field of convolutional kernels. Based on this key observation, we propose a simple yet effective re-dilation that can dynamically adjust the convolutional perception field during inference. We further propose the dispersed convolution and noise-damped classifier-free guidance, which can enable ultra-high-resolution image generation (e.g., 4096 x 4096). Notably, our approach does not require any training or optimization. Extensive experiments demonstrate that our approach can address the repetition issue well and achieve state-of-the-art performance on higher-resolution image synthesis, especially in texture details. Our work also suggests that a pre-trained diffusion model trained on low-resolution images can be directly used for high-resolution visual generation without further tuning, which may provide insights for future research on ultra-high-resolution image and video synthesis. More results are available at the anonymous website: https://scalecrafter.github.io/ScaleCrafter/
****
Nithin Chalapathi,Yiheng Du,Aditi S. Krishnapriyan

Scaling physics-informed hard constraints with mixture-of-experts
Imposing known physical constraints, such as conservation laws, during neural network training introduces an inductive bias that can improve accuracy, reliability, convergence, and data efficiency for modeling physical dynamics. While such constraints can be softly imposed via loss function penalties, recent advancements in differentiable physics and optimization improve performance by incorporating PDE-constrained optimization as individual layers in neural networks. This enables a stricter adherence to physical constraints. However, imposing hard constraints significantly increases computational and memory costs, especially for complex dynamical systems. This is because it requires solving an optimization problem over a large number of points in a mesh, representing spatial and temporal discretizations, which greatly increases the complexity of the constraint. To address this challenge, we develop a scalable approach to enforce hard physical constraints using Mixture-of-Experts (MoE), which can be used with any neural network architecture. Our approach imposes the constraint over smaller decomposed domains, each of which is solved by an ``expert'' through differentiable optimization. During training, each expert independently performs a localized backpropagation step by leveraging the implicit function theorem; the independence of each expert allows for parallelization across multiple GPUs. Compared to standard differentiable optimization, our scalable approach achieves greater accuracy in the neural PDE solver setting for predicting the dynamics of challenging non-linear systems. We also improve training stability and require significantly less computation time during both training and inference stages.
****
Daniel Goldfarb,Itay Evron,Nir Weinberger,Daniel Soudry,PAul HAnd
The Joint Effect of Task Similarity and Overparameterization on Catastrophic Forgetting — An Analytical Model
In continual learning, catastrophic forgetting is affected by multiple aspects of the tasks. Previous works have analyzed separately how forgetting is affected by either task similarity or overparameterization. In contrast, our paper examines how task similarity and overparameterization jointly affect forgetting in an analyzable model. Specifically, we focus on two-task continual linear regression, where the second task is a random orthogonal transformation of an arbitrary first task (an abstraction of random permutation tasks). We derive an exact analytical expression for the expected forgetting — and uncover a nuanced pattern. In highly overparameterized models, intermediate task similarity causes the most forgetting. However, near the interpolation threshold, forgetting decreases monotonically with the expected task similarity. We validate our findings with linear regression on synthetic data, and with neural networks on established permutation task benchmarks.
****
Vladimir R Kostic,Pietro Novelli,Riccardo Grazzi,Karim Lounici,massimiliano pontil
Learning invariant representations of time-homogeneous stochastic dynamical systems
We consider the general class of time-homogeneous stochastic dynamical systems, both discrete and continuous, and study the problem of learning a representation of the state that faithfully captures its dynamics. This is instrumental to learning the transfer operator or the generator of the system, which in turn can be used for numerous tasks, such as forecasting and interpreting the system dynamics. We show that the search for a good representation can be cast as an optimization problem over neural networks. Our approach is supported by recent results in statistical learning theory, highlighting the role of approximation error and metric distortion in the learning problem. The objective function we propose is associated with projection operators from the representation space to the data space, overcomes metric distortion, and can be empirically estimated from data. In the discrete-time setting, we further derive a relaxed objective function that is differentiable and numerically well-conditioned. We compare our method against state-of-the-art approaches on different datasets, showing better performance across the board.

****
Mrinank Sharma,Meg Tong,Tomasz Korbak,David Duvenaud,Amanda Askell,Samuel R. Bowman,Esin DURMUS,Zac Hatfield-Dodds,Scott R Johnston,Shauna M Kravec,Timothy Maxwell,Sam McCandlish,Kamal Ndousse,Oliver Rausch,Nicholas Schiefer,Da Yan,Miranda Zhang,Ethan Perez
Towards Understanding Sycophancy in Language Models
Reinforcement learning from human feedback (RLHF) is a popular technique for training high-quality AI assistants. However, RLHF may also encourage model responses that match user beliefs over truthful responses, a behavior known as sycophancy. We investigate the prevalence of sycophancy in RLHF-trained models and whether human preference judgments are responsible. We first demonstrate that five state-of-the-art AI assistants consistently exhibit sycophancy behavior across four varied free-form text-generation tasks. To understand if human preferences drive this broadly observed behavior of RLHF models, we analyze existing human preference data. We find that when a response matches a user's views, it is more likely to be preferred. Moreover, both humans and preference models (PMs) prefer convincingly-written sycophantic responses over correct ones a non-negligible fraction of the time. Optimizing model outputs against PMs also sometimes sacrifices truthfulness in favor of sycophancy. Overall, our results indicate that sycophancy is a general behavior of RLHF models, likely driven in part by human preference judgments favoring sycophantic responses.
****
Yameng Peng,Andy Song,Haytham M. Fayek,Vic Ciesielski,Xiaojun Chang
SWAP-NAS: Sample-Wise Activation Patterns for Ultra-fast NAS
Training-free metrics (a.k.a. zero-cost proxies) are widely used to avoid resource-intensive neural network training, especially in Neural Architecture Search (NAS). Recent studies show that existing training-free metrics have several limitations, such as limited correlation and poor generalisation across different search spaces and tasks. Hence, we propose Sample-Wise Activation Patterns and its derivative, SWAP-Score, a novel high-performance training-free metric. It measures the expressivity of networks over a batch of input samples. The SWAP-Score is strongly correlated with ground-truth performance across various search spaces and tasks, outperforming 15 existing training-free metrics on NAS-Bench-101/201/301 and TransNAS-Bench-101. The SWAP-Score can be further enhanced by regularisation, which leads to even higher correlations in cell-based search space and enables model size control during the search. For example, Spearman’s rank correlation coefficient between regularised SWAP-Score and CIFAR-100 validation accuracies on NAS-Bench-201 networks is 0.90, significantly higher than 0.80 from the second-best metric, NWOT. When integrated with an evolutionary algorithm for NAS, our SWAP-NAS achieves competitive performance on CIFAR-10 and ImageNet in approximately 6 minutes and 9 minutes of GPU time respectively.
****
Penghui Qi,Xinyi Wan,Guangxing Huang,Min Lin
Zero Bubble (Almost) Pipeline Parallelism
Pipeline parallelism is one of the key components for large-scale distributed training, yet its efficiency suffers from pipeline bubbles which were deemed inevitable. In this work, we introduce a scheduling strategy that, to our knowledge, is the first to successfully achieve zero pipeline bubbles under synchronous training semantics. The key idea behind this improvement is to split the backward computation into two parts, one that computes gradient for the input and another that computes for the parameters. Based on this idea, we handcraft novel pipeline schedules that significantly outperform the baseline methods. We further develop an algorithm that automatically finds an optimal schedule based on specific model configuration and memory limit. Additionally, to truly achieve zero bubble, we introduce a novel technique to bypass synchronizations during the optimizer step. Experimental evaluations show that our method outperforms the 1F1B schedule up to 15\% in throughput under a similar memory limit. This number can be further pushed to 30\% when the memory constraint is relaxed. We believe our results mark a major step forward in harnessing the true potential of pipeline parallelism. The source code based on Megatron-LM is publicly avaiable at \url{https://g

ithub.com/sail-sg/zero-bubble-pipeline-parallelism}.
****
Yuhui Zhang,Elaine Sui,Serena Yeung
Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data
Building cross-modal applications is challenging due to limited paired multi-mod
al data. Recent works have shown that leveraging a pre-trained multi-modal contr
astive representation space enables cross-modal tasks to be learned from uni-mod
al data. This is based on the assumption that contrastive optimization makes emb
eddings from different modalities interchangeable. However, this assumption is u
nder-explored due to the poorly understood geometry of the multi-modal contrasti
ve space, where a modality gap exists. In our study, we provide a theoretical ex
planation of this space's geometry and introduce a three-step method, C^3 (Con
nect, Collapse, Corrupt), to bridge the modality gap, enhancing the interchangea
bility of embeddings. Our C^3 method significantly improves cross-modal learni
ng from uni-modal data, achieving state-of-the-art results on zero-shot image /
audio / video captioning and text-to-image generation.
****
Weidong Huang,Jiaming Ji,Borong Zhang,Chunhe Xia,Yaodong Yang
SafeDreamer: Safe Reinforcement Learning with World Models
The deployment of Reinforcement Learning (RL) in real-world applications is cons
trained by its failure to satisfy safety criteria.
Existing Safe Reinforcement Learning (SafeRL) methods, which rely on cost functi
ons to enforce safety, often fail to achieve zero-cost performance in complex sc
enarios, especially vision-only tasks. These limitations are primarily due to mo
del inaccuracies and inadequate sample efficiency. The integration of the world
model has proven effective in mitigating these shortcomings. In this work, we in
troduce SafeDreamer, a novel algorithm incorporating Lagrangian-based methods in
to world model planning processes within the superior Dreamer framework. Our met
hod achieves nearly zero-cost performance on various tasks, spanning low-dimensi
onal and vision-only input, within the Safety-Gymnasium benchmark, showcasing it
s efficacy in balancing performance and safety in RL tasks. Further details can
be seen on our project website: https://sites.google.com/view/safedreamer.
****
Zhiyuan Zeng,Jiatong Yu,Tianyu Gao,Yu Meng,Tanya Goyal,Danqi Chen
Evaluating Large Language Models at Evaluating Instruction Following
As research in large language models (LLMs) continues to accelerate, LLM-based e
valuation has emerged as a scalable and cost-effective alternative to human eval
uations for comparing the ever-increasing list of models. This paper investigate
s the efficacy of these “LLM evaluators”, particularly in using them to assess i
nstruction following, a metric that gauges how closely generated text adheres to
 the
instructions. We introduce a challenging meta-evaluation benchmark, LLMBAR, desi
gned to test the ability of an LLM evaluator to discern instruction-following ou
tputs. The authors curated 419 pairs of outputs, one adhering to instructions wh
ile the other diverging, yet may possess deceptive qualities that could mislead
an LLM evaluator. Contrary to existing meta-evaluation, we discover that differe
nt evaluators (i.e., combinations of LLMs and prompts) exhibit distinct performa
nce on LLMBAR and even the highest-scoring LLM evaluators have substantial room
for improvement. We also present a novel suite of prompting strategies that furt
her close the gap between LLM and human evaluators. With LLMBAR, we hope to offe
r more insight into the behavior of LLM evaluators and foster research in develo
ping better instruction-following models.
****
Sebastian Pineda Arango,Fabio Ferreira,Arlind Kadra,Frank Hutter,Josif Grabocka
Quick-Tune: Quickly Learning Which Pretrained Model to Finetune and How
With the ever-increasing number of pretrained models, machine learning practitio
ners are continuously faced with which pretrained model to use, and how to finet
une it for a new dataset. In this paper, we propose a methodology that jointly s
earches for the optimal pretrained model and the hyperparameters for finetuning
it. Our method transfers knowledge about the performance of many pretrained mode

ls with multiple hyperparameter configurations on a series of datasets. To this aim, we evaluated over 20k hyperparameter configurations for finetuning 24 pretrained image classification models on 87 datasets to generate a large-scale meta-dataset. We meta-learn a gray-box performance predictor on the learning curves of this meta-dataset and use it for fast hyperparameter optimization on new datasets. We empirically demonstrate that our resulting approach can quickly select an accurate pretrained model for a new dataset together with its optimal hyperparameters.
****
Zehao Dou,Yang Song
Diffusion Posterior Sampling for Linear Inverse Problem Solving: A Filtering Perspective
Diffusion models have achieved tremendous success in generating high-dimensional data like images, videos and audio. These models provide powerful data priors that can solve linear inverse problems in zero shot through Bayesian posterior sampling.
However, exact posterior sampling for diffusion models is intractable. Current solutions often hinge on approximations that are either computationally expensive or lack strong theoretical guarantees. In this work, we introduce an efficient diffusion sampling algorithm for linear inverse problems that is guaranteed to be asymptotically accurate. We reveal a link between Bayesian posterior sampling and Bayesian filtering in diffusion models, proving the former as a specific instance of the latter. Our method, termed filtering posterior sampling, leverages sequential Monte Carlo methods to solve the corresponding filtering problem. It seamlessly integrates with all Markovian diffusion samplers, requires no model re-training, and guarantees accurate samples from the Bayesian posterior as particle counts rise. Empirical tests demonstrate that our method generates better or comparable results than leading zero-shot diffusion posterior samplers on tasks like image inpainting, super-resolution, and deblurring.
****
Prasanna Mayilvahanan,Thaddäus Wiedemer,Evgenia Rusak,Matthias Bethge,Wieland Brendel
Does CLIP’s generalization performance mainly stem from high train-test similarity?
Foundation models like CLIP are trained on hundreds of millions of samples and effortlessly generalize to new tasks and inputs. Out of the box, CLIP shows stellar zero-shot and few-shot capabilities on a wide range of out-of-distribution (OOD) benchmarks, which prior works attribute mainly to today's large and comprehensive training dataset (like LAION). However, it is questionable how meaningful terms like out-of-distribution generalization are for CLIP as it seems likely that web-scale datasets like LAION simply contain many samples that are similar to common OOD benchmarks originally designed for ImageNet. To test this hypothesis, we retrain CLIP on pruned LAION splits that replicate ImageNet’s train-test similarity with respect to common OOD benchmarks. While we observe a performance drop on some benchmarks, surprisingly, CLIP’s overall performance remains high. This shows that high train-test similarity is insufficient to explain CLIP’s OOD performance, and other properties of the training data must drive CLIP to learn more generalizable representations. Additionally, by pruning data points that are dissimilar to the OOD benchmarks, we uncover a 100M split of LAION (¼ of its original size) on which CLIP can be trained to match its original OOD performance.
****
Xiao Zhang,Ji Wu
Dissecting learning and forgetting in language model finetuning
Finetuning language models on domain-specific corpus is a common approach to enhance their domain knowledge and capability. While improving performance on domain tasks, it often brings a side-effect of forgetting of the model's general abilities. In this study, we analyze the effects of finetuning on language models by dissecting its impacts on the modeling of topic, style, and factual knowledge in text. Our method uses instruction-following LLMs such as ChatGPT to auto-gener

ate controlled-variable text examples which we use to probe the model. Our findi ngs reveal that finetuning results in significant shifts in the language model's topic and style priors, while actual knowledge learning only contributes to a s mall fraction of the total probability change. Analysis shows that the adaptatio n of topic and style priors behave akin to learning simple features: they are le arned rapidly and require little model capacity. They are also learned independe ntly and primarily at the beginning of a text sequence. In contrast, factual kno wledge is learned stably but slowly and requires significant model capacity to l earn. The research offers insights and understanding into the finer dynamics of learning and forgetting in language models, and can potentially inform future re search on improving domain adaptation and addressing the challenges of forgettin g in continual learning of language models.
****
Jiarong Liu,Yifan Zhong,Siyi Hu,Haobo Fu,QIANG FU,Xiaojun Chang,Yaodong Yang
Maximum Entropy Heterogeneous-Agent Reinforcement Learning
Multi-agent reinforcement learning (MARL) has been shown effective for coopera tive games in recent years. However, existing state-of-the-art methods face chal lenges related to sample complexity, training instability, and the risk of conve rging to a suboptimal Nash Equilibrium. In this paper, we propose a unified fram ework for learning \emph{stochastic} policies to resolve these issues. We embed cooperative MARL problems into probabilistic graphical models, from which we der ive the maximum entropy (MaxEnt) objective for MARL. Based on the MaxEnt framewo rk, we propose *Heterogeneous-Agent Soft Actor-Critic* (HASAC) algorithm. Theore tically, we prove the monotonic improvement and convergence to *quantal response equilibrium* (QRE) properties of HASAC. Furthermore, we generalize a unified te mplate for MaxEnt algorithmic design named *Maximum Entropy Heterogeneous-Agent Mirror Learning* (MEHAML), which provides any induced method with the same guara ntees as HASAC. We evaluate HASAC on six benchmarks: Bi-DexHands, Multi-Agent Mu JoCo, StarCraft Multi-Agent Challenge, Google Research Football, Multi-Agent Par ticle Environment, and Light Aircraft Game. Results show that HASAC consistently outperforms strong baselines, exhibiting better sample efficiency, robustness, and sufficient exploration.
****
Martin Klissarov,Pierluca D'Oro,Shagun Sodhani,Roberta Raileanu,Pierre-Luc Bacon ,Pascal Vincent,Amy Zhang,Mikael Henaff
Motif: Intrinsic Motivation from Artificial Intelligence Feedback
Exploring rich environments and evaluating one's actions without prior knowledge is immensely challenging. In this paper, we propose Motif, a general method to interface such prior knowledge from a Large Language Model (LLM) with an agent. Motif is based on the idea of grounding LLMs for decision-making without requiri ng them to interact with the environment: it elicits preferences from an LLM ove r pairs of captions to construct an intrinsic reward, which is then used to trai n agents with reinforcement learning. We evaluate Motif's performance and behavi or on the challenging, open-ended and procedurally-generated NetHack game. Surpr isingly, by only learning to maximize its intrinsic reward, Motif achieves a hig her game score than an algorithm directly trained to maximize the score itself. When combining Motif's intrinsic reward with the environment reward, our method significantly outperforms existing approaches and makes progress on tasks where no advancements have ever been made without demonstrations. Finally, we show tha t Motif mostly generates intuitive human-aligned behaviors which can be steered easily through prompt modifications, while scaling well with the LLM size and th e amount of information given in the prompt.
****
Rong Dai,Yonggang Zhang,Ang Li,Tongliang Liu,Xun Yang,Bo Han
Enhancing One-Shot Federated Learning Through Data and Ensemble Co-Boosting
One-shot Federated Learning (OFL) has become a promising learning paradigm, enab ling the training of a global server model via a single communication round. In OFL, the server model is aggregated by distilling knowledge from all client mode ls (the ensemble), which are also responsible for synthesizing samples for disti llation. In this regard, advanced works show that the performance of the server

model is intrinsically related to the quality of the synthesized data and the en semble model. To promote OFL, we introduce a novel framework, Co-Boosting, in wh ich synthesized data and the ensemble model mutually enhance each other progress ively. Specifically, Co-Boosting leverages the current ensemble model to synthes ize higher-quality samples in an adversarial attack manner. These hard samples a re then employed to promote the quality of the ensemble model by adjusting the e nsembling weights for each client model. Consequently, Co-Boosting periodically achieves high-quality data and ensemble models. Extensive experiments demonstrat e that Co-Boosting can substantially outperform existing baselines under various settings. Moreover, Co-Boosting eliminates the need for adjustments to the clie nt's local training, requires no additional data or model transmission, and allo ws client models to have heterogeneous architectures.
****
Arman Isajanyan,Artur Shatveryan,David Kocharian,Zhangyang Wang,Humphrey Shi
Social Reward: Evaluating and Enhancing Generative AI through Million-User Feedb ack from an Online Creative Community
Social reward as a form of community recognition provides a strong source of motivation for users of online platforms to actively engage and contribute with content to accumulate peers approval. In the realm of text-conditioned image synthesis, the recent surge in progress has ushered in a collaborative era where users and AI systems coalesce to refine visual creations. This co-creative pro-cess in the landscape of online social networks empowers users to craft original visual artworks seeking for community validation. Nevertheless, assessing these models in the context of collective community preference introduces distinct cha l-
lenges. Existing evaluation methods predominantly center on limited size user studies guided by image quality and alignment with prompts. This work pio-neers a paradigm shift, unveiling Social Reward - an innovative reward modeling framework that leverages implicit feedback from social network users engaged in creative editing of generated images. We embark on an extensive journey of dataset curation and refinement, drawing from Picsart: an online visual creation and editing platform, yielding a first million-user-scale dataset of implicit hu man
preferences for user-generated visual art named Picsart Image-Social. Our anal-ysis exposes the shortcomings of current metrics in modeling community creative preference of text-to-image models’ outputs, compelling us to introduce a novel predictive model explicitly tailored to address these limitations. Rigorous quan -
titative experiments and user study show that our Social Reward model aligns better with social popularity than existing metrics. Furthermore, we utilize So-cial Reward to fine-tune text-to-image models, yielding images that are more fa-vored by not only Social Reward, but also other established metrics. These find-ings highlight the relevance and effectiveness of Social Reward in assessing com -
munity appreciation for AI-generated artworks, establishing a closer alignment with users’ creative goals: creating popular visual art. Codes can be accessed a t
https://github.com/Picsart-AI-Research/Social-Reward
****
Dongyoung Go,Tomasz Korbak,Germán Kruszewski,Jos Rozen,Marc Dymetman
Compositional Preference Models for Aligning LMs
As language models (LMs) become more capable, it is increasingly important to al ign them with human preferences. However, the dominant paradigm for training Pre ference Models (PMs) for that purpose suffers from fundamental limitations, such as lack of transparency and scalability, along with susceptibility to overfitti ng the preference dataset.
We propose Compositional Preference Models (CPMs), a novel PM framework that dec omposes one global preference assessment into several interpretable features, ob tains scalar scores for these features from a prompted LM, and aggregates these scores using a logistic regression classifier. Through these simple steps, CPMs

allow to control which properties of the preference data are used to train the p
reference model and to build it based on features that are believed to underlie
the human preference judgment.
Our experiments show that CPMs not only improve generalization and are more robu
st to overoptimization than standard PMs, but also that best-of-n samples obtain
ed using CPMs tend to be preferred over samples obtained using conventional PMs.
Overall, our approach demonstrates the benefits of endowing PMs with priors abou
t which features determine human preferences while relying on LM capabilities to
 extract those features in a scalable and robust way.
****
Adyasha Maharana,Prateek Yadav,Mohit Bansal
$\mathbb{D}^2$ Pruning: Message Passing for Balancing Diversity & Difficulty in
Data Pruning
In recent years, data quality has emerged as an important factor for training ma
ssive models. Analytical theories suggest that higher-quality data can lead to l
ower test errors in models trained on a fixed data budget. Moreover, a model can
 be trained on a lower compute budget without compromising performance if a data
set can be stripped of its redundancies. Coreset selection (or data pruning) see
ks to select a subset of the training data so as to maximize the performance of
models trained on this subset, also referred to as coreset. There are two domina
nt approaches: (1) geometry-based data selection for maximizing *data diversity*
 in the coreset, and (2) functions that assign *difficulty scores* to samples ba
sed on training dynamics. Optimizing for data diversity leads to a coreset that
is biased towards easier samples, whereas, selection by difficulty ranking omits
 easy samples that are necessary for the training of deep learning models. This
demonstrates that data diversity and importance scores are two complementary fac
tors that need to be jointly considered during coreset selection. In this work,
we represent a dataset as an undirected graph and propose a novel pruning algori
thm, $\mathbb{D}^2$ Pruning, that uses message passing over this dataset graph f
or coreset selection. $\mathbb{D}^2$ Pruning updates the difficulty scores of ea
ch example by incorporating the difficulty of its neighboring examples in the da
taset graph. Then, these updated difficulty scores direct a graph-based sampling
 method to select a coreset that encapsulates both diverse and difficult regions
 of the dataset space. We evaluate supervised and self-supervised versions of ou
r method on various vision and NLP datasets. Results show that $\mathbb{D}^2$ Pr
uning improves coreset selection over previous state-of-the-art methods at low-t
o-medium pruning rates. Additionally, we find that using $\mathbb{D}^2$ Pruning
for filtering large multimodal datasets leads to increased diversity in the data
set and improved generalization of pretrained models. Our work shows that $\mathbb{D}^2$ Pruning is a versatile framework for understanding and processing datas
ets.
****
Garrett Tanzer,Mirac Suzgun,Eline Visser,Dan Jurafsky,Luke Melas-Kyriazi
A Benchmark for Learning to Translate a New Language from One Grammar Book
Large language models (LLMs) can perform impressive feats with in-context learni
ng or lightweight finetuning. It is natural to wonder how well these models adap
t to genuinely new tasks, but how does one find tasks that are unseen in interne
t-scale training sets? We turn to a field that is explicitly motivated and bottl
enecked by a scarcity of web data: low-resource languages. In this paper, we int
roduce MTOB (Machine Translation from One Book), a benchmark for learning to tra
nslate between English and Kalamang—a language with less than 200 speakers and t
herefore virtually no presence on the web—using several hundred pages of field l
inguistics reference materials. This task framing is novel in that it asks a mod
el to learn a language from a single human-readable book of grammar explanations
, rather than a large mined corpus of in-domain data, more akin to L2 language l
earning than L1 language acquisition. We demonstrate that baselines using curren
t LLMs are promising but fall short of human performance, achieving 44.7 chrF on
 Kalamang to English translation and 45.8 chrF on English to Kalamang translatio
n, compared to 51.6 and 57.0 chrF by a human who learned Kalamang from the same
reference materials. We hope that MTOB will help measure LLM capabilities along

a new dimension, and that the methods developed to solve it could help expand access to language technology for underserved communities by leveraging qualitatively different kinds of data than traditional machine translation.
****
Kun LEI,Zhengmao He,Chenhao Lu,Kaizhe Hu,Yang Gao,Huazhe Xu
Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization
Combining offline and online reinforcement learning (RL) is crucial for efficient and safe learning. However, previous approaches treat offline and online learning as separate procedures, resulting in redundant designs and limited performance. We ask: *Can we achieve straightforward yet effective offline and online learning without introducing extra conservatism or regularization?* In this study, we propose Uni-O4, which utilizes an on-policy objective for both offline and online learning. Owning to the alignment of objectives in two phases, the RL agent can transfer between offline and online learning seamlessly. This property enhances the flexibility of the learning paradigm, allowing for arbitrary combinations of pretraining, fine-tuning, offline, and online learning. In the offline phase, specifically, Uni-O4 leverages diverse ensemble policies to address the mismatch issues between the estimated behavior policy and the offline dataset. Through a simple offline policy evaluation (OPE) approach, Uni-O4 can achieve multi-step policy improvement safely. We demonstrate that by employing the method above, the fusion of these two paradigms can yield superior offline initialization as well as stable and rapid online fine-tuning capabilities.
Through real-world robot tasks, we highlight the benefits of this paradigm for rapid deployment in challenging, previously unseen real-world environments. Additionally, through comprehensive evaluations using numerous simulated benchmarks, we substantiate that our method achieves state-of-the-art performance in both offline and offline-to-online fine-tuning learning. [Our website](uni-o4.github.io)
****
Harikrishna Narasimhan,Aditya Krishna Menon,Wittawat Jitkrittum,Neha Gupta,Sanjiv Kumar
Learning to Reject for Balanced Error and Beyond
Learning to reject (L2R) is a classical problem where one seeks a classifier capable of abstaining on low-confidence samples. Most prior work on L2R has focused on minimizing the standard misclassification error. However, in many real-world applications, the label distribution is highly imbalanced, necessitating alternate evaluation metrics such as the balanced error or the worst-group error that enforce equitable performance across both the head and tail classes. In this paper, we establish that traditional L2R methods can be grossly sub-optimal for such metrics, and show that this is due to an intricate dependence in the objective between the label costs and the rejector. We then derive the form of the Bayes-optimal classifier and rejector for the balanced error, propose a novel plug-in approach to mimic this solution, and extend our results to general evaluation metrics. Through experiments on benchmark image classification tasks, we show that our approach yields better trade-offs in both the balanced and worst-group error compared to L2R baselines.
****
Pengcheng Jiang,Cao Xiao,Adam Richard Cross,Jimeng Sun
GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs
Clinical predictive models often rely on patients’ electronic health records (EHR), but integrating medical knowledge to enhance predictions and decision-making is challenging. This is because personalized predictions require personalized knowledge
graphs (KGs), which are difficult to generate from patient EHR data. To address this, we propose GraphCare, an open-world framework that uses external KGs to improve EHR-based predictions. Our method extracts knowledge from large language models (LLMs) and external biomedical KGs to build patient-specific KGs, which are then used to train our proposed Bi-attention AugmenTed
(BAT) graph neural network (GNN) for healthcare predictions. On two public datas

ets, MIMIC-III and MIMIC-IV, GraphCare surpasses baselines in four vital healthcare prediction tasks: mortality, readmission, length of stay (LOS), and drug recommendation. On MIMIC-III, it boosts AUROC by 17.6% and 6.6% for mortality and readmission, and F1-score by 7.9% and 10.8% for LOS and drug recommendation, respectively. Notably, GraphCare demonstrates a substantial edge in scenarios with limited data availability. Our findings highlight the potential of using external KGs in healthcare prediction tasks and demonstrate the promise of GraphCare in generating personalized KGs for promoting personalized medicine.
****
Wenhao Zhan,Masatoshi Uehara,Nathan Kallus,Jason D. Lee,Wen Sun
Provable Offline Preference-Based Reinforcement Learning
In this paper, we investigate the problem of offline Preference-based Reinforcement Learning (PbRL) with human feedback where feedback is available in the form of preference between trajectory pairs rather than explicit rewards. Our proposed algorithm consists of two main steps: (1) estimate the implicit reward using Maximum Likelihood Estimation (MLE) with general function approximation from offline data and (2) solve a distributionally robust planning problem over a confidence set around the MLE. We consider the general reward setting where the reward can be defined over the whole trajectory and provide a novel guarantee that allows us to learn any target policy with a polynomial number of samples, as long as the target policy is covered by the offline data. This guarantee is the first of its kind with general function approximation. To measure the coverage of the target policy, we introduce a new single-policy concentrability coefficient, which can be upper bounded by the per-trajectory concentrability coefficient. We also establish lower bounds that highlight the necessity of such concentrability and the difference from standard RL, where state-action-wise rewards are directly observed. We further extend and analyze our algorithm when the feedback is given over action pairs.
****
Tianrong Chen,Jiatao Gu,Laurent Dinh,Evangelos Theodorou,Joshua M. Susskind,Shuangfei Zhai
Generative Modeling with Phase Stochastic Bridge
Diffusion models (DMs) represent state-of-the-art generative models for continuous inputs. DMs work by constructing a Stochastic Differential Equation (SDE) in the input space (ie, position space), and using a neural network to reverse it. In this work, we introduce a novel generative modeling framework grounded in \textbf{phase space dynamics}, where a phase space is defined as {an augmented space encompassing both position and velocity.} Leveraging insights from Stochastic Optimal Control, we construct a path measure in the phase space that enables efficient sampling. {In contrast to DMs, our framework demonstrates the capability to generate realistic data points at an early stage of dynamics propagation.} This early prediction sets the stage for efficient data generation by leveraging additional velocity information along the trajectory. On standard image generation benchmarks, our model yields favorable performance over baselines in the regime of small Number of Function Evaluations (NFEs). Furthermore, our approach rivals the performance of diffusion models equipped with efficient sampling techniques, underscoring its potential as a new tool generative modeling.
****
Abhra Chaudhuri,Serban Georgescu,Anjan Dutta
Learning Conditional Invariances through Non-Commutativity
Invariance learning algorithms that conditionally filter out domain-specific random variables as distractors, do so based only on the data semantics, and not the target domain under evaluation. We show that a provably optimal and sample-efficient way of learning conditional invariances is by relaxing the invariance criterion to be non-commutatively directed towards the target domain. Under domain asymmetry, i.e., when the target domain contains semantically relevant information absent in the source, the risk of the encoder φ^* that is optimal on average across domains is strictly lower-bounded by the risk of the target-specific optimal encoder Φ^*_τ. We prove that non-commutativity steers the optimization towards Φ^*_τ instead of φ^*, bringing the $\mathcal{

H}$-divergence between domains down to zero, leading to a stricter bound on the target risk. Both our theory and experiments demonstrate that non-commutative invariance (NCI) can leverage source domain samples to meet the sample complexity needs of learning Φ^*_τ, surpassing SOTA invariance learning algorithms for domain adaptation, at times by over 2\%, approaching the performance of an oracle. Implementation is available at https://github.com/abhrac/nci.
****
Tianbao Xie,Siheng Zhao,Chen Henry Wu,Yitao Liu,Qian Luo,Victor Zhong,Yanchao Yang,Tao Yu
Text2Reward: Reward Shaping with Language Models for Reinforcement Learning
Designing reward functions is a longstanding challenge in reinforcement learning (RL); it requires specialized knowledge or domain data, leading to high costs for development. To address this, we introduce Text2Reward, a data-free framework that automates the generation and shaping of dense reward functions based on large language models (LLMs). Given a goal described in natural language, Text2Reward generates shaped dense reward functions as an executable program grounded in a compact representation of the environment. Unlike inverse RL and recent work that uses LLMs to write sparse reward codes or unshaped dense rewards with a constant function across timesteps, Text2Reward produces interpretable, free-form dense reward codes that cover a wide range of tasks, utilize existing packages, and allow iterative refinement with human feedback. We evaluate Text2Reward on two robotic manipulation benchmarks (ManiSkill2, MetaWorld) and two locomotion environments of MuJoCo. On 13 of the 17 manipulation tasks, policies trained with generated reward codes achieve similar or better task success rates and convergence speed than expert-written reward codes. For locomotion tasks, our method learns six novel locomotion behaviors with a success rate exceeding 94%. Furthermore, we show that the policies trained in the simulator with our method can be deployed in the real world. Finally, Text2Reward further improves the policies by refining their reward functions with human feedback. Video results are available at https://text-to-reward.github.io/
****
Yingtian Zou,Kenji Kawaguchi,Yingnan Liu,Jiashuo Liu,Mong-Li Lee,Wynne Hsu
Towards Robust Out-of-Distribution Generalization Bounds via Sharpness
Generalizing to out-of-distribution (OOD) data or unseen domain, termed OOD generalization, still lacks appropriate theoretical guarantees. Canonical OOD bounds focus on different distance measurements between source and target domains but fail to consider the optimization property of the learned model. As empirically shown in recent work, sharpness of learned minimum influences OOD generalization. To bridge this gap between optimization and OOD generalization, we study the effect of sharpness on how a model tolerates data change in domain shift which is usually captured by "robustness" in generalization. In this paper, we give a rigorous connection between sharpness and robustness, which gives better OOD guarantees for robust algorithms. It also provides a theoretical backing for "flat minima leads to better OOD generalization". Overall, we propose a sharpness-based OOD generalization bound by taking robustness into consideration, resulting in a tighter bound than non-robust guarantees. Our findings are supported by the experiments on a ridge regression model, as well as the experiments on deep learning classification tasks.
****
Jiatao Gu,Shuangfei Zhai,Yizhe Zhang,Joshua M. Susskind,Navdeep Jaitly
Matryoshka Diffusion Models
Diffusion models are the de-facto approach for generating high-quality images and videos, but learning high-dimensional models remains a formidable task due to computational and optimization challenges. Existing methods often resort to training cascaded models in pixel space, or using a downsampled latent space of a separately trained auto-encoder. In this paper, we introduce Matryoshka Diffusion (MDM), an end-to-end framework for high-resolution image and video synthesis. We propose a diffusion process that denoises inputs at multiple resolutions jointly and uses a NestedUNet architecture where features and parameters for small-scale inputs are nested within those of large scales. In addition, MDM enables a pr

ogressive training schedule from lower to higher resolutions, which leads to sig
nificant improvements in optimization for high-resolution generation. We demonst
rate the effectiveness of our approach on various benchmarks, including class-co
nditioned image generation, high-resolution text-to-image, and text-to-video app
lications. Remarkably, we can train a single pixel-space model at resolutions of
 up to 1024x1024 pixels, demonstrating strong zero-shot generalization using the
 CC12M dataset, which contains only 12 million images.
****
Yihang Chen,Fanghui Liu,Yiping Lu,Grigorios Chrysos,Volkan Cevher
Generalization of Scaled Deep ResNets in the Mean-Field Regime
Despite the widespread empirical success of ResNet, the generalization propertie
s of deep ResNet are rarely explored beyond the lazy training regime. In this wo
rk, we investigate scaled ResNet in the limit of infinitely deep and wide neural
 networks, of which the gradient flow is described by a partial differential equ
ation in the large-neural network limit, i.e., the mean-field regime. To derive
the generalization bounds under this setting, our analysis necessitates a shift
from the conventional time-invariant Gram matrix employed in the lazy training r
egime to a time-variant, distribution-dependent version. To this end, we provide
 a global lower bound on the minimum eigenvalue of the Gram matrix under the mea
n-field regime. Besides, for the traceability of the dynamic of Kullback-Leibler
 (KL) divergence, we establish the linear convergence of the empirical error and
 estimate the upper bound of the KL divergence over parameters distribution. Fin
ally, we build the uniform convergence for generalization bound via Rademacher c
omplexity. Our results offer new insights into the generalization ability of dee
p ResNet beyond the lazy training regime and contribute to advancing the underst
anding of the fundamental properties of deep neural networks.
****
Siddarth Venkatraman,Shivesh Khaitan,Ravi Tej Akella,John Dolan,Jeff Schneider,G
len Berseth
Reasoning with Latent Diffusion in Offline Reinforcement Learning
Offline reinforcement learning (RL) holds promise as a means to learn high-rewar
d policies from a static dataset, without the need for further environment inter
actions. However, a key challenge in offline RL lies in effectively stitching po
rtions of suboptimal trajectories from the static dataset while avoiding extrapo
lation errors arising due to a lack of support in the dataset. Existing approach
es use conservative methods that are tricky to tune and struggle with multi-moda
l data or rely on noisy Monte Carlo return-to-go samples for reward conditioning
. In this work, we propose a novel approach that leverages the expressiveness of
 latent diffusion to model in-support trajectory sequences as compressed latent
skills. This facilitates learning a Q-function while avoiding extrapolation erro
r via batch-constraining. The latent space is also expressive and gracefully cop
es with multi-modal data. We show that the learned temporally-abstract latent sp
ace encodes richer task-specific information for offline RL tasks as compared to
 raw state-actions. This improves credit assignment and facilitates faster rewar
d propagation during Q-learning. Our method demonstrates state-of-the-art perfor
mance on the D4RL benchmarks, particularly excelling in long-horizon, sparse-rew
ard tasks.
****
Wei Zhuo,Zemin Liu,Bryan Hooi,Bingsheng He,Guang Tan,Rizal Fathony,Jia Chen
Partitioning Message Passing for Graph Fraud Detection
Label imbalance and homophily-heterophily mixture are the fundamental problems e
ncountered when applying Graph Neural Networks (GNNs) to Graph Fraud Detection (
GFD) tasks. Existing GNN-based GFD models are designed to augment graph structur
e to accommodate the inductive bias of GNNs towards homophily, by excluding hete
rophilic neighbors during message passing. In our work, we argue that the key to
 applying GNNs for GFD is not to exclude but to {\em distinguish} neighbors with
 different labels. Grounded in this perspective, we introduce Partitioning Messa
ge Passing (PMP), an intuitive yet effective message passing paradigm expressly
crafted for GFD. Specifically, in the neighbor aggregation stage of PMP, neighbo
rs with different classes are aggregated with distinct node-specific aggregation

functions. By this means, the center node can adaptively adjust the information aggregated from its heterophilic and homophilic neighbors, thus avoiding the model gradient being dominated by benign nodes which occupy the majority of the population. We theoretically establish a connection between the spatial formulation of PMP and spectral analysis to characterize that PMP operates an adaptive node-specific spectral graph filter, which demonstrates the capability of PMP to handle heterophily-homophily mixed graphs. Extensive experimental results show that PMP can significantly boost the performance on GFD tasks.
****
Mihaela C Stoian,Salijona Dyrmishi,Maxime Cordy,Thomas Lukasiewicz,Eleonora Giunchiglia
How Realistic Is Your Synthetic Data? Constraining Deep Generative Models for Tabular Data
Deep Generative Models (DGMs) have been shown to be powerful tools for generating tabular data, as they have been increasingly able to capture the complex distributions that characterize them. However, to generate realistic synthetic data, it is often not enough to have a good approximation of their distribution, as it also requires compliance with constraints that encode essential background knowledge on the problem at hand. In this paper, we address this limitation and show how DGMs for tabular data can be transformed into Constrained Deep Generative Models (C-DGMs), whose generated samples are guaranteed to be compliant with the given constraints. This is achieved by automatically parsing the constraints and transforming them into a Constraint Layer (CL) seamlessly integrated with the DGM. Our extensive experimental analysis with various DGMs and tasks reveals that standard DGMs often violate constraints, some exceeding 95% non-compliance, while their corresponding C-DGMs are never non-compliant. Then, we quantitatively demonstrate that, at training time, C-DGMs are able to exploit the background knowledge expressed by the constraints to outperform their standard counterparts with up to 4.5% improvement in utility and detection. Further, we show how our CL does not necessarily need to be integrated at training time, as it can be also used as a guardrail at inference time, still producing some improvements in the overall performance of the models. Finally, we show that our CL does not hinder the sample generation time of the models.
****
John Xavier Morris,Wenting Zhao,Justin T Chiu,Vitaly Shmatikov,Alexander M Rush
Language Model Inversion
Given a prompt, language models produce a distribution over all possible next tokens; when the prompt is unknown, can we use this distributional information to recover the prompt? We consider the problem of anguage model inversion and show that next-token probabilities contain a surprising amount of information about the preceding text. Often we can recover the text in cases where it is hidden from the user, motivating a method for recovering unknown prompts given only the model's current distribution output. We consider a variety of model access scenarios, and show how even without predictions for every token in the vocabulary we can recover the probability vector through search and reconstruction of the input. On LLAMA-7B, our inversion method reconstructs prompts with a BLEU of 59 and token-level F1 of 77 and recovers 23% of prompts exactly
****
Sahana Ramnath,Brihi Joshi,Skyler Hallinan,Ximing Lu,Liunian Harold Li,Aaron Chan,Jack Hessel,Yejin Choi,Xiang Ren
Tailoring Self-Rationalizers with Multi-Reward Distillation
Large language models (LMs) are capable of generating free-text rationales to aid question answering. However, prior work 1) suggests that useful self-rationalization is emergent only at significant scales (e.g., 175B parameter GPT-3); and 2) focuses largely on downstream performance, ignoring the semantics of the rationales themselves, e.g., are they faithful, true, and helpful for humans? In this work, we enable small-scale LMs (~200x smaller than GPT-3) to generate rationales that not only improve downstream task performance, but are also more plausible, consistent, and diverse, assessed both by automatic and human evaluation. Our method, MaRio (Multi-rewArd RatIOnalization), is a multi-reward conditioned se

lf-rationalization algorithm that optimizes multiple distinct properties like pl ausibility, diversity and consistency. Results on three difficult question-answe ring datasets StrategyQA, QuaRel and OpenBookQA show that not only does MaRio im prove task accuracy, but it also improves the self-rationalization quality of sm all LMs across the aforementioned axes better than a supervised fine-tuning (SFT) baseline. Extensive human evaluations confirm that MaRio rationales are prefer red vs. SFT rationales, as well as qualitative improvements in plausibility and consistency.
****
Hailey Joren,Charles Thomas Marx,Berk Ustun
Classification with Conceptual Safeguards
Machine learning models are often used to automate routine tasks. In settings wh ere mistakes are costly, we can trade off accuracy for coverage by abstaining fr om making a prediction on instances for which the model is uncertain. In this wo rk, we present a new approach to selective classification in deep learning with concepts. Our approach constructs a concept bottleneck model where the front-end model can make predictions given soft concepts and leverage concept confirmatio n to improve coverage and performance under abstention. We develop techniques to propagate uncertainty and identify concepts for confirmation. We evaluate our a pproach on real-world and synthetic datasets, showing that it can improve covera ge while maintaining performance across a range of tasks.
****
Ziwei Luo,Fredrik K. Gustafsson,Zheng Zhao,Jens Sjölund,Thomas B. Schön
Controlling Vision-Language Models for Multi-Task Image Restoration
Vision-language models such as CLIP have shown great impact on diverse downstrea m tasks for zero-shot or label-free predictions. However, when it comes to low-l evel vision such as image restoration their performance deteriorates dramaticall y due to corrupted inputs. In this paper, we present a degradation-aware vision-language model (DA-CLIP) to better transfer pretrained vision-language models to low-level vision tasks as a multi-task framework for image restoration. More sp ecifically, DA-CLIP trains an additional controller that adapts the fixed CLIP i mage encoder to predict high-quality feature embeddings. By integrating the embe dding into an image restoration network via cross-attention, we are able to pilo t the model to learn a high-fidelity image reconstruction. The controller itself will also output a degradation feature that matches the real corruptions of the input, yielding a natural classifier for different degradation types. In additi on, we construct a mixed degradation dataset with synthetic captions for DA-CLIP training. Our approach advances state-of-the-art performance on both degradatio n-specific and unified image restoration tasks, showing a promising direction of prompting image restoration with large-scale pretrained vision-language models. Our code is available at https://github.com/Algolzw/daclip-uir.
****
Ziqi Pang,Ziyang Xie,Yunze Man,Yu-Xiong Wang
Frozen Transformers in Language Models Are Effective Visual Encoder Layers
This paper reveals that large language models (LLMs), despite being trained sole ly on text data, are \emph{surprisingly} strong encoders for \emph{purely} visua l tasks in the absence of language. Even more intriguingly, this can be achieved by a simple yet previously overlooked strategy -- employing a \emph{frozen} tra nsformer block from \emph{pre-trained} LLMs as a constituent encoder layer to di rectly process visual tokens. Our work pushes the boundaries of leveraging LLMs for computer vision tasks, significantly departing from conventional practices t hat typically necessitate a multi-modal vision-language setup with associated la nguage prompts, inputs, or outputs. We demonstrate that our approach consistentl y enhances performance across \emph{a diverse range of tasks}, encompassing pure 2D or 3D visual recognition tasks (e.g., image and point cloud classification), temporal modeling tasks (e.g., action recognition), non-semantic tasks (e.g., m otion forecasting), and multi-modal tasks (e.g., 2D/3D visual question answering and image-text retrieval). Such improvements are a general phenomenon, applicab le to various types of LLMs (e.g., LLaMA and OPT) and different LLM transformer blocks. We additionally propose the \emph{information filtering} hypothesis to e

xplain the effectiveness of pre-trained LLMs in visual encoding -- the pre-train
ed LLM transformer blocks discern informative visual tokens and further amplify
their effect. This hypothesis is empirically supported by the observation that t
he feature activation, after training with LLM transformer blocks, exhibits a st
ronger focus on relevant regions. We hope that our work inspires new perspective
s on utilizing LLMs and deepening our understanding of their underlying mechanis
ms.
****
Han Zhang,Xiaofan Gui,Shun Zheng,Ziheng Lu,Yuqi Li,Jiang Bian
BatteryML: An Open-source Platform for Machine Learning on Battery Degradation
Battery degradation remains a pivotal concern in the energy storage domain, with
 machine learning emerging as a potent tool to drive forward insights and soluti
ons. However, this intersection of electrochemical science and machine learning
poses complex challenges. Machine learning experts often grapple with the intric
acies of battery science, while battery researchers face hurdles in adapting int
ricate models tailored to specific datasets. Beyond this, a cohesive standard fo
r battery degradation modeling, inclusive of data formats and evaluative benchma
rks, is conspicuously absent. Recognizing these impediments, we present Battery
ML—a one-step, all-encompass, and open-source platform designed to unify data p
reprocessing, feature extraction, and the implementation of both traditional and
 state-of-the-art models. This streamlined approach promises to enhance the prac
ticality and efficiency of research applications. BatteryML seeks to fill this v
oid, fostering an environment where experts from diverse specializations can col
laboratively contribute, thus elevating the collective understanding and advance
ment of battery research.
****
Jianhao Yuan,Jie Zhang,Shuyang Sun,Philip Torr,Bo Zhao
Real-Fake: Effective Training Data Synthesis Through Distribution Matching
Synthetic training data has gained prominence in numerous learning tasks and sce
narios, offering advantages such as dataset augmentation, generalization evalua■
tion, and privacy preservation. Despite these benefits, the efficiency of synthe
tic data generated by
current methodologies remains inferior when training advanced deep models exclus
ively, limiting its practical utility. To address this challenge, we analyze the
 principles underlying training data synthesis for supervised learning and eluci
date a principled theoretical framework from the distribution-matching perspecti
ve that explicates the mechanisms governing synthesis efficacy. Through extensiv
e experiments, we demonstrate the effectiveness of our synthetic data across div
erse image classification tasks, both as a replacement for and augmenta■tion to
real datasets, while also benefits such as out-of-distribution generalization, p
rivacy preservation, and scalability. Specifically, we achieve 70.9% top1 classi
fication accuracy on ImageNet1K when training solely with synthetic data equival
ent
to 1 × the original real data size, which increases to 76.0% when scaling up to
10 × synthetic data.
****
Tianyuan Zou,Zixuan GU,Yu He,Hideaki Takahashi,Yang Liu,Ya-Qin Zhang
VFLAIR: A Research Library and Benchmark for Vertical Federated Learning
Vertical Federated Learning (VFL) has emerged as a collaborative training paradi
gm that allows participants with different features of the same group of users t
o accomplish cooperative training without exposing their raw data or model param
eters. VFL has gained significant attention for its research potential and real-
world applications in recent years, but still faces substantial challenges, such
 as in defending various kinds of data inference and backdoor attacks. Moreover,
 most of existing VFL projects are industry-facing and not easily used for keepi
ng track of the current research progress. To address this need, we present an e
xtensible and lightweight VFL framework VFLAIR (available at https://github.com/
FLAIR-THU/VFLAIR), which supports VFL training with a variety of models, dataset
s and protocols, along with standardized modules for comprehensive evaluations o
f attacks and defense strategies. We also benchmark 11 attacks and 8 defense

s performance under different communication and model partition settings and draw concrete insights and recommendations on the choice of defense strategies for different practical VFL deployment scenarios.
****
Jianhao Shen,Ye Yuan,Srbuhi Mirzoyan,Ming Zhang,Chenguang Wang
Measuring Vision-Language STEM Skills of Neural Models
We introduce a new challenge to test the STEM skills of neural models. The problems in the real world often require solutions, combining knowledge from STEM (science, technology, engineering, and math). Unlike existing datasets, our dataset requires the understanding of multimodal vision-language information of STEM. Our dataset features one of the largest and most comprehensive datasets for the challenge. It includes 448 skills and $1,073,146$ questions spanning all STEM subjects. Compared to existing datasets that often focus on examining expert-level ability, our dataset includes fundamental skills and questions designed based on the K-12 curriculum. We also add state-of-the-art foundation models such as CLIP and GPT-3.5-Turbo to our benchmark. Results show that the recent model advances only help master a very limited number of lower grade-level skills (2.5% in the third grade) in our dataset. In fact, these models are still well below (averaging 54.7%) the performance of elementary students, not to mention near expert-level performance. To understand and increase the performance on our dataset, we teach the models on a training split of our dataset.
Even though we observe improved performance, the model performance remains relatively low compared to average elementary students. To solve STEM problems, we will need novel algorithmic innovations from the community.
****
Henry Li,Ronen Basri,Yuval Kluger
Likelihood Training of Cascaded Diffusion Models via Hierarchical Volume-preserving Maps
Cascaded models are multi-scale generative models with a marked capacity for producing perceptually impressive samples at high resolutions. In this work, we show that they can also be excellent likelihood models, so long as we overcome a fundamental difficulty with probabilistic multi-scale models: the intractability of the likelihood function. Chiefly, in cascaded models each intermediary scale introduces extraneous variables that cannot be tractably marginalized out for likelihood evaluation. This issue vanishes by modeling the diffusion process on latent spaces induced by a class of transformations we call hierarchical volume-preserving maps, which decompose spatially structured data in a hierarchical fashion without introducing local distortions in the latent space. We demonstrate that two such maps are well-known in the literature for multiscale modeling: Laplacian pyramids and wavelet transforms. Not only do such reparameterizations allow the likelihood function to be directly expressed as a joint likelihood over the scales, we show that the Laplacian pyramid and wavelet transform also produces significant improvements to the state-of-the-art on a selection of benchmarks in likelihood modeling, including density estimation, lossless compression, and out-of-distribution detection. Investigating the theoretical basis of our empirical gains we uncover deep connections to score matching under the Earth Mover's Distance (EMD), which is a well-known surrogate for perceptual similarity.
****
Ziyu Wang,Lejun Min,Gus Xia
Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models
Recent deep music generation studies have put much emphasis on *music structure* and *long-term* generation. However, we are yet to see high-quality, well-structured whole-song generation. In this paper, we make the first attempt to model a full music piece under the realization of *compositional hierarchy*. With a focus on symbolic representations of pop songs, we define a hierarchical language, in which each level of hierarchy focuses on the context dependency at a certain music scope. The high-level languages reveal whole-song form, phrase, and cadence, whereas the low-level languages focus on notes, chords, and their local patterns. A cascaded diffusion model is trained to model the hierarchical language, w

here each level is conditioned on its upper levels. Experiments and analysis sho
w that our model is capable of generating full-piece music with recognizable glo
bal verse-chorus structure and cadences, and the music quality is higher than th
e baselines. Additionally, we show that the proposed model is *controllable* in
a flexible way. By sampling from the interpretable hierarchical languages or adj
usting external controls, users can control the music flow via various features
such as phrase harmonic structures, rhythmic patterns, and accompaniment texture
.
****
Ivan Grega,Ilyes Batatia,Gabor Csanyi,Sri Karlapati,Vikram Deshpande
Energy-conserving equivariant GNN for elasticity of lattice architected metamate
rials
Lattices are architected metamaterials whose properties strongly depend on their
 geometrical design. The analogy between lattices and graphs enables the use of
graph neural networks (GNNs) as a faster surrogate model compared to traditional
 methods such as finite element modelling. In this work, we generate a big datas
et of structure-property relationships for strut-based lattices. The dataset is
made available to the community which can fuel the development of methods anchor
ed in physical principles for the fitting of fourth-order tensors. In addition,
we present a higher-order GNN model trained on this dataset. The key features of
 the model are (i) SE(3) equivariance, and (ii) consistency with the thermodynam
ic law of conservation of energy. We compare the model to non-equivariant models
 based on a number of error metrics and demonstrate its benefits in terms of pre
dictive performance and reduced training requirements. Finally, we demonstrate a
n example application of the model to an architected material design task. The m
ethods which we developed are applicable to fourth-order tensors beyond elastici
ty such as piezo-optical tensor etc.
****
Yicong Hong,Kai Zhang,Jiuxiang Gu,Sai Bi,Yang Zhou,Difan Liu,Feng Liu,Kalyan Sun
kavalli,Trung Bui,Hao Tan
LRM: Large Reconstruction Model for Single Image to 3D
We propose the first Large Reconstruction Model (LRM) that predicts the 3D model
 of an object from a single input image within just 5 seconds. In contrast to ma
ny previous methods that are trained on small-scale datasets such as ShapeNet in
 a category-specific fashion, LRM adopts a highly scalable transformer-based arc
hitecture with 500 million learnable parameters to directly predict a neural rad
iance field (NeRF) from the input image. We train our model in an end-to-end man
ner on massive multi-view data containing around 1 million objects, including bo
th synthetic renderings from Objaverse and real captures from MVImgNet. This com
bination of a high-capacity model and large-scale training data empowers our mod
el to be highly generalizable and produce high-quality 3D reconstructions from v
arious testing inputs, including real-world in-the-wild captures and images crea
ted by generative models. Video demos and interactable 3D meshes can be found on
 our LRM project webpage: https://yiconghong.me/LRM.
****
Yuhan Helena Liu,Aristide Baratin,Jonathan Cornford,Stefan Mihalas,Eric Todd She
aBrown,Guillaume Lajoie
How connectivity structure shapes rich and lazy learning in neural circuits
In theoretical neuroscience, recent work leverages deep learning tools to explor
e how some network attributes critically influence its learning dynamics. Notabl
y, initial weight distributions with small (resp. large) variance may yield a ri
ch (resp. lazy) regime, where significant (resp. minor) changes to network state
s and representation are observed over the course of learning. However, in biolo
gy, neural circuit connectivity generally has a low-rank structure and therefore
 differs markedly from the random initializations generally used for these studi
es. As such, here we investigate how the structure of the initial weights — in p
articular their effective rank — influences the network learning regime. Through
 both empirical and theoretical analyses, we discover that high-rank initializat
ions typically yield smaller network changes indicative of lazier learning, a fi
nding we also confirm with experimentally-driven initial connectivity in recurre

nt neural networks. Conversely, low-rank initialization biases learning towards richer learning. Importantly, however, as an exception to this rule, we find lazier learning can still occur with a low-rank initialization that aligns with task and data statistics. Our research highlights the pivotal role of initial weight structures in shaping learning regimes, with implications for metabolic costs of plasticity and risks of catastrophic forgetting.
****
Amirhossein Vahidi,Simon Schosser,Lisa Wimmer,Yawei Li,Bernd Bischl,Eyke Hüllermeier,Mina Rezaei
Probabilistic Self-supervised Representation Learning via Scoring Rules Minimization
%
Self-supervised learning methods have shown promising results across a wide range of tasks in computer vision, natural language processing, and multimodal analysis. However, self-supervised approaches come with a notable limitation, dimensional collapse, where a model doesn't fully utilize its capacity to encode information optimally. Motivated by this, we propose ProSMin, a novel probabilistic self-supervised learning approach that leverages the power of probabilistic models to enhance representation quality and mitigate collapsing representations. Our proposed approach involves two neural networks, the online network and the target network, which collaborate and learn the diverse distribution of representations from each other through probabilistic knowledge distillation. The two networks are trained via our new loss function based on proper scoring rules. We provide a theoretical justification for ProSMin and demonstrate its modified scoring rule. This insight validates the method's optimization process and contributes to its robustness and effectiveness in improving representation quality. We evaluate our probabilistic model on various downstream tasks, such as in-distribution generalization, out-of-distribution detection, dataset corruption, low-shot learning, and transfer learning. Our method achieves superior accuracy and calibration, outperforming the self-supervised baseline in a variety of experiments on large datasets such as ImageNet-O and ImageNet-C. ProSMin thus demonstrates its scalability and real-world applicability. Our code is publicly available: https://github.com/amirvhd/SSL-sore-rule.
****
Chujie Zheng,Hao Zhou,Fandong Meng,Jie Zhou,Minlie Huang
Large Language Models Are Not Robust Multiple Choice Selectors
Multiple choice questions (MCQs) serve as a common yet important task format in the evaluation of large language models (LLMs). This work shows that modern LLMs are vulnerable to option position changes in MCQs due to their inherent “selection bias”, namely, they prefer to select specific option IDs as answers (like “Option A”). Through extensive empirical analyses with 20 LLMs on three benchmarks, we pinpoint that this behavioral bias primarily stems from LLMs’ token bias, where the model a priori assigns more probabilistic mass to specific option ID tokens (e.g., A/B/C/D) when predicting answers from the option IDs. To mitigate selection bias, we propose a label-free, inference-time debiasing method, called PriDe, which separates the model’s prior bias for option IDs from the overall prediction distribution. PriDe first estimates the prior by permutating option contents on a small number of test samples, and then applies the estimated prior to debias the remaining samples. We demonstrate that it achieves interpretable and transferable debiasing with high computational efficiency. We hope this work can draw broader research attention to the bias and robustness of modern LLMs.
****
Maxim Khanov,Jirayu Burapacheep,Yixuan Li
ARGS: Alignment as Reward-Guided Search
Aligning large language models with human objectives is paramount, yet common approaches including RLHF suffer from unstable and resource-intensive training. In response to this challenge, we introduce ARGS, Alignment as Reward-Guided Search, a novel framework that integrates alignment into the decoding process, eliminating the need for expensive RL training. By adjusting the model’s probabilistic predictions using a reward signal, ARGS generates texts with semantic diversity

while being aligned with human preferences, offering a promising and flexible solution for aligning language models. Notably, our method demonstrates consistent enhancements in average reward compared to baselines across diverse alignment tasks and various model dimensions. For example, under the same greedy-based decoding strategy, our method improves the average reward by 19.56% relative to the baseline and secures a preference or tie score of 64.33% in GPT-4 evaluation. We believe that our framework, emphasizing test-time alignment, paves the way for more responsive language models in the future. Code is publicly available at: https://github.com/deeplearning-wisc/args.
****
Chau Pham,Boyi Liu,Yingxiang Yang,Zhengyu Chen,Tianyi Liu,Jianbo Yuan,Bryan A. Plummer,Zhaoran Wang,Hongxia Yang
Let Models Speak Ciphers: Multiagent Debate through Embeddings
Discussion and debate among Large Language Models (LLMs) have gained considerable attention due to their potential to enhance the reasoning ability of LLMs. Although natural language is an obvious choice for communication due to LLM's language understanding capability, the token sampling step needed when generating natural language poses a potential risk of information loss, as it uses only one token to represent the model's belief across the entire vocabulary. In this paper, we introduce a communication regime named CIPHER (Communicative Inter-Model Protocol Through Embedding Representation) to address this issue. Specifically, we remove the token sampling step from LLMs and let them communicate their beliefs across the vocabulary through the expectation of the raw transformer output embeddings. Remarkably, by deviating from natural language, CIPHER offers an advantage of encoding a broader spectrum of information without any modification to the model weights, outperforming the state-of-the-art LLM debate methods using natural language by 0.5-5.0% across five reasoning tasks and multiple open-source LLMs of varying sizes. This showcases the superiority and robustness of embeddings as an alternative "language" for communication among LLMs. We anticipate that CIPHER will inspire further exploration for the design of interactions within LLM agent systems, offering a new direction that could significantly influence future developments in the field.
****
Wenxi Wang,Yang Hu,Mohit Tiwari,Sarfraz Khurshid,Kenneth McMillan,Risto Miikkulainen
NeuroBack: Improving CDCL SAT Solving using Graph Neural Networks
Propositional satisfiability (SAT) is an NP-complete problem that impacts many research fields, such as planning, verification, and security. Mainstream modern SAT solvers are based on the Conflict-Driven Clause Learning (CDCL) algorithm. Recent work aimed to enhance CDCL SAT solvers using Graph Neural Networks (GNNs). However, so far this approach either has not made solving more effective, or required substantial GPU resources for frequent online model inferences. Aiming to make GNN improvements practical, this paper proposes an approach called NeuroBack, which builds on two insights: (1) predicting phases (i.e., values) of variables appearing in the majority (or even all) of the satisfying assignments are essential for CDCL SAT solving, and (2) it is sufficient to query the neural model only once for the predictions before the SAT solving starts. Once trained, the offline model inference allows NeuroBack to execute exclusively on the CPU, removing its reliance on GPU resources. To train NeuroBack, a new dataset called DataBack containing 120,286 data samples is created. Finally, NeuroBack is implemented as an enhancement to a state-of-the-art SAT solver called Kissat. As a result, it allowed Kissat to solve 5.2% more problems on the recent SAT competition problem set, SATCOMP-2022. NeuroBack therefore shows how machine learning can be harnessed to improve SAT solving in an effective and practical manner.
****

Omar Khattab,Arnav Singhvi,Paridhi Maheshwari,Zhiyuan Zhang,Keshav Santhanam,Sri Vardhamanan A,Saiful Haq,Ashutosh Sharma,Thomas T. Joshi,Hanna Moazam,Heather Miller,Matei Zaharia,Christopher Potts
DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines
The ML community is rapidly exploring techniques for prompting language models (LMs) and for stacking them into pipelines that solve complex tasks. Unfortunately, existing LM pipelines are typically implemented using hard-coded “prompt templates”, i.e. lengthy strings discovered via trial and error. Toward a more systematic approach for developing and optimizing LM pipelines, we introduce DSPy, a programming model that abstracts LM pipelines as text transformation graphs, or imperative computational graphs where LMs are invoked through declarative modules. DSPy modules are parameterized, meaning they can learn how to apply compositions of prompting, finetuning, augmentation, and reasoning techniques. We design a compiler that will optimize any DSPy pipeline to maximize a given metric, by creating and collecting demonstrations. We conduct two case studies, showing that succinct DSPy programs can express and optimize pipelines that reason about math word problems, tackle multi-hop retrieval, answer complex questions, and control agent loops. Within minutes of compiling, DSPy can automatically produce pipelines that outperform out-of-the-box few-shot prompting as well as expert-created demonstrations for GPT-3.5 and Llama2-13b-chat. On top of that, DSPy programs compiled for relatively small LMs like 770M parameter T5 and Llama2-13b-chat are competitive with many approaches that rely on large and proprietary LMs like GPT-3.5 and on expert-written prompt chains. DSPy is available at https://github.com/stanfordnlp/dspy
****
Nicholas Corrado,Josiah P. Hanna
Understanding when Dynamics-Invariant Data Augmentations Benefit Model-free Reinforcement Learning Updates
Recently, data augmentation (DA) has emerged as a method for leveraging domain knowledge to inexpensively generate additional data in reinforcement learning (RL) tasks, often yielding substantial improvements in data efficiency.
While prior work has demonstrated the utility of incorporating augmented data directly into model-free RL updates,
it is not well-understood when a particular DA strategy will improve data efficiency.
In this paper, we seek to identify general aspects of DA responsible for observed learning improvements.
Our study focuses on sparse-reward tasks with dynamics-invariant data augmentation functions, serving as an initial step towards a more general understanding of DA and its integration into RL training.
Experimentally, we isolate three relevant aspects of DA: state-action coverage, reward density, and the number of augmented transitions generated per update (the augmented replay ratio).
From our experiments, we draw two conclusions: (1) increasing state-action coverage often has a much greater impact on data efficiency than increasing reward density, and (2) decreasing the augmented replay ratio substantially improves data efficiency.
In fact, certain tasks in our empirical study are solvable only when the replay ratio is sufficiently low.
****
Arian Rokkum Jamasb,Alex Morehead,Zuobai Zhang,Chaitanya K. Joshi,Kieran Didi,Simon V Mathis,Charles Harris,Jian Tang,Jianlin Cheng,Pietro Lio,Tom Leon Blundell
Evaluating Representation Learning on the Protein Structure Universe
Protein structure representation learning is the foundation for promising applications in drug discovery, protein design, and protein function prediction. However, there remains a need for a robust, standardised benchmark to track the progress of new and established methods with greater granularity and relevance to downstream applications. In this work, we introduce a comprehensive and open benchmark suite for evaluating protein structure representation learning methods.

We provide several pre-training methods, downstream tasks and pre-training corpo
ra comprised of both experimental and predicted structures, offering a balanced
challenge to representation learning algorithms. These tasks enable the systemat
ic evaluation of the quality of the learned embeddings, the structural and funct
ional relationships captured, and their usefulness in downstream tasks. We bench
mark state-of-the-art protein-specific and generic geometric Graph Neural Networ
ks and the extent to which they benefit from different types of pre-training. We
 find that pre-training consistently improves the performance of both rotation-i
nvariant and equivariant models, and that equivariant models seem to benefit eve
n more from pre-training compared to invariant models.

We aim to establish a common ground for the machine learning and computational b
iology communities to collaborate, compare, and advance protein structure repres
entation learning. By providing a standardised and rigorous evaluation platform,
 we expect to accelerate the development of novel methodologies and improve our
understanding of protein structures and their functions. The codebase incorporat
es several engineering contributions which considerably reduces the barrier to e
ntry for pre-training and working with large structure-based datasets. Our bench
mark is available at: https://anonymous.4open.science/r/ProteinWorkshop-B8F5/
****
Grzegorz Rype■■,Sebastian Cygert,Valeriya Khan,Tomasz Trzcinski,Bartosz Micha■ Z
ieli■ski,Bart■omiej Twardowski
Divide and not forget: Ensemble of selectively trained experts in Continual Le
arning
Class-incremental learning is becoming more popular as it helps models widen the
ir applicability while not forgetting what they already know. A trend in this ar
ea is to use a mixture-of-expert technique, where different models work together
 to solve the task. However, the experts are usually trained all at once using w
hole task data, which makes them all prone to forgetting and increasing computat
ional burden. To address this limitation, we introduce a novel approach named SE
ED. SEED selects only one, the most optimal expert for a considered task, and us
es data from this task to fine-tune only this expert. For this purpose, each exp
ert represents each class with a Gaussian distribution, and the optimal expert i
s selected based on the similarity of those distributions. Consequently, SEED in
creases diversity and heterogeneity within the experts while maintaining the hig
h stability of this ensemble method. The extensive experiments demonstrate that
SEED achieves state-of-the-art performance in exemplar-free settings across vari
ous scenarios, showing the potential of expert diversification through data in c
ontinual learning.
****
Zhixiang Chi,Li Gu,Tao Zhong,Huan Liu,YUANHAO YU,Konstantinos N Plataniotis,Yang
 Wang
Adapting to Distribution Shift by Visual Domain Prompt Generation
In this paper, we aim to adapt a model at test-time using a few unlabeled data t
o address distribution shifts.
To tackle the challenges of extracting domain knowledge from a limited amount of
 data, it is crucial to utilize correlated information from pre-trained backbone
s and source domains. Previous studies fail to utilize recent foundation models
with strong out-of-distribution generalization. Additionally, domain-centric des
igns are not flavored in their works. Furthermore, they employ the process of mo
delling source domains and the process of learning to adapt independently into d
isjoint training stages. In this work, we propose an approach on top of the pre-
computed features of the foundation model. Specifically, we build a knowledge ba
nk to learn the transferable knowledge from source domains. Conditioned on few-s
hot target data, we introduce a domain prompt generator to condense the knowledg
e bank into a domain-specific prompt. The domain prompt then directs the visual
features towards a particular domain via a guidance module. Moreover, we propose
 a domain-aware contrastive loss and employ meta-learning to facilitate domain k
nowledge extraction. Extensive experiments are conducted to validate the domain
knowledge extraction. The proposed method outperforms previous work on 5 large-s

cale benchmarks including WILDS and DomainNet.
****
Yanqin Jiang,Li Zhang,Jin Gao,Weiming Hu,Yao Yao
Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video
In this paper, we present Consistent4D, a novel approach for generating 4D dynamic objects from uncalibrated monocular videos. Uniquely, we cast the 360-degree dynamic object reconstruction as a 4D generation problem, eliminating the need for tedious multi-view data collection and camera calibration. This is achieved by leveraging the object-level 3D-aware image diffusion model as the primary supervision signal for training dynamic Neural Radiance Fields (DyNeRF). Specifically, we propose a cascade DyNeRF to facilitate stable convergence and temporal continuity under the supervision signal which is discrete along the time axis. To achieve spatial and temporal consistency, we further introduce an interpolation-driven consistency loss. It is optimized by minimizing the L2 distance between rendered frames from DyNeRF and interpolated frames from a pre-trained video interpolation model. Extensive experiments show that our Consistent4D can perform competitively to prior art alternatives, opening up new possibilities for 4D dynamic object generation from monocular videos, whilst also demonstrating advantage for conventional text-to-3D generation tasks
****
Nick Richardson,Deniz Oktay,Yaniv Ovadia,James C Bowden,Ryan P Adams
Fiber Monte Carlo
Integrals with discontinuous integrands are ubiquitous, arising from discrete structure in applications like topology optimization, graphics, and computational geometry.

These integrals are often part of a forward model in an inverse problem where it is necessary to reason backwards about the parameters, ideally using gradient-based optimization.

Monte Carlo methods are widely used to estimate the value of integrals, but this results in a non-differentiable approximation that is amenable to neither conventional automatic differentiation nor reparameterization-based gradient methods.

This significantly disrupts efforts to integrate machine learning methods in areas that exhibit these discontinuities: physical simulation and robotics, design, graphics, and computational geometry.

Although bespoke domain-specific techniques can handle special cases, a general methodology to wield automatic differentiation in these discrete contexts is wanting.

We introduce a differentiable variant of the simple Monte Carlo estimator which samples line segments rather than points from the domain.

We justify our estimator analytically as conditional Monte Carlo and demonstrate the diverse functionality of the method as applied to image stylization, topology optimization, and computational geometry.
****
Dong Wei,Huaijiang Sun,Bin Li,Xiaoning Sun,Shengxiang Hu,Weiqing Li,Jianfeng Lu
NeRM: Learning Neural Representations for High-Framerate Human Motion Synthesis
Generating realistic human motions with high framerate is an underexplored task, due to the varied framerates of training data, huge memory burden brought by high framerates and slow sampling speed of generative models. Recent advances make a compromise for training by downsampling high-framerate details away and discarding low-framerate samples, which suffer from severe information loss and restricted-framerate generation. In this paper, we found that the recent emerging paradigm of Implicit Neural Representations (INRs) that encode a signal into a continuous function can effectively tackle this challenging problem. To this end, we introduce NeRM, a generative model capable of taking advantage of varied-size data and capturing variational distribution of motions for high-framerate motion synthesis. By optimizing latent representation and a auto-decoder conditioned on temporal coordinates, NeRM learns continuous motion fields of sampled motion clips that ingeniously avoid explicit modeling of raw varied-size motions. This expressive latent representation is then used to learn a diffusion model that enab

les both unconditional and conditional generation of human motions. We demonstrate that our approach achieves competitive results with state-of-the-art methods, and can generate arbitrary framerate motions. Additionally, we show that NeRM is not only memory-friendly, but also highly efficient even when generating high-framerate motions.
****
Xiyuan Wang,Haotong Yang,Muhan Zhang
Neural Common Neighbor with Completion for Link Prediction
In this work, we propose a novel link prediction model and further boost it by studying graph incompleteness. First, We introduce MPNN-then-SF, an innovative architecture leveraging structural feature (SF) to guide MPNN's representation pooling, with its implementation, namely Neural Common Neighbor (NCN). NCN exhibits superior expressiveness and scalability compared with existing models, which can be classified into two categories: SF-then-MPNN, augmenting MPNN's input with SF, and SF-and-MPNN, decoupling SF and MPNN. Second, we investigate the impact of graph incompleteness---the phenomenon that some links are unobserved in the input graph---on SF, like the common neighbor. Through dataset visualization, we observe that incompleteness reduces common neighbors and induces distribution shifts, significantly affecting model performance. To address this issue, we propose to use a link prediction model to complete the common neighbor structure. Combining this method with NCN, we propose Neural Common Neighbor with Completion (NCNC). NCN and NCNC outperform recent strong baselines by large margins, and NCNC further surpasses state-of-the-art models in standard link prediction benchmarks.
****
Alexander Theus,Olin Geimer,Friedrich Wicke,Thomas Hofmann,Sotiris Anagnostidis, Sidak Pal Singh
Towards Meta-Pruning via Optimal Transport
Structural pruning of neural networks conventionally relies on identifying and discarding less important neurons, a practice often resulting in significant accuracy loss that necessitates subsequent fine-tuning efforts. This paper introduces a novel approach named Intra-Fusion, challenging this prevailing pruning paradigm.
Unlike existing methods that focus on designing meaningful neuron importance metrics, Intra-Fusion redefines the overlying pruning procedure.
Through utilizing the concepts of model fusion and Optimal Transport, we leverage an agnostically given importance metric to arrive at a more effective sparse model representation.
Notably, our approach achieves substantial accuracy recovery without the need for resource-intensive fine-tuning, making it an efficient and promising tool for neural network compression.
Additionally, we explore how fusion can be added to the pruning process to significantly decrease the training time while maintaining competitive performance. We benchmark our results for various networks on commonly used datasets such as CIFAR-10, CIFAR-100, and ImageNet. More broadly, we hope that the proposed Intra-Fusion approach invigorates exploration into a fresh alternative to the predominant compression approaches.
Our code is available [here](https://github.com/alexandertheus/Intra-Fusion).
****
Chaohua Shi,Kexin Huang,Lu GAN,Hongqing Liu,Mingrui Zhu,Nannan Wang,Xinbo Gao
On the Analysis of GAN-based Image-to-Image Translation with Gaussian Noise Injection
Image-to-image (I2I) translation is vital in computer vision tasks like style transfer and domain adaptation. While recent advances in GAN have enabled high-quality sample generation, real-world challenges such as noise and distortion remain significant obstacles. Although Gaussian noise injection during training has been utilized, its theoretical underpinnings have been unclear. This work provides a robust theoretical framework elucidating the role of Gaussian noise injection in I2I translation models. We address critical questions on the influence of noise variance on distribution divergence, resilience to unseen noise types, and

optimal noise intensity selection. Our contributions include connecting f-dive rgence and score matching, unveiling insights into the impact of Gaussian noise on aligning probability distributions, and demonstrating generalized robustness implications. We also explore choosing an optimal training noise level for consi stent performance in noisy environments. Extensive experiments validate our theo retical findings, showing substantial improvements over various I2I baseline mod els in noisy settings. Our research rigorously grounds Gaussian noise injection for I2I translation, offering a sophisticated theoretical understanding beyond h euristic applications.
****
Reese Pathak,Rajat Sen,Weihao Kong,Abhimanyu Das
Transformers can optimally learn regression mixture models
Mixture models arise in many regression problems, but most methods have seen lim ited adoption partly due to these algorithms' highly-tailored and model-specific nature. On the other hand, transformers are flexible, neural sequence models th at present the intriguing possibility of providing general-purpose prediction me thods, even in this mixture setting. In this work, we investigate the hypothesis that transformers can learn an optimal predictor for mixtures of regressions. W e construct a generative process for a mixture of linear regressions for which t he decision-theoretic optimal procedure is given by data-driven exponential weig hts on a finite set of parameters. We observe that transformers achieve low mean -squared error on data generated via this process. By probing the transformer's output at inference time, we also show that transformers typically make predicti ons that are close to the optimal predictor. Our experiments also demonstrate th at transformers can learn mixtures of regressions in a sample-efficient fashion and are somewhat robust to distribution shifts. We complement our experimental o bservations by proving constructively that the decision-theoretic optimal proced ure is indeed implementable by a transformer.
****
Kun Wang,Hao Wu,Yifan Duan,Guibin Zhang,Kai Wang,Xiaojiang Peng,Yu Zheng,Yuxuan Liang,Yang Wang
NuwaDynamics: Discovering and Updating in Causal Spatio-Temporal Modeling
Spatio-temporal (ST) prediction plays a pivotal role in earth sciences, such as meteorological prediction, urban computing. Adequate high-quality data, coupled with deep models capable of inference, are both indispensable and prerequisite f or achieving meaningful results. However, the sparsity of data and the high cost s associated with deploying sensors lead to significant data imbalances. Models that are overly tailored and lack causal relationships further compromise the ge neralizabilities of inference methods. Towards this end, we first establish a ca usal concept for ST predictions, named NuwaDynamics, which targets to identify causal regions in data and endow model with causal reasoning ability in a two-st age process. Concretely, we initially leverage upstream self-supervision to disc ern causal important patches, imbuing the model with generalized information and conducting informed interventions on complementary trivial patches to extrapola te potential test distributions. This phase is referred to as the discovery step . Advancing beyond discovery step, we transfer the data to downstream tasks for targeted ST objectives, aiding the model in recognizing a broader potential dist ribution and fostering its causal perceptual capabilities (refer as Update step) . Our concept aligns seamlessly with the contemporary backdoor adjustment mechan ism in causality theory. Extensive experiments on six real-world ST benchmarks s howcase that models can gain outcomes upon the integration of the NuwaDynamics c oncept. NuwaDynamics also can significantly benefit a wide range of changeable S T tasks like extreme weather and long temporal step super-resolution predictions .
****
Zhenan Fan,Huang Fang,Xinglu Wang,Zirui Zhou,Jian Pei,Michael Friedlander,Yong Z hang
Fair and Efficient Contribution Valuation for Vertical Federated Learning
Federated learning is an emerging technology for training machine learning model s across decentralized data sources without sharing data. Vertical federated lea

rning, also known as feature-based federated learning, applies to scenarios where data sources have the same sample IDs but different feature sets. To ensure fairness among data owners, it is critical to objectively assess the contributions from different data sources and compensate the corresponding data owners accordingly. The Shapley value is a provably fair contribution valuation metric originating from cooperative game theory. However, its straight-forward computation requires extensively retraining a model on each potential combination of data sources, leading to prohibitively high communication and computation overheads due to multiple rounds of federated learning. To tackle this challenge, we propose a contribution valuation metric called vertical federated Shapley value (VerFedSV) based on the classic Shapley value. We show that VerFedSV not only satisfies many desirable properties of fairness but is also efficient to compute. Moreover, VerFedSV can be adapted to both synchronous and asynchronous vertical federated learning algorithms. Both theoretical analysis and extensive experimental results demonstrate the fairness, efficiency, adaptability, and effectiveness of VerFedSV.
****
Mohamed Elsayed,A. Rupam Mahmood
Addressing Loss of Plasticity and Catastrophic Forgetting in Continual Learning
Deep representation learning methods struggle with continual learning, suffering from both catastrophic forgetting of useful units and loss of plasticity, often due to rigid and unuseful units. While many methods address these two issues separately, only a few currently deal with both simultaneously. In this paper, we introduce Utility-based Perturbed Gradient Descent (UPGD) as a novel approach for the continual learning of representations. UPGD combines gradient updates with perturbations, where it applies smaller modifications to more useful units, protecting them from forgetting, and larger modifications to less useful units, rejuvenating their plasticity. We use a challenging streaming learning setup where continual learning problems have hundreds of non-stationarities and unknown task boundaries. We show that many existing methods suffer from at least one of the issues, predominantly manifested by their decreasing accuracy over tasks. On the other hand, UPGD continues to improve performance and surpasses or is competitive with all methods in all problems. Finally, in extended reinforcement learning experiments with PPO, we show that while Adam exhibits a performance drop after initial learning, UPGD avoids it by addressing both continual learning issues.
****
Jiaming Liu,Senqiao Yang,Peidong Jia,Renrui Zhang,Ming Lu,Yandong Guo,Wei Xue,Shanghang Zhang
ViDA: Homeostatic Visual Domain Adapter for Continual Test Time Adaptation
Since real-world machine systems are running in non-stationary environments, Continual Test-Time Adaptation (CTTA) task is proposed to adapt the pre-trained model to continually changing target domains. Recently, existing methods mainly focus on model-based adaptation, which aims to leverage a self-training manner to extract the target domain knowledge. However, pseudo labels can be noisy and the updated model parameters are unreliable under dynamic data distributions, leading to error accumulation and catastrophic forgetting in the continual adaptation process. To tackle these challenges and maintain the model plasticity, we design a Visual Domain Adapter (ViDA) for CTTA, explicitly handling both domain-specific and domain-shared knowledge. Specifically, we first comprehensively explore the different domain representations of the adapters with trainable high-rank or low-rank embedding spaces. Then we inject ViDAs into the pre-trained model, which leverages high-rank and low-rank features to adapt the current domain distribution and maintain the continual domain-shared knowledge, respectively. To exploit the low-rank and high-rank ViDAs more effectively, we further propose a Homeostatic Knowledge Allotment (HKA) strategy, which adaptively combines different knowledge from each ViDA. Extensive experiments conducted on four widely used benchmarks demonstrate that our proposed method achieves state-of-the-art performance in both classification and segmentation CTTA tasks. Note that, our method can be regarded as a novel transfer paradigm for large-scale models, delivering promising results in adaptation to continually changing distributions.

****
Fredrik Carlsson,Johan Broberg,Erik Hillbom,Magnus Sahlgren,Joakim Nivre

Branch-GAN: Improving Text Generation with (not so) Large Language Models

The current advancements in open domain text generation have been spearheaded by Transformer-based large language models. Leveraging efficient parallelization and vast training datasets, these models achieve unparalleled text generation capabilities. Even so, current models are known to suffer from deficiencies such as repetitive texts, looping issues, and lack of robustness. While adversarial training through generative adversarial networks (GAN) is a proposed solution, earlier research in this direction has predominantly focused on older architectures, or narrow tasks. As a result, this approach is not yet compatible with modern language models for open-ended text generation, leading to diminished interest within the broader research community. We propose a computationally efficient GAN approach for sequential data that utilizes the parallelization capabilities of Transformer models. Our method revolves around generating multiple branching sequences from each training sample, while also incorporating the typical next-step prediction loss on the original data. In this way, we achieve a dense reward and loss signal for both the generator and the discriminator, resulting in a stable training dynamic. We apply our training method to pre-trained language models, using data from their original training set but less than 0.01% of the available data. A comprehensive human evaluation shows that our method significantly improves the quality of texts generated by the model while avoiding the previously reported sparsity problems of GAN approaches. Even our smaller models outperform larger original baseline models with more than 16 times the number of parameters. Finally, we corroborate previous claims that perplexity on held-out data is not a sufficient metric for measuring the quality of generated texts.

****
Yuxiang Lai,Yi Zhou,Xinghong Liu,Tao Zhou

Memory-Assisted Sub-Prototype Mining for Universal Domain Adaptation

Universal domain adaptation aims to align the classes and reduce the feature gap between the same category of the source and target domains. The target private category is set as the unknown class during the adaptation process, as it is not included in the source domain. However, most existing methods overlook the intra-class structure within a category, especially in cases where there exists significant concept shift between the samples belonging to the same category. When samples with large concept shift are forced to be pushed together, it may negatively affect the adaptation performance. Moreover, from the interpretability aspect, it is unreasonable to align visual features with significant differences, such as fighter jets and civil aircraft, into the same category. Unfortunately, due to such semantic ambiguity and annotation cost, categories are not always classified in detail, making it difficult for the model to perform precise adaptation. To address these issues, we propose a novel Memory-Assisted Sub-Prototype Mining (MemSPM) method that can learn the differences between samples belonging to the same category and mine sub-classes when there exists significant concept shift between them. By doing so, our model learns a more reasonable feature space that enhances the transferability and reflects the inherent differences among samples annotated as the same category. We evaluate the effectiveness of our MemSPM method over multiple scenarios, including UniDA, OSDA, and PDA. Our method achieves state-of-the-art performance on four benchmarks in most cases.

****
Sherry Yang,Yilun Du,Seyed Kamyar Seyed Ghasemipour,Jonathan Tompson,Leslie Pack Kaelbling,Dale Schuurmans,Pieter Abbeel

Learning Interactive Real-World Simulators

Generative models trained on internet data have revolutionized how text, image, and video content can be created. Perhaps the next milestone for generative models is to simulate realistic experience in response to actions taken by humans, robots, and other interactive agents. Applications of a real-world simulator range from controllable content creation in games and movies, to training embodied agents purely in simulation that can be directly deployed in the real world. We explore the possibility of learning a universal simulator (UniSim) of real-world

interaction through generative modeling. We first make the important observation
 that natural datasets available for learning a real-world simulator are often r
ich along different axes (e.g., abundant objects in image data, densely sampled
actions in robotics data, and diverse movements in navigation data). With carefu
l orchestration of diverse datasets, each providing a different aspect of the ov
erall experience, UniSim can emulate how humans and agents interact with the wor
ld by simulating the visual outcome of both high-level instructions such as “ope
n the drawer” and low-level controls such as “move by x,y” from otherwise static
 scenes and objects. There are numerous use cases for such a real-world simulato
r. As an example, we use UniSim to train both high-level vision-language planner
s and low-level reinforcement learning policies, each of which exhibit zero-shot
 real-world transfer after training purely in a learned real-world simulator. We
 also show that other types of intelligence such as video captioning models can
benefit from training with simulated experience in UniSim, opening up even wider
 applications.
****
Ruiyuan Gao,Kai Chen,Enze Xie,Lanqing HONG,Zhenguo Li,Dit-Yan Yeung,Qiang Xu
MagicDrive: Street View Generation with Diverse 3D Geometry Control
Recent advancements in diffusion models have significantly enhanced the data syn
thesis with 2D control. Yet, precise 3D control in street view generation, cruci
al for 3D perception tasks, remains elusive. Specifically, utilizing Bird's-Eye
View (BEV) as the primary condition often leads to challenges in geometry contro
l (e.g., height), affecting the representation of object shapes, occlusion patte
rns, and road surface elevations, all of which are essential to perception data
synthesis, especially for 3D object detection tasks. In this paper, we introduce
 MagicDrive, a novel street view generation framework, offering diverse 3D geome
try controls including camera poses, road maps, and 3D bounding boxes, together
with textual descriptions, achieved through tailored encoding strategies. Beside
s, our design incorporates a cross-view attention module, ensuring consistency a
cross multiple camera views. With MagicDrive, we achieve high-fidelity street-vi
ew image & video synthesis that captures nuanced 3D geometry and various scene d
escriptions, enhancing tasks like BEV segmentation and 3D object detection. Proj
ect Website: https://flymin.github.io/magicdrive
****
Junyan Cheng,Peter Chin
SocioDojo: Building Lifelong Analytical Agents with Real-world Text and Time Ser
ies
We introduce SocioDojo, an open-ended lifelong learning environment for developi
ng ready-to-deploy autonomous agents capable of performing human-like analysis a
nd decision-making on societal topics such as economics, finance, politics, and
culture. It consists of (1) information sources from news, social media, reports
, etc., (2) a knowledge base built from books, journals, and encyclopedias, plus
 a toolbox of Internet and knowledge graph search interfaces, (3) 30K high-quali
ty time series in finance, economy, society, and polls, which support a novel ta
sk called "hyperportfolio", that can reliably and scalably evaluate societal ana
lysis and decision-making power of agents, inspired by portfolio optimization wi
th time series as assets to "invest". We also propose a novel Analyst-Assistant-
Actuator architecture for the hyperportfolio task, and a Hypothesis & Proof prom
pting for producing in-depth analyses on input news, articles, etc. to assist de
cision-making. We perform experiments and ablation studies to explore the factor
s that impact performance. The results show that our proposed method achieves im
provements of 32.4% and 30.4% compared to the state-of-the-art method in the two
 experimental settings.
****
Sina Baharlouei,Shivam Patel,Meisam Razaviyayn
f-FERM: A Scalable Framework for Robust Fair Empirical Risk Minimization
Training and deploying machine learning models that meet fairness criteria for p
rotected groups are fundamental in modern artificial intelligence.
While numerous constraints and regularization terms have been proposed in the li
terature to promote fairness in machine learning tasks, most of these approaches

are not amenable to stochastic optimization due to the complex and nonlinear st ructure of constraints and regularizers. Here, the term ``stochastic'' refers to the ability of the algorithm to work with small mini-batches of data. Motivated by the limitation of existing literature, this paper presents a unified stochas tic optimization framework for fair empirical risk minimization based on f-div ergence measures (f-FERM). The proposed stochastic algorithm enjoys theoretica l convergence guarantees. In addition, our experiments demonstrate the superiori ty of fairness-accuracy tradeoffs offered by f-FERM for almost all batch sizes (ranging from full-batch to batch size of one). Moreover, we show that our fram ework can be extended to the case where there is a distribution shift from train ing to the test data.
Our extension is based on a distributionally robust optimization reformulation o f f-FERM objective under ℓ_p norms as uncertainty sets. Again, in this di stributionally robust setting, f-FERM not only enjoys theoretical convergence guarantees but also outperforms other baselines in the literature in the tasks i nvolving distribution shifts.
An efficient stochastic implementation of f-FERM is publicly available.
****
Yanbo Wang,Jian Liang,Ran He
Towards Eliminating Hard Label Constraints in Gradient Inversion Attacks
Gradient inversion attacks aim to reconstruct local training data from intermedi ate gradients exposed in the federated learning framework. Despite successful at tacks, all previous methods, starting from reconstructing a single data point an d then relaxing the single-image limit to batch level, are only tested under har d label constraints. Even for single-image reconstruction, we still lack an anal ysis-based algorithm to recover augmented soft labels. In this work, we change t he focus from enlarging batchsize to investigating the hard label constraints, c onsidering a more realistic circumstance where label smoothing and mixup techniq ues are used in the training process. In particular, we are the first to initiat e a novel algorithm to simultaneously recover the ground-truth augmented label a nd the input feature of the last fully-connected layer from single-input gradien ts, and provide a necessary condition for any analytical-based label recovery me thods. Extensive experiments testify to the label recovery accuracy, as well as the benefits to the following image reconstruction. We believe soft labels in cl assification tasks are worth further attention in gradient inversion attacks.
****
Tom Yan,Chicheng Zhang
The Human-AI Substitution game: active learning from a strategic labeler
The standard active learning setting assumes a willing labeler, who provides lab els on informative examples to speed up learning. However, if the labeler wishes to be compensated for as many labels as possible before learning finishes, the labeler may benefit from actually slowing down learning. This incentive arises f or instance if the labeler is to be replaced by the ML model, once it is learned . In this paper, we initiate the study of learning from a strategic labeler, who selectively abstains from labeling to slow down learning. We first prove that s trategic abstention can prolong learning, and propose novel complexity measures to analyze the query cost of the learning game. Next, we develop a near-optimal deterministic algorithm, prove its robustness to strategic labeling, and contras t it with other active learning algorithms. We also provide extensions that enco mpass other learning setups/goals. Finally, we characterize the query cost of mu lti-task active learning, with and without abstention. Our first exploration of strategic labeling aims to add to our theoretical understanding of the imitative nature of ML in human-AI interaction.
****
Tinghao Xie,Xiangyu Qi,Ping He,Yiming Li,Jiachen T. Wang,Prateek Mittal
BaDExpert: Extracting Backdoor Functionality for Accurate Backdoor Input Detecti on
We present a novel defense, against backdoor attacks on Deep Neural Networks (DN Ns), wherein adversaries covertly implant malicious behaviors (backdoors) into D NNs. Our defense falls within the category of post-development defenses that ope

rate independently of how the model was generated. The proposed defense is built upon a novel reverse engineering approach that can directly extract **backdoor functionality** of a given backdoored model to a *backdoor expert* model. The approach is straightforward --- finetuning the backdoored model over a small set of intentionally mislabeled clean samples, such that it unlearns the normal functionality while still preserving the backdoor functionality, and thus resulting in a model~(dubbed a backdoor expert model) that can only recognize backdoor inputs. Based on the extracted backdoor expert model, we show the feasibility of devising highly accurate backdoor input detectors that filter out the backdoor inputs during model inference. Further augmented by an ensemble strategy with a finetuned auxiliary model, our defense, **BaDExpert** (**Ba**ckdoor Input **D**etection with Backdoor **Expert**), effectively mitigates 17 SOTA backdoor attacks while minimally impacting clean utility. The effectiveness of BaDExpert has been verified on multiple datasets (CIFAR10, GTSRB and ImageNet) across various model architectures (ResNet, VGG, MobileNetV2 and Vision Transformer). Our code is integrated into our research toolbox: https://github.com/vtu81/backdoor-toolbox.
****
Seungcheol Park,Hojun Choi,U Kang
Accurate Retraining-free Pruning for Pretrained Encoder-based Language Models
Given a pretrained encoder-based language model, how can we accurately compress it without retraining? Retraining-free structured pruning algorithms are crucial in pretrained language model compression due to their significantly reduced pruning cost and capability to prune large language models. However, existing retraining-free algorithms encounter severe accuracy degradation, as they fail to handle pruning errors, especially at high compression rates. In this paper, we propose KPrune (Knowledge-preserving pruning), an accurate retraining-free structured pruning algorithm for pretrained encoder-based language models.
KPrune focuses on preserving the useful knowledge of the pretrained model to minimize pruning errors through a carefully designed iterative pruning process composed of knowledge measurement, knowledge-preserving mask search, and knowledge-preserving weight-tuning. As a result, KPrune shows significant accuracy improvements up to 58.02%p higher F1 score compared to existing retraining-free pruning algorithms under a high compression rate of 80% on the SQuAD benchmark without any retraining process.
****
Wei Mao,Richard Hartley,Mathieu Salzmann,miaomiao Liu
Neural SDF Flow for 3D Reconstruction of Dynamic Scenes
In this paper, we tackle the problem of 3D reconstruction of dynamic scenes from multi-view videos. Previous dynamic scene reconstruction works either attempt to model the motion of 3D points in space, which constrains them to handle a single articulated object or require depth maps as input. By contrast, we propose to directly estimate the change of Signed Distance Function (SDF), namely SDF flow, of the dynamic scene. We show that the SDF flow captures the evolution of the scene surface. We further derive the mathematical relation between the SDF flow and the scene flow, which allows us to calculate the scene flow from the SDF flow analytically by solving linear equations. Our experiments on real-world multi-view video datasets show that our reconstructions are better than those of the state-of-the-art methods. Our code is available at https://github.com/wei-mao-2019/SDFFlow.git.
****
Olga Fourkioti,Matt De Vries,Chris Bakal
CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images
The visual examination of tissue biopsy sections is fundamental for cancer diagnosis, with pathologists analyzing sections at multiple magnifications to discern tumor cells and their subtypes. However, existing attention-based multiple instance learning (MIL) models, used for analyzing Whole Slide Images (WSIs) in cancer diagnostics, often overlook the contextual information of tumor and neighboring tiles, leading to misclassifications. To address this, we propose the Context

-Aware Multiple Instance Learning (CAMIL) architecture. CAMIL incorporates neighbor-constrained attention to consider dependencies among tiles within a WSI and integrates contextual constraints as prior knowledge into the MIL model. We evaluated CAMIL on subtyping non-small cell lung cancer (TCGA-NSCLC) and detecting lymph node (CAMELYON16) metastasis, achieving test AUCs of 0.959\% and 0.975\%, respectively, outperforming other state-of-the-art methods. Additionally, CAMIL enhances model interpretability by identifying regions of high diagnostic value
****
Andrew Kirjner,Jason Yim,Raman Samusevich,Shahar Bracha,Tommi S. Jaakkola,Regina Barzilay,Ila R Fiete
Improving protein optimization with smoothed fitness landscapes
The ability to engineer novel proteins with higher fitness for a desired property would be revolutionary for biotechnology and medicine. Modeling the combinatorially large space of sequences is infeasible; prior methods often constrain optimization to a small mutational radius, but this drastically limits the design space. Instead of heuristics, we propose smoothing the fitness landscape to facilitate protein optimization. First, we formulate protein fitness as a graph signal then use Tikunov regularization to smooth the fitness landscape. We find optimizing in this smoothed landscape leads to improved performance across multiple methods in the GFP and AAV benchmarks. Second, we achieve state-of-the-art results utilizing discrete energy-based models and MCMC in the smoothed landscape. Our method, called Gibbs sampling with Graph-based Smoothing (GGS), demonstrates a unique ability to achieve 2.5 fold fitness improvement (with in-silico evaluation) over its training set. GGS demonstrates potential to optimize proteins in the limited data regime. Code: https://github.com/kirjner/GGS
****
Shrinivas Ramasubramanian,Harsh Rangwani,Sho Takemori,Kunal Samanta,Yuhei Umeda,Venkatesh Babu Radhakrishnan
Selective Mixup Fine-Tuning for Optimizing Non-Decomposable Objectives
The rise in internet usage has led to the generation of massive amounts of data, resulting in the adoption of various supervised and semi-supervised machine learning algorithms, which can effectively utilize the colossal amount of data to train models. However, before deploying these models in the real world, these must be strictly evaluated on performance measures like worst-case recall and satisfy constraints such as fairness. We find that current state-of-the-art empirical techniques offer sub-optimal performance on these practical, non-decomposable performance objectives. On the other hand, the theoretical techniques necessitate training a new model from scratch for each performance objective. To bridge the gap, we propose SelMix, a selective mixup-based inexpensive fine-tuning technique for pre-trained models, to optimize for the desired objective. The core idea of our framework is to determine a sampling distribution to perform a mixup of features between samples from particular classes such that it optimizes the given objective. We comprehensively evaluate our technique against the existing empirical and theoretically principled methods on standard benchmark datasets for imbalanced classification. We find that proposed SelMix fine-tuning significantly improves the performance for various practical non-decomposable objectives across benchmarks.
****
Dominik Schmidt,Minqi Jiang
Learning to Act without Actions
Pre-training large models on vast amounts of web data has proven to be an effective approach for obtaining powerful, general models in domains such as language and vision. However, this paradigm has not yet taken hold in reinforcement learning. This is because videos, the most abundant form of embodied behavioral data on the web, lack the action labels required by existing methods for imitating behavior from demonstrations. We introduce *Latent Action Policies* (LAPO), a method for recovering latent action information—and obtaining latent-action policies, world models, and inverse dynamics models—purely from videos. LAPO is the first method able to recover the structure of the true action space purely from observed dynamics, even in challenging procedurally-generated environments. Further,

LAPO's latent-action policies can be rapidly turned into regular, expert-level policies, either offline using a small action-labeled dataset, or online via rewards. LAPO is the first step towards pre-training powerful, generalist policies and world models on the vast amounts of videos readily available on the web.
****
Yunyang Li,Yusong Wang,Lin Huang,Han Yang,Xinran Wei,Jia Zhang,Tong Wang,Zun Wang,Bin Shao,Tie-Yan Liu
Long-Short-Range Message-Passing: A Physics-Informed Framework to Capture Non-Local Interaction for Scalable Molecular Dynamics Simulation
Computational simulation of chemical and biological systems using *ab initio* molecular dynamics has been a challenge over decades. Researchers have attempted to address the problem with machine learning and fragmentation-based methods. However, the two approaches fail to give a satisfactory description of long-range and many-body interactions, respectively. Inspired by fragmentation-based methods, we propose the Long-Short-Range Message-Passing (LSR-MP) framework as a generalization of the existing equivariant graph neural networks (EGNNs) with the intent to incorporate long-range interactions efficiently and effectively. We apply the LSR-MP framework to the recently proposed ViSNet and demonstrate the state-of-the-art results with up to 40% MAE reduction for molecules in MD22 and Chignolin datasets. Consistent improvements to various EGNNs will also be discussed to illustrate the general applicability and robustness of our LSR-MP framework. The code for our experiments and trained model weights could be found at https://github.com/liyy2/LSR-MP.
****
Etash Kumar Guha,Shlok Natarajan,Thomas Möllenhoff,Mohammad Emtiyaz Khan,Eugene Ndiaye
Conformal Prediction via Regression-as-Classification
Conformal prediction (CP) for regression can be challenging, especially when the output distribution is heteroscedastic, multimodal, or skewed. Some of the issues can be addressed by estimating a distribution over the output, but in reality, such approaches can be sensitive to estimation error and yield unstable intervals. Here, we circumvent the challenges by converting regression to a classification problem and then use CP for classification to obtain CP sets for regression. To preserve the ordering of the continuous-output space, we design a new loss function and present necessary modifications to the CP classification techniques. Empirical results on many benchmarks show that this simple approach gives surprisingly good results on many practical problems.
****
Sewon Min,Suchin Gururangan,Eric Wallace,Weijia Shi,Hannaneh Hajishirzi,Noah A. Smith,Luke Zettlemoyer
SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
The legality of training language models (LMs) on copyrighted or otherwise restricted data is under intense debate. However, as we show, model performance significantly degrades if trained only on low-risk text (e.g., out-of-copyright books or government documents), due to its limited size and domain coverage. We present SILO, a new language model that manages this risk-performance tradeoff during inference. SILO is built by (1) training a parametric LM on the Open License Corpus (OLC), a new corpus we curate with 228B tokens of public domain and permissively licensed text and (2) augmenting it with a more general and easily modifiable nonparametric datastore (e.g., containing copyrighted books or news) that is only queried during inference. The datastore allows use of high-risk data without training on it, supports sentence-level data attribution, and enables data producers to opt out from the model by removing content from the store. These capabilities can foster compliance with data-use regulations such as the fair use doctrine in the United States and the GDPR in the European Union. Our experiments show that the parametric LM struggles on its own with domains not covered by OLC. However, access to the datastore greatly improves out of domain performance, closing 90% of the performance gap with an LM trained on the Pile, a more diverse corpus with mostly high-risk text. We also analyze which nonparametric approach works best, where the remaining errors lie, and how performance scales with dat

astore size. Our results suggest that it is possible to build high quality langu age models while mitigating legal risk.
****
Eduardo Dadalto Câmara Gomes,Marco Romanelli,Georg Pichler,Pablo Piantanida
A Data-Driven Measure of Relative Uncertainty for Misclassification Detection
Misclassification detection is an important problem in machine learning, as it a llows for the identification of instances where the model's predictions are unre liable. However, conventional uncertainty measures such as Shannon entropy do no t provide an effective way to infer the real uncertainty associated with the mod el's predictions. In this paper, we introduce a novel data-driven measure of unc ertainty relative to an observer for misclassification detection. By learning pa tterns in the distribution of soft-predictions, our uncertainty measure can iden tify misclassified samples based on the predicted class probabilities. Interest ingly, according to the proposed measure, soft-predictions corresponding to misc lassified instances can carry a large amount of uncertainty, even though they ma y have low Shannon entropy. We demonstrate empirical improvements over multiple image classification tasks, outperforming state-of-the-art misclassification det ection methods.
****
Rembert Daems,Manfred Opper,Guillaume Crevecoeur,Tolga Birdal
Variational Inference for SDEs Driven by Fractional Noise
We present a novel variational framework for performing inference in (neural) st ochastic differential equations (SDEs) driven by Markov-approximate fractional B rownian motion (fBM). SDEs offer a versatile tool for modeling real-world contin uous-time dynamic systems with inherent noise and randomness. Combining SDEs wit h the powerful inference capabilities of variational methods, enables the learni ng of representative distributions through stochastic gradient descent. However, conventional SDEs typically assume the underlying noise to follow a Brownian m otion (BM), which hinders their ability to capture long-term dependencies. In co ntrast, fractional Brownian motion (fBM) extends BM to encompass non-Markovian d ynamics, but existing methods for inferring fBM parameters are either computatio nally demanding or statistically inefficient.

In this paper, building upon the Markov approximation of fBM, we derive the evid ence lower bound essential for efficient variational inference of posterior path measures, drawing from the well-established field of stochastic analysis. Addit ionally, we provide a closed-form expression for optimal approximation coefficie nts and propose to use neural networks to learn the drift, diffusion and control terms within our variational posterior, leading to the variational training of neural-SDEs. In this framework, we also optimize the Hurst index, governing the nature of our fractional noise. Beyond validation on synthetic data, we contribu te a novel architecture for variational latent video prediction,—an approach tha t, to the best of our knowledge, enables the first variational neural-SDE applic ation to video perception.
****
Yihao Xue,Siddharth Joshi,Dang Nguyen,Baharan Mirzasoleiman
Understanding the Robustness of Multi-modal Contrastive Learning to Distribution Shift
Recently, multimodal contrastive learning (MMCL) approaches, such as CLIP, have achieved a remarkable success in learning representations that are robust agains t distribution shift and generalize to new domains. Despite the empirical succes s, the mechanism behind learning such generalizable representations is not under stood. In this work, we rigorously analyze this problem and
uncover two mechanisms behind MMCL's robustness: \emph{intra-class contrasting}, which allows the model to learn features with a high variance, and \emph{inter-class feature sharing}, where annotated details in one class help learning other classes better. Both mechanisms prevent spurious features that are over-represe nted in the training data to overshadow the generalizable core features. This yi elds superior zero-shot classification accuracy under distribution shift. Furthe rmore, we theoretically demonstrate the benefits of using rich captions on robus

tness and explore the effect of annotating different types of details in the captions. We validate our theoretical findings through experiments, including a well-designed synthetic experiment and an experiment involving training CLIP models on MSCOCO/Conceptual Captions and evaluating them on shifted ImageNets.
****
Alain Rakotomamonjy,Kimia Nadjahi,Liva Ralaivola
Federated Wasserstein Distance
We introduce a principled way of computing the Wasserstein distance between two distributions in a federated manner.
Namely, we show how to estimate the Wasserstein distance between two samples stored and
kept on different devices/clients whilst a central entity/server orchestrates the computations
(again, without having access to the samples). To achieve this feat, we take advantage of the geometric
properties of the Wasserstein distance -- in particular, the triangle inequality --
and that of the associated {\em geodesics}: our algorithm, FedWad (for Federated Wasserstein Distance), iteratively approximates
the Wasserstein distance by manipulating and exchanging distributions from the space of geodesics in lieu of the input samples.
In addition to establishing the convergence properties of FedWad,
we provide empirical results on federated coresets and federate
optimal transport dataset distance, that we respectively exploit for
building a novel federated model and for boosting performance of popular federated learning algorithms.
****
Yongchao Zhou,Kaifeng Lyu,Ankit Singh Rawat,Aditya Krishna Menon,Afshin Rostamizadeh,Sanjiv Kumar,Jean-François Kagy,Rishabh Agarwal
DistillSpec: Improving Speculative Decoding via Knowledge Distillation
Speculative decoding~(SD) accelerates large language model inference by employing a faster {\em draft} model for generating multiple tokens, which are then verified in parallel by the larger {\em target} model, resulting in the text generated according to the target model distribution. However, identifying a compact draft model that is well-aligned with the target model is challenging. To tackle this issue, we propose {\em DistillSpec} that uses knowledge distillation to better align the draft model with the target model, before applying SD. DistillSpec makes two key design choices, which we demonstrate via systematic study to be crucial to improve the draft and target alignment: utilizing \emph{on-policy} data generation from the draft model, and \emph{tailoring the divergence function} to the task and decoding strategy. Notably, DistillSpec yields impressive $10 - 45\%$ speedups over standard SD on a range of standard benchmarks, using both greedy and non-greedy sampling. Furthermore, we combine DistillSpec with lossy SD to achieve fine-grained control over the latency vs. task performance trade-off. Finally, in practical scenarios with models of varying sizes, first using distillation to boost the performance of the target model and then applying DistillSpec to train a well-aligned draft model can reduce decoding latency by $6 - 10\times$ with minimal performance drop, compared to standard decoding without distillation.
****
Fei Kong,Jinhao Duan,RuiPeng Ma,Heng Tao Shen,Xiaoshuang Shi,Xiaofeng Zhu,Kaidi Xu
An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization
Recently, diffusion models have achieved remarkable success in generating tasks, including image and audio generation. However, like other generative models, diffusion models are prone to privacy issues. In this paper, we propose an efficient query-based membership inference attack (MIA), namely Proximal Initialization Attack (PIA), which utilizes groundtruth trajectory obtained by ϵ initialized in $t=0$ and predicted point to infer memberships. Experimental results

indicate that the proposed method can achieve competitive performance with only two queries that achieve at least 6$\times$ efficiency than the previous SOTA baseline on both discrete-time and continuous-time diffusion models. Moreover, previous works on the privacy of diffusion models have focused on vision tasks without considering audio tasks. Therefore, we also explore the robustness of diffusion models to MIA in the text-to-speech (TTS) task, which is an audio generation task. To the best of our knowledge, this work is the first to study the robustness of diffusion models to MIA in the TTS task. Experimental results indicate that models with mel-spectrogram (image-like) output are vulnerable to MIA, while models with audio output are relatively robust to MIA. Code is available at https://github.com/kong13661/PIA.
****
Songning Lai,Lijie Hu,Junxiao Wang,Laure Berti-Equille,Di Wang
Faithful Vision-Language Interpretation via Concept Bottleneck Models
The demand for transparency in healthcare and finance has led to interpretable machine learning (IML) models, notably the concept bottleneck models (CBMs), valued for their potential in performance and insights into deep neural networks. However, CBM's reliance on manually annotated data poses challenges. Label-free CBMs have emerged to address this, but they remain unstable, affecting their faithfulness as explanatory tools. To address this issue of inherent instability, we introduce a formal definition for an alternative concept called the Faithful Vision-Language Concept (FVLC) model. We present a methodology for constructing an FVLC that satisfies four critical properties. Our extensive experiments on four benchmark datasets using Label-free CBM model architectures demonstrate that our FVLC outperforms other baselines regarding stability against input and concept set perturbations. Our approach incurs minimal accuracy degradation compared to the vanilla CBM, making it a promising solution for reliable and faithful model interpretation.
****
Yihuan Mao,Chengjie Wu,Xi Chen,Hao Hu,Ji Jiang,Tianze Zhou,Tangjie Lv,Changjie Fan,Zhipeng Hu,Yi Wu,Yujing Hu,Chongjie Zhang
Stylized Offline Reinforcement Learning: Extracting Diverse High-Quality Behaviors from Heterogeneous Datasets
Previous literature on policy diversity in reinforcement learning (RL) either focuses on the online setting or ignores the policy performance. In contrast, offline RL, which aims to learn high-quality policies from batched data, has yet to fully leverage the intrinsic diversity of the offline dataset. Addressing this dichotomy and aiming to balance quality and diversity poses a significant challenge to extant methodologies. This paper introduces a novel approach, termed Stylized Offline RL (SORL), which is designed to extract high-performing, stylistically diverse policies from a dataset characterized by distinct behavioral patterns. Drawing inspiration from the venerable Expectation-Maximization (EM) algorithm, SORL innovatively alternates between policy learning and trajectory clustering, a mechanism that promotes policy diversification. To further augment policy performance, we introduce advantage-weighted style learning into the SORL framework. Experimental evaluations across multiple environments demonstrate the significant superiority of SORL over previous methods in extracting high-quality policies with diverse behaviors. A case in point is that SORL successfully learns strong policies with markedly distinct playing patterns from a real-world human dataset of a popular basketball video game "Dunk City Dynasty."
****
Priyank Jaini,Kevin Clark,Robert Geirhos
Intriguing Properties of Generative Classifiers
What is the best paradigm to recognize objects---discriminative inference (fast but potentially prone to shortcut learning) or using a generative model (slow but potentially more robust)? We build on recent advances in generative modeling that turn text-to-image models into classifiers. This allows us to study their behavior and to compare them against discriminative models and human psychophysical data.
We report four intriguing emergent properties of generative classifiers: they sh

ow a record-breaking human-like shape bias (99% for Imagen), near human-level out-of-distribution accuracy, state-of-the-art alignment with human classification errors, and they understand certain perceptual illusions. Our results indicate that while the current dominant paradigm for modeling human object recognition is discriminative inference, zero-shot generative models approximate human object recognition data surprisingly well.
****
Jiawei Ge,Shange Tang,Jianqing Fan,Chi Jin
On the Provable Advantage of Unsupervised Pretraining
Unsupervised pretraining, which learns a useful representation using a large amount of unlabeled data to facilitate the learning of downstream tasks, is a critical component of modern large-scale machine learning systems. Despite its tremendous empirical success, the rigorous theoretical understanding of why unsupervised pretraining generally helps remains rather limited---most existing results are restricted to particular methods or approaches for unsupervised pretraining with specialized structural assumptions. This paper studies a generic framework, where the unsupervised representation learning task is specified by an abstract class of latent variable models Φ and the downstream task is specified by a class of prediction functions Ψ. We consider a natural approach of using Maximum Likelihood Estimation (MLE) for unsupervised pretraining and Empirical Risk Minimization (ERM) for learning downstream tasks. We prove that, under a mild ``informative'' condition, our algorithm achieves an excess risk of $\\tilde{\\mathcal{O}}(\sqrt{\mathcal{C}\_\Phi/m} + \sqrt{\mathcal{C}\_\Psi/n})$ for downstream tasks, where $\mathcal{C}\_\Phi, \mathcal{C}\_\Psi$ are complexity measures of function classes Φ, Ψ, and m, n are the number of unlabeled and labeled data respectively. Comparing to the baseline of $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}\_{\Phi \circ \Psi}/n})$ achieved by performing supervised learning using only the labeled data, our result rigorously shows the benefit of unsupervised pretraining when $m \gg n$ and $\mathcal{C}\_{\Phi\circ \Psi} > \mathcal{C}\_\Psi$. This paper further shows that our generic framework covers a wide range of approaches for unsupervised pretraining, including factor models, Gaussian mixture models, and contrastive learning.
****
Sachin Kumar,Chan Young Park,Yulia Tsvetkov
Gen-Z: Generative Zero-Shot Text Classification with Contextualized Label Descriptions
Language model (LM) prompting—a popular paradigm for solving NLP tasks—has been shown to be susceptible to miscalibration and brittleness to slight prompt variations, caused by its discriminative prompting approach, i.e., predicting the label given the input. To address these issues, we propose Gen-Z—a generative prompting framework for zero-shot text classification. GEN-Z is generative, as it measures the LM likelihood of input text, conditioned on natural language descriptions of labels. The framework is multivariate, as label descriptions allow us to seamlessly integrate additional contextual information about the labels to improve task performance. On various standard classification benchmarks, with six open-source LM families, we show that zero-shot classification with simple contextualization of the data source of the evaluation set consistently outperforms both zero-shot and few-shot baselines while improving robustness to prompt variations. Further, our approach enables personalizing classification in a zero-shot manner by incorporating author, subject, or reader information in the label descriptions.
****
Zixian Huang,Gengyang Xiao,Yu Gu,Gong Cheng
A Branching Decoder for Set Generation
Generating a set of text is a common challenge for many NLP applications, for example, automatically providing multiple keyphrases for a document to facilitate user reading. Existing generative models use a sequential decoder that generates a single sequence successively, and the set generation problem is converted to sequence generation via concatenating multiple text into a long text sequence. However, the elements of a set are unordered, which makes this scheme suffer from

biased or conflicting training signals. In this paper, we propose a branching decoder, which can generate a dynamic number of tokens at each time-step and branch multiple generation paths. In particular, paths are generated individually so that no order dependence is required. Moreover, multiple paths can be generated in parallel which greatly reduces the inference time. Experiments on several keyphrase generation datasets demonstrate that the branching decoder is more effective and efficient than the existing sequential decoder.
****
Yury Gorishniy,Ivan Rubachev,Nikolay Kartashev,Daniil Shlenskii,Akim Kotelnikov,Artem Babenko
TabR: Tabular Deep Learning Meets Nearest Neighbors
Deep learning (DL) models for tabular data problems (e.g. classification, regression) are currently receiving increasingly more attention from researchers.
However, despite the recent efforts, the non-DL algorithms based on gradient-boosted decision trees (GBDT) remain a strong go-to solution for these problems.
One of the research directions aimed at improving the position of tabular DL involves designing so-called retrieval-augmented models.
For a target object, such models retrieve other objects (e.g. the nearest neighbors) from the available training data and use their features and labels to make a better prediction.

In this work, we present TabR -- essentially, a feed-forward network with a custom k-Nearest-Neighbors-like component in the middle.
On a set of public benchmarks with datasets up to several million objects, TabR marks a big step forward for tabular DL: it demonstrates the best average performance among tabular DL models, becomes the new state-of-the-art on several datasets, and even outperforms GBDT models on the recently proposed "GBDT-friendly" benchmark (see Figure 1).
Among the important findings and technical details powering TabR, the main ones lie in the attention-like mechanism that is responsible for retrieving the nearest neighbors and extracting valuable signal from them.
In addition to the higher performance, TabR is simple and significantly more efficient compared to prior retrieval-based tabular DL models.
****
Megan Richards,Polina Kirichenko,Diane Bouchacourt,Mark Ibrahim
Does Progress On Object Recognition Benchmarks Improve Generalization on Crowdsourced, Global Data?
For more than a decade, researchers have measured progress in object recognition on the ImageNet dataset along with its associated generalization benchmarks such as ImageNet-A, -C, and -R. Recent advances in foundation models, trained on orders of magnitude more data, have begun to saturate performance on these benchmarks. Despite this progress, even today’s best models are brittle in practice. As a step toward more holistic measurement of model reliability, we propose studying performance on crowdsourced, global datasets, which contain natural distribution shifts seen practically in deployment. We perform a comprehensive empirical study on two crowdsourced, globally representative datasets, evaluating nearly 100 vision models to uncover several concerning empirical trends: first, that progress on crowdsourced, global data has significantly lagged behind standard benchmarks, with advances on ImageNet occurring at $2.5x$ the rate of progress on crowdsourced, global data. Second, we find that progress on standard benchmarks has failed to improve or exacerbated geographic disparities: \textit{geographic disparities between the least performant models and today's best models have more than tripled}. We showcase the promise of using more curated and/or representative training datasets for mitigating these trends, and emphasize curation of web-scale, geographically representative training datasets as a critical open problem for the research community.
****
Wenbo Li,Xin Yu,Kun Zhou,Yibing Song,Zhe Lin
Image Inpainting via Iteratively Decoupled Probabilistic Modeling
Generative adversarial networks (GANs) have made great success in image inpainti

ng yet still have difficulties tackling large missing regions. In contrast, iterative probabilistic algorithms, such as autoregressive and denoising diffusion models, have to be deployed with massive computing resources for decent effect. To achieve high-quality results with low computational cost, we present a novel pixel spread model (PSM) that iteratively employs decoupled probabilistic modeling, combining the optimization efficiency of GANs with the prediction tractability of probabilistic models. As a result, our model selectively spreads informative pixels throughout the image in a few iterations, largely enhancing the completion quality and efficiency. On multiple benchmarks, we achieve new state-of-the-art performance. Our code and models will be publicly available.
****

Zeren Chen,Ziqin Wang,Zhen Wang,Huayang Liu,Zhenfei Yin,Si Liu,Lu Sheng,Wanli Ouyang,Jing Shao

Octavius: Mitigating Task Interference in MLLMs via LoRA-MoE

Recent studies have demonstrated Large Language Models (LLMs) can extend their zero-shot generalization capabilities to multimodal learning through instruction tuning. As more modalities and downstream tasks are introduced, negative conflicts and interference may have a worse impact on performance. While this phenomenon has been overlooked in previous work, we propose a novel and extensible framework, called Octavius, for comprehensive studies and experimentation on multimodal learning with Multimodal Large Language Models (MLLMs). Specifically, to mitigate the interference, we combine the concept of Mixture-of-Experts (MoE) with LoRA and design a multimodal LoRA-MoE decoder for task- and modality-specific learning. To the best of our knowledge, we are one of the pioneering efforts to introduce MoE into MLLMs to address this problem. The experimental results (about 20% improvement) have shown the effectiveness and versatility of our design in various 2D and 3D downstream tasks. Code and corresponding dataset will be available
soon.
****

Ziyao Guo,Kai Wang,George Cazenavette,HUI LI,Kaipeng Zhang,Yang You

Towards Lossless Dataset Distillation via Difficulty-Aligned Trajectory Matching

The ultimate goal of Dataset Distillation is to synthesize a small synthetic dataset such that a model trained on this synthetic set will perform equally well as a model trained on the full, real dataset. Until now, no method of Dataset Distillation has reached this completely lossless goal, in part due to the fact that previous methods only remain effective when the total number of synthetic samples is extremely small. Since only so much information can be contained in such a small number of samples, it seems that to achieve truly loss dataset distillation, we must develop a distillation method that remains effective as the size of the synthetic dataset grows. In this work, we present such an algorithm and elucidate why existing methods fail to generate larger, high-quality synthetic sets. Current state-of-the-art methods rely on trajectory-matching, or optimizing the synthetic data to induce similar long-term training dynamics as the real data. We empirically find that the training stage of the trajectories we choose to match (i.e., early or late) greatly affects the effectiveness of the distilled dataset. Specifically, early trajectories (where the teacher network learns easy patterns) work well for a low-cardinality synthetic set since there are fewer examples wherein to distribute the necessary information. Conversely, late trajectories (where the teacher network learns hard patterns) provide better signals for larger synthetic sets since there are now enough samples to represent the necessary complex patterns. Based on our findings, we propose to align the difficulty of the generated patterns with the size of the synthetic dataset. In doing so, we successfully scale trajectory matching-based methods to larger synthetic datasets, achieving lossless dataset distillation for the very first time. Code and distilled datasets will be released.
****

Shanda Li,Chong You,Guru Guruganesh,Joshua Ainslie,Santiago Ontanon,Manzil Zaheer,Sumit Sanghai,Yiming Yang,Sanjiv Kumar,Srinadh Bhojanapalli

Functional Interpolation for Relative Positions improves Long Context Transforme

rs
Preventing the performance decay of Transformers on inputs longer than those use
d for training has been an important challenge in extending the context length o
f these models. Though the Transformer architecture has fundamentally no limits
on the input sequence lengths it can process, the choice of position encoding us
ed during training can limit the performance of these models on longer inputs. W
e propose a novel functional relative position encoding with progressive interpo
lation, FIRE, to improve Transformer generalization to longer contexts. We theor
etically prove that this can represent some of the popular relative position enc
odings, such as T5's RPE, Alibi, and Kerple. We next empirically show that FIRE
models have better generalization to longer contexts on both zero-shot language
modeling and long text benchmarks.
****
Michael-Andrei Panaitescu-Liess,Yigitcan Kaya,Sicheng Zhu,Furong Huang,Tudor Dum
itras
Like Oil and Water: Group Robustness Methods and Poisoning Defenses Don't Mix
Group robustness has become a major concern in machine learning (ML) as conventi
onal training paradigms were found to produce high error on minority groups. Wit
hout explicit group annotations, proposed solutions rely on heuristics that aim
to identify and then amplify the minority samples during training. In our work,
we first uncover a critical shortcoming of these methods: an inability to distin
guish legitimate minority samples from poison samples in the training set. By am
plifying poison samples as well, group robustness methods inadvertently boost th
e success rate of an adversary---e.g., from 0\% without amplification to over 97
\% with it. Notably, we supplement our empirical evidence with an impossibility
result proving this inability of a standard heuristic under some assumptions. Mo
reover, scrutinizing recent poisoning defenses both in centralized and federated
 learning, we observe that they rely on similar heuristics to identify which sam
ples should be eliminated as poisons. In consequence, minority samples are elimi
nated along with poisons, which damages group robustness---e.g., from 55\% witho
ut the removal of the minority samples to 41\% with it. Finally, as they pursue
opposing goals using similar heuristics, our attempt to alleviate the trade-off
by combining group robustness methods and poisoning defenses falls short. By exp
osing this tension, we also hope to highlight how benchmark-driven ML scholarshi
p can obscure the trade-offs among different metrics with potentially detrimenta
l consequences.
****
Yiqun Yao,Zheng Zhang,Jing Li,Yequan Wang
Masked Structural Growth for 2x Faster Language Model Pre-training
Accelerating large language model pre-training is a critical issue in present re
search. In this paper, we focus on speeding up pre-training by progressively gro
wing from a small Transformer structure to a large one. There are two main resea
rch problems associated with progressive growth: determining the optimal growth
schedule, and designing efficient growth operators. In terms of growth schedule,
 the impact of each single dimension on a schedule’s efficiency is underexplored
 by existing work. Regarding the growth operators, existing methods rely on the
initialization of new weights to inherit knowledge, and achieve only non-strict
function preservation, limiting further improvements on training dynamics. To ad
dress these issues, we propose Masked Structural Growth (MSG), including (i) gro
wth schedules involving all possible dimensions and (ii) strictly function-prese
rving growth operators that is independent of the initialization of new weights.
 Experiments show that MSG is significantly faster than related work: we achieve
 up to 2.2x speedup in pre-training different types of language models while mai
ntaining comparable or better downstream performances. Code is publicly availabl
e at https://github.com/cofe-ai/MSG.
****
Philip Quirke,Fazl Barez
Understanding Addition in Transformers
Understanding the inner workings of machine learning models like Transformers is
 vital for their safe and ethical use. This paper provides a comprehensive analy

sis of a one-layer Transformer model trained to perform n-digit integer addition. Our findings suggests that the model dissects the task into parallel streams dedicated to individual digits, employing varied algorithms tailored to different positions within the digits. Furthermore, we identify a rare scenario characterized by high loss, which we explain. By thoroughly elucidating the model's algorithm, we provide new insights into its functioning. These findings are validated through rigorous testing and mathematical modeling, thereby contributing to the broader fields of model understanding and interpretability. Our approach opens the door for analyzing more complex tasks and multi-layer Transformer models.
****
Yash J. Patel,Akash Kundu,Mateusz Ostaszewski,Xavier Bonet-Monroig,Vedran Dunjko,Onur Danaci
Curriculum reinforcement learning for quantum architecture search under hardware errors
The key challenge in the noisy intermediate-scale quantum era is finding useful circuits compatible with current device limitations.
Variational quantum algorithms (VQAs) offer a potential solution by fixing the circuit architecture and optimizing individual gate parameters in an external loop. However, parameter optimization can become intractable, and the overall performance of the algorithm depends heavily on the initially chosen circuit architecture. Several quantum architecture search (QAS) algorithms have been developed to design useful circuit architectures automatically. In the case of parameter optimization alone, noise effects have been observed to dramatically influence the performance of the optimizer and final outcomes, which is a key line of study. However, the effects of noise on the architecture search, which could be just as critical, are poorly understood. This work addresses this gap by introducing a curriculum-based reinforcement learning QAS (CRLQAS) algorithm designed to tackle challenges in realistic VQA deployment. The algorithm incorporates (i) a 3D architecture encoding and restrictions on environment dynamics to explore the search space of possible circuits efficiently, (ii) an episode halting scheme to steer the agent to find shorter circuits, and (iii) a novel variant of simultaneous perturbation stochastic approximation as an optimizer for faster convergence. To facilitate studies, we developed an optimized simulator for our algorithm, significantly improving computational efficiency in simulating noisy quantum circuits by employing the Pauli-transfer matrix formalism in the Pauli-Liouville basis. Numerical experiments focusing on quantum chemistry tasks demonstrate that CRLQAS outperforms existing QAS algorithms across several metrics in both noiseless and noisy environments.
****
Fei Shen,Hu Ye,Jun Zhang,Cong Wang,Xiao Han,Yang Wei
Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models
Recent work has showcased the significant potential of diffusion models in pose-guided person image synthesis.
However, owing to the inconsistency in pose between the source and target images, synthesizing an image with a distinct pose, relying exclusively on the source image and target pose information, remains a formidable challenge.
This paper presents Progressive Conditional Diffusion Models (PCDMs) that incrementally bridge the gap between person images under the target and source poses through three stages.
Specifically, in the first stage, we design a simple prior conditional diffusion model that predicts the global features of the target image by mining the global alignment relationship between pose coordinates and image appearance.
Then, the second stage establishes a dense correspondence between the source and target images using the global features from the previous stage, and an inpainting conditional diffusion model is proposed to further align and enhance the contextual features, generating a coarse-grained person image.
In the third stage, we propose a refining conditional diffusion model to utilize the coarsely generated image from the previous stage as a condition, achieving texture restoration and enhancing fine-detail consistency.

The three-stage PCDMs work progressively to generate the final high-quality and high-fidelity synthesized image.
Both qualitative and quantitative results demonstrate the consistency and photorealism of our proposed PCDMs under challenging scenarios.
The code and model will be available at https://github.com/tencent-ailab/PCDMs.
****
Haeyong Kang,Jaehong Yoon,DaHyun Kim,Sung Ju Hwang,Chang D. Yoo
Progressive Fourier Neural Representation for Sequential Video Compilation
Neural Implicit Representation (NIR) has recently gained significant attention due to its remarkable ability to encode complex and high-dimensional data into representation space and easily reconstruct it through a trainable mapping function. However, NIR methods assume a one-to-one mapping between the target data and representation models regardless of data relevancy or similarity. This results in poor generalization over multiple complex data and limits their efficiency and scalability. Motivated by continual learning, this work investigates how to accumulate and transfer neural implicit representations for multiple complex video data over sequential encoding sessions. To overcome the limitation of NIR, we propose a novel method, Progressive Fourier Neural Representation (PFNR), that aims to find an adaptive and compact sub-module in Fourier space to encode videos in each training session. This sparsified neural encoding allows the neural network to hold free weights, enabling an improved adaptation for future videos. In addition, when learning a representation for a new video, PFNR transfers the representation of previous videos with frozen weights. This design allows the model to continuously accumulate high-quality neural representations for multiple videos while ensuring lossless decoding that perfectly preserves the learned representations for previous videos. We validate our PFNR method on the UVG8/17 and DAVIS50 video sequence benchmarks and achieve impressive performance gains over strong continual learning baselines.
****
Qiwei Di,Tao Jin,Yue Wu,Heyang Zhao,Farzad Farnoud,Quanquan Gu
Variance-aware Regret Bounds for Stochastic Contextual Dueling Bandits
Dueling bandits is a prominent framework for decision-making involving preferential feedback, a valuable feature that fits various applications involving human interaction, such as ranking, information retrieval, and recommendation systems. While substantial efforts have been made to minimize the cumulative regret in dueling bandits, a notable gap in the current research is the absence of regret bounds that account for the inherent uncertainty in pairwise comparisons between the dueling arms. Intuitively, greater uncertainty suggests a higher level of difficulty in the problem. To bridge this gap, this paper studies the problem of contextual dueling bandits, where the binary comparison of dueling arms is generated from a generalized linear model (GLM). We propose a new SupLinUCB-type algorithm that enjoys computational efficiency and a variance-aware regret bound $\tilde O\big(d\sqrt{\sum_{t=1}^T\sigma_t^2} + d\big)$, where σ_t is the variance of the pairwise comparison at round t, d is the dimension of the context vectors, and T is the time horizon. Our regret bound naturally aligns with the intuitive expectation — in scenarios where the comparison is deterministic, the algorithm only suffers from an $\tilde O(d)$ regret. We perform empirical experiments on synthetic data to confirm the advantage of our method over previous variance-agnostic algorithms.
****
David A. R. Robin,Kevin Scaman,Marc Lelarge
Random Sparse Lifts: Construction, Analysis and Convergence of finite sparse networks
We present a framework to define a large class of neural networks for which, by construction, training by gradient flow provably reaches arbitrarily low loss when the number of parameters grows. Distinct from the fixed-space global optimality of non-convex optimization, this new form of convergence, and the techniques introduced to prove such convergence, pave the way for a usable deep learning convergence theory in the near future, without overparameterization assumptions relating the number of parameters and training samples. We define these architectu

res from a simple computation graph and a mechanism to lift it, thus increasing the number of parameters, generalizing the idea of increasing the widths of multi-layer perceptrons. We show that architectures similar to most common deep learning models are present in this class, obtained by sparsifying the weight tensors of usual architectures at initialization. Leveraging tools of algebraic topology and random graph theory, we use the computation graph's geometry to propagate properties guaranteeing convergence to any precision for these large sparse models.
****
Safa Messaoud,Billel Mokeddem,Zhenghai Xue,Linsey Pang,Bo An,Haipeng Chen,Sanjay Chawla
S2AC: Energy-Based Reinforcement Learning with Stein Soft Actor Critic
Learning expressive stochastic policies instead of deterministic ones has been proposed to achieve better stability, sample complexity and robustness. Notably, in Maximum Entropy reinforcement learning (MaxEnt RL), the policy is modeled as an expressive energy-based model (EBM) over the Q-values. However, this formulation requires the estimation of the entropy of such EBM distributions which is an open problem. To address this, previous MaxEnt RL methods either implicitly estimate the entropy, yielding high computational complexity and variance (SQL), or follow a variational inference approach that fits simplified distributions (e.g., Gaussian) for tractability (SAC). We propose Sein Soft Actor-Critic (S^{2}AC), a MaxEnt RL algorithm that learns expressive policies without compromising efficiency. S^{2}AC uses parameterized Stein Variational Gradient Descent (SVGD) as the underlying policy. At the core of S^{2}AC is a new solution to the above open challenge of entropy computation for EBMs. Our entropy formula is computationally efficient and only depends on first-order derivatives and vector products. Empirical results show that S^{2}AC yields more optimal solutions to the MaxEnt objective than SQL and SAC in the multi-goal environment, and outperforms SAC and SQL on the MuJoCo benchmark. Our code is available at: https://anonymous.4open.science/r/Stein-Soft-Actor-Critic/
****
Amin Mansouri,Jason Hartford,Yan Zhang,Yoshua Bengio
Object centric architectures enable efficient causal representation learning
Causal representation learning has showed a variety of settings in which we can disentangle latent variables with identifiability guarantees (up to some reasonable equivalence class). Common to all of these approaches is the assumption that (1) the latent variables are represented as d-dimensional vectors, and (2) that the observations are the output of some injective generative function of these latent variables. While these assumptions appear benign, we show that when the observations are of multiple objects, the generative function is no longer injective and disentanglement fails in practice. We can address this failure by combining recent developments in object-centric learning and causal representation learning. By modifying the Slot Attention architecture (Locatello et al., 2020), we develop an object-centric architecture that leverages weak supervision from sparse perturbations to disentangle each object's properties. This approach is more data-efficient in the sense that it requires significantly fewer perturbations than a comparable approach that encodes to a Euclidean space and we show that this approach successfully disentangles the properties of a set of objects in a series of simple image-based disentanglement experiments.
****
Wenxuan Zhou,Sheng Zhang,Yu Gu,Muhao Chen,Hoifung Poon
UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition
Large language models (LLMs) have demonstrated remarkable generalizability, such as understanding arbitrary entities and relations. Instruction tuning has proven effective for distilling LLMs into more cost-efficient models such as Alpaca and Vicuna. Yet such student models still trail the original LLMs by large margins in downstream applications. In this paper, we explore targeted distillation with mission-focused instruction tuning to train student models that can excel in a broad application class such as open information extraction. Using named entit

y recognition (NER) for case study, we show how ChatGPT can be distilled into much smaller UniversalNER models for open NER. For evaluation, we assemble the largest NER benchmark to date, comprising 43 datasets across 9 diverse domains such as biomedicine, programming, social media, law, finance. Without using any direct supervision, UniversalNER attains remarkable NER accuracy across tens of thousands of entity types, outperforming general instruction-tuned models such as Alpaca and Vicuna by over 30 absolute F1 points in average. With a tiny fraction of parameters, UniversalNER not only acquires ChatGPT's capability in recognizing arbitrary entity types, but also outperforms its NER accuracy by 7-9 absolute F1 points in average. Remarkably, UniversalNER even outperforms by a large margin state-of-the-art multi-task instruction-tuned systems such as InstructUIE, which uses supervised NER examples. We also conduct thorough ablation studies to assess the impact of various components in our distillation approach. We release the distillation recipe, data, and UniversalNER models to facilitate future research on targeted distillation.
****
Joe Benton,Valentin De Bortoli,Arnaud Doucet,George Deligiannidis
Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization
Denoising diffusions are a powerful method to generate approximate samples from high-dimensional data distributions. Recent results provide polynomial bounds on their convergence rate, assuming L^2-accurate scores. Until now, the tightest bounds were either superlinear in the data dimension or required strong smoothness assumptions. We provide the first convergence bounds which are linear in the data dimension (up to logarithmic factors) assuming only finite second moments of the data distribution. We show that diffusion models require at most $\tilde O(\frac{d \log^2(1/\delta)}{\varepsilon^2})$ steps to approximate an arbitrary distribution on $\mathbb{R}^d$ corrupted with Gaussian noise of variance δ to within ε^2 in KL divergence. Our proof extends the Girsanov-based methods of previous works. We introduce a refined treatment of the error from discretizing the reverse SDE inspired by stochastic localization.
****
Yangsibo Huang,Samyak Gupta,Mengzhou Xia,Kai Li,Danqi Chen
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
The rapid progress in open-source large language models (LLMs) is significantly advancing AI development. Extensive efforts have been made before model release to align their behavior with human values, with the primary goal of ensuring their helpfulness and harmlessness. However, even carefully aligned models can be manipulated maliciously, leading to unintended behaviors, known as ``jailbreaks". These jailbreaks are typically triggered by specific text inputs, often referred to as adversarial prompts. In this work, we propose the generation exploitation attack, an extremely simple approach that disrupts model alignment by only manipulating variations of decoding methods. By exploiting different generation strategies, including varying decoding hyper-parameters and sampling methods, we increase the attack success rate from 0% to more than 95% across 11 language models including LLaMA2, Vicuna, Falcon, and MPT families, outperforming state-of-the-art attacks with $30\times$ lower computational cost. Finally, we propose an effective alignment method that explores diverse generation strategies, which can reasonably reduce the attack success rate under our attack. Altogether, our study underscores a major failure in current safety evaluation and alignment procedures for open-source LLMs, strongly advocating for more comprehensive red teaming and better alignment before releasing such models.
****
Kai Hu,Klas Leino,Zifan Wang,Matt Fredrikson
Effectively Leveraging Capacity for Improved Deterministic Robustness Certification
Recent studies have highlighted the potential of Lipschitz-based methods for training certifiably robust neural networks against adversarial attacks.
A key challenge, supported both theoretically and empirically, is that robustness demands greater network capacity and more data than standard training.

However, effectively adding capacity under stringent Lipschitz constraints has p
roven more difficult than it may seem, evident by the fact that state-of-the-art
 approach tend more towards \emph{underfitting} than overfitting.
Moreover, we posit that a lack of careful exploration of the design space for Li
pshitz-based approaches has left potential performance gains on the table.
In this work, we provide a more comprehensive evaluation to better uncover the p
otential of Lipschitz-based certification methods.
Using a combination of novel techniques, design optimizations, and synthesis of
prior work, we are able to significantly improve the state-of-the-art VRA for de
terministic certification on a variety of benchmark datasets, and over a range o
f perturbation sizes.
Of particular note, we discover that the addition of large ``Cholesky-orthogonal
ized residual dense'' layers to the end of existing state-of-the-art Lipschitz-c
ontrolled ResNet architectures is especially effective for increasing network ca
pacity and performance.
Combined with filtered generative data augmentation, our final results further t
he state of the art deterministic VRA by up to 8.5 percentage points.
****
Yonggang Zhang,Zhiqin Yang,Xinmei Tian,Nannan Wang,Tongliang Liu,Bo Han
Robust Training of Federated Models with Extremely Label Deficiency
Federated semi-supervised learning (FSSL) has emerged as a powerful paradigm for
 collaboratively training machine learning models using distributed data with la
bel deficiency. Advanced FSSL methods predominantly focus on training a single m
odel on each client. However, this approach could lead to a discrepancy between
the objective functions of labeled and unlabeled data, resulting in gradient con
flicts. To alleviate gradient conflict, we propose a novel twin-model paradigm,
called **Twinsight**, designed to enhance mutual guidance by providing insights
from different perspectives of labeled and unlabeled data. In particular, Twinsi
ght concurrently trains a supervised model with a supervised objective function
while training an unsupervised model using an unsupervised objective function. T
o enhance the synergy between these two models, Twinsight introduces a neighborh
ood-preserving constraint, which encourages the preservation of the neighborhood
 relationship among data features extracted by both models. Our comprehensive ex
periments on four benchmark datasets provide substantial evidence that Twinsight
 can significantly outperform state-of-the-art methods across various experiment
al settings, demonstrating the efficacy of the proposed Twinsight.
****
Bojan Karlaš,David Dao,Matteo Interlandi,Sebastian Schelter,Wentao Wu,Ce Zhang
Data Debugging with Shapley Importance over Machine Learning Pipelines
When a machine learning (ML) model exhibits poor quality (e.g., poor accuracy or
 fairness), the problem can often be traced back to errors in the training data.
 Being able to discover the data examples that are the most likely culprits is a
 fundamental concern that has received a lot of attention recently. One prominen
t way to measure "data importance" with respect to model quality is the Shapley
value. Unfortunately, existing methods only focus on the ML model in isolation,
without considering the broader ML pipeline for data preparation and feature ext
raction, which appears in the majority of real-world ML code. This presents a ma
jor limitation to applying existing methods in practical settings. In this paper
, we propose Datascope, a method for efficiently computing Shapley-based data im
portance over ML pipelines. We introduce several approximations that lead to dra
matic improvements in terms of computational speed. Finally, our experimental ev
aluation demonstrates that our methods are capable of data error discovery that
is as effective as existing Monte Carlo baselines, and in some cases even outper
form them. We release our code as an open-source data debugging library availabl
e at https://github.com/easeml/datascope.
****
Yanbang Wang,Jon Kleinberg
From Graphs to Hypergraphs: Hypergraph Projection and its Reconstruction
We study the implications of the modeling choice to use a graph, instead of a hy
pergraph, to represent real-world interconnected systems whose constituent relat

ionships are of higher order by nature. Such a modeling choice typically involve
s an underlying projection process that maps the original hypergraph onto a grap
h, and is prevalent in graph-based analysis. While hypergraph projection can po
tentially lead to loss of higher-order relations, there exists very limited stud
ies on the consequences of doing so, as well as its remediation. This work fills
 this gap by doing two things: (1) we develop analysis based on graph and set th
eory, showing two ubiquitous patterns of hyperedges that are root to structural
information loss in all hypergraph projections; we also quantify the combinatori
al impossibility of recovering the lost higher-order structures if no extra help
 is provided; (2) we still seek to recover the lost higher-order structures in h
ypergraph projection, and in light of (1)'s findings we make reasonable assumpti
ons to allow the help of some prior knowledge of the application domain. Under t
his problem setting, we develop a learning-based hypergraph reconstruction metho
d based on an important statistic of hyperedge distributions that we find. Our r
econstruction method is systematically evaluated on 8 real-world datasets under
different settings, and exhibits consistently top performance. We also demonstra
te benefits of the reconstructed hypergraphs through use cases of protein rankin
gs and link predictions.
****
Andrei Lupu,Chris Lu,Jarek Luca Liesen,Robert Tjarko Lange,Jakob Nicolaus Foerst
er
Behaviour Distillation
Dataset distillation aims to condense large datasets into a small number of synt
hetic examples that can be used as drop-in replacements when training new models
. It has applications to interpretability, neural architecture search, privacy,
and continual learning. Despite strong successes in supervised domains, such met
hods have not yet been extended to reinforcement learning, where the lack of fix
ed dataset renders most distillation methods unusable.
Filling the gap, we formalize $\textit{behaviour distillation}$, a setting that
aims to discover and then condense the information required for training an expe
rt policy into a synthetic dataset of state-action pairs, $\textit{without acces
s to expert data}$.
We then introduce Hallucinating Datasets with Evolution Strategies (HaDES), a me
thod for behaviour distillation that can discover datasets of $\textit{just four
}$ state-action pairs which, under supervised learning, train agents to competit
ive performance levels in continuous control tasks.
We show that these datasets generalize out of distribution to training policies
with a wide range of architectures and hyperparameters. We also demonstrate appl
ication to a downstream task, namely training multi-task agents in a zero-shot f
ashion.
Beyond behaviour distillation, HaDES provides significant improvements in neuroe
volution for RL over previous approaches and achieves SoTA results on one standa
rd supervised dataset distillation task. Finally, we show that visualizing the s
ynthetic datasets can provide human-interpretable task insights.
****
Marin Scalbert,Maria Vakalopoulou,Florent Couzinie-Devy
Towards domain-invariant Self-Supervised Learning with Batch Styles Standardizat
ion
In Self-Supervised Learning (SSL), models are typically pretrained, fine-tuned,
and evaluated on the same domains. However, they tend to perform poorly when eva
luated on unseen domains, a challenge that Unsupervised Domain Generalization (U
DG) seeks to address. Current UDG methods rely on domain labels, which are often
 challenging to collect, and domain-specific architectures that lack scalability
 when confronted with numerous domains, making the current methodology impractic
al and rigid. Inspired by contrastive-based UDG methods that mitigate spurious c
orrelations by restricting comparisons to examples from the same domain, we hypo
thesize that eliminating style variability within a batch could provide a more c
onvenient and flexible way to reduce spurious correlations without requiring dom
ain labels. To verify this hypothesis, we introduce Batch Styles Standardization
 (BSS), a relatively simple yet powerful Fourier-based method to standardize the

style of images in a batch specifically designed for integration with SSL methods to tackle UDG. Combining BSS with existing SSL methods offers serious advantages over prior UDG methods: (1) It eliminates the need for domain labels or domain-specific network components to enhance domain-invariance in SSL representations, and (2) offers flexibility as BSS can be seamlessly integrated with diverse contrastive-based but also non-contrastive-based SSL methods. Experiments on several UDG datasets demonstrate that it significantly improves downstream task performances on unseen domains, often outperforming or rivaling UDG methods. Finally, this work clarifies the underlying mechanisms contributing to BSS's effectiveness in improving domain-invariance in SSL representations and performances on unseen domains. Implementations of the extended SSL methods and BSS are provided at this [url](https://gitlab.com/vitadx/articles/towards-domain-invariant-ssl-through-bss).
****
Rohan Subramani,Marcus Williams,Max Heitmann,Halfdan Holm,Charlie Griffin,Joar Max Viktor Skalse
On the Expressivity of Objective-Specification Formalisms in Reinforcement Learning
Most algorithms in reinforcement learning (RL) require that the objective is formalised with a Markovian reward function. However, it is well-known that certain tasks cannot be expressed by means of an objective in the Markov rewards formalism, motivating the study of alternative objective-specification formalisms in RL such as Linear Temporal Logic and Multi-Objective Reinforcement Learning. To date, there has not yet been any thorough analysis of how these formalisms relate to each other in terms of their expressivity. We fill this gap in the existing literature by providing a comprehensive comparison of 17 salient objective-specification formalisms. We place these formalisms in a preorder based on their expressive power, and present this preorder as a Hasse diagram. We find a variety of limitations for the different formalisms, and argue that no formalism is both dominantly expressive and straightforward to optimise with current techniques. For example, we prove that each of Regularised RL, (Outer) Nonlinear Markov Rewards, Reward Machines, Linear Temporal Logic, and Limit Average Rewards can express a task that the others cannot. The significance of our results is twofold. First, we identify important expressivity limitations to consider when specifying objectives for policy optimization. Second, our results highlight the need for future research which adapts reward learning to work with a greater variety of formalisms, since many existing reward learning methods assume that the desired objective takes a Markovian form. Our work contributes towards a more cohesive understanding of the costs and benefits of different RL objective-specification formalisms.
****
Guy Tennenholtz,Yinlam Chow,ChihWei Hsu,Jihwan Jeong,Lior Shani,Azamat Tulepbergenov,Deepak Ramachandran,Martin Mladenov,Craig Boutilier
Demystifying Embedding Spaces using Large Language Models
Embeddings have become a pivotal means to represent complex, multi-faceted information about entities, concepts, and relationships in a condensed and useful format. Nevertheless, they often preclude direct interpretation. While downstream tasks make use of these compressed representations, meaningful interpretation usually requires visualization using dimensionality reduction or specialized machine learning interpretability methods. This paper addresses the challenge of making such embeddings more interpretable and broadly useful, by employing large language models (LLMs) to directly interact with embeddings -- transforming abstract vectors into understandable narratives. By injecting embeddings into LLMs, we enable querying and exploration of complex embedding data. We demonstrate our approach on a variety of diverse tasks, including: enhancing concept activation vectors (CAVs), communicating novel embedded entities, and decoding user preferences in recommender systems. Our work couples the immense information potential of embeddings with the interpretative power of LLMs.
****
Yanwei Wang,Tsun-Hsuan Wang,Jiayuan Mao,Michael Hagenow,Julie Shah

Grounding Language Plans in Demonstrations Through Counter-Factual Perturbations
Grounding the abstract knowledge captured by Large Language Models (LLMs) in physical domains remains a pivotal yet unsolved problem. Whereas prior works have largely focused on leveraging LLMs for generating abstract plans in symbolic spaces, this work uses LLMs to guide the learning for structures and constraints in robot manipulation tasks. Specifically, we borrow from manipulation planning literature the concept of mode families, defining specific types of motion constraints among sets of objects, to serve as an intermediate layer that connects high-level language representations with low-level physical trajectories. By locally perturbing a small set of successful human demonstrations, we augment the dataset with additional successful executions as well as counterfactuals that fail the task. Our explanation-based learning framework trains neural network-based classifiers to differentiate success task executions from failures and as a by-product learns classifiers that ground low-level states into mode families without dense labeling. This further enables us to learn structured policies for the target task.
Experimental validation in both 2D continuous-space and robotic manipulation environments demonstrates the robustness of our mode-based imitation methods under external perturbations.
****
Ashwinee Panda,Christopher A. Choquette-Choo,Zhengming Zhang,Yaoqing Yang,Prateek Mittal
Teach LLMs to Phish: Stealing Private Information from Language Models
When large language models are trained on private data, it can be a _significant_ privacy risk for them to memorize and regurgitate sensitive information. In this work, we propose a new _practical_ data extraction attack that we call ``neural phishing''. This attack enables an adversary to target and extract sensitive or personally identifiable information (PII), e.g., credit card numbers, from a model trained on user data with upwards of 10% secret extraction rates, at times, as high as 80%. Our attack assumes only that an adversary can insert only 10s of benign-appearing sentences into the training dataset
using only vague priors on the structure of the user data.
****
Matteo Pirotta,Andrea Tirinzoni,Ahmed Touati,Alessandro Lazaric,Yann Ollivier
Fast Imitation via Behavior Foundation Models
Imitation learning (IL) aims at producing agents that can imitate any behavior given a few expert demonstrations. Yet existing approaches require many demonstrations and/or running (online or offline) reinforcement learning (RL) algorithms for each new imitation task. Here we show that recent RL foundation models based on successor measures can imitate any expert behavior almost instantly with just a few demonstrations and no need for RL or fine-tuning, while accommodating several IL principles (behavioral cloning, feature matching, reward-based, and goal-based reductions). In our experiments, imitation via RL foundation models matches, and often surpasses, the performance of SOTA offline IL algorithms, and produces imitation policies from new demonstrations within seconds instead of hours.
****
Huangjie Zheng,Zhendong Wang,Jianbo Yuan,Guanghan Ning,Pengcheng He,Quanzeng You,Hongxia Yang,Mingyuan Zhou
Learning Stackable and Skippable LEGO Bricks for Efficient, Reconfigurable, and Variable-Resolution Diffusion Modeling
Diffusion models excel at generating photo-realistic images but come with significant computational costs in both training and sampling. While various techniques address these computational challenges, a less-explored issue is designing an efficient and adaptable network backbone for iterative refinement. Current options like U-Net and Vision Transformer often rely on resource-intensive deep netwo

rks and lack the flexibility needed for generating images at variable resolutions or with a smaller network than used in training.
This study introduces LEGO bricks, which seamlessly integrate Local-feature Enrichment and Global-content Orchestration. These bricks can be stacked to create a test-time reconfigurable diffusion backbone, allowing selective skipping of bricks to reduce sampling costs and generate higher-resolution images than the training data. LEGO bricks enrich local regions with an MLP and transform them using a Transformer block while maintaining a consistent full-resolution image across all bricks. Experimental results demonstrate that LEGO bricks enhance training efficiency, expedite convergence, and facilitate variable-resolution image generation while maintaining strong generative performance. Moreover, LEGO significantly reduces sampling time compared to other methods, establishing it as a valuable enhancement for diffusion models.
****
Geraud Nangue Tasse,Devon Jarvis,Steven James,Benjamin Rosman
Skill Machines: Temporal Logic Skill Composition in Reinforcement Learning
It is desirable for an agent to be able to solve a rich variety of problems that can be specified through language in the same environment. A popular approach towards obtaining such agents is to reuse skills learned in prior tasks to generalise compositionally to new ones. However, this is a challenging problem due to the curse of dimensionality induced by the combinatorially large number of ways high-level goals can be combined both logically and temporally in language. To address this problem, we propose a framework where an agent first learns a sufficient set of skill primitives to achieve all high-level goals in its environment. The agent can then flexibly compose them both logically and temporally to provably achieve temporal logic specifications in any regular language, such as regular fragments of linear temporal logic. This provides the agent with the ability to map from complex temporal logic task specifications to near-optimal behaviours zero-shot. We demonstrate this experimentally in a tabular setting, as well as in a high-dimensional video game and continuous control environment. Finally, we also demonstrate that the performance of skill machines can be improved with regular off-policy reinforcement learning algorithms when optimal behaviours are desired.
****
Andi Peng,Ilia Sucholutsky,Belinda Z. Li,Theodore Sumers,Thomas L. Griffiths,Jacob Andreas,Julie Shah
Learning with Language-Guided State Abstractions
We describe a framework for using natural language to design state abstractions for imitation learning.
Generalizable policy learning in high-dimensional observation spaces is facilitated by well-designed state representations, which can surface important features of an environment and hide irrelevant ones.
These state representations are typically manually specified, or derived from other labor-intensive labeling procedures.
Our method, LGA (\textit{language-guided abstraction}), uses a combination of natural language supervision and background knowledge from language models (LMs) to automatically build state representations tailored to unseen tasks.
In LGA, a user first provides a (possibly incomplete) description of a target task in natural language; next, a pre-trained LM translates this task description into a state abstraction function that masks out irrelevant features; finally, an imitation policy is trained using a small number of demonstrations and LGA-generated abstract states.
Experiments on simulated robotic tasks show that LGA yields state abstractions similar to those designed by humans, but in a fraction of the time, and that these abstractions improve generalization and robustness in the presence of spurious correlations and ambiguous specifications.
We illustrate the utility of the learned abstractions on mobile manipulation tasks with a Spot robot.
****
Manju Garimella,Denizhan Pak,Justin Newell Wood,Samantha Marie Waters Wood

A Newborn Embodied Turing Test for Comparing Object Segmentation Across Animals and Machines
Newborn brains rapidly learn to solve challenging object recognition tasks, incl uding segmenting objects from backgrounds and recognizing objects across novel b ackgrounds and viewpoints. Conversely, modern machine-learning (ML) algorithms a re "data hungry," requiring more training data than brains to reach similar perf ormance levels. How do we close this learning gap between brains and machines? H ere we introduce a new benchmark—a Newborn Embodied Turing Test (NETT) for objec t segmentation—in which newborn animals and machines are raised in the same envi ronments and tested with the same tasks, permitting direct comparison of their l earning abilities. First, we raised newborn chicks in controlled environments co ntaining a single object rotating on a single background, then tested their abil ity to recognize that object across new backgrounds and viewpoints. Second, we p erformed “digital twin” experiments in which we reared and tested artificial chi cks in virtual environments that mimicked the rearing and testing conditions of the biological chicks. We inserted a variety of ML “brains” into the artificial chicks and measured whether those algorithms learned common object recognition b ehavior as biological chicks. All biological chicks solved this one-shot object segmentation task, successfully learning background-invariant object representat ions that generalized across new backgrounds and viewpoints. In contrast, none o f the artificial chicks solved this object segmentation task, instead learning b ackground-dependent representations that failed to generalize across new backgro unds and viewpoints. This digital twin design exposes core limitations in curren t ML algorithms in achieving brain-like object perception. Our NETT is publicly available for comparing ML algorithms with newborn chicks. Ultimately, we antici pate that NETT benchmarks will allow researchers to build embodied AI systems th at learn as efficiently and robustly as newborn brains.
****
Zhiyuan Li,Yi Wang,Zhiren Wang
Fast Equilibrium of SGD in Generic Situations
Normalization layers are ubiquitous in deep learning, greatly accelerating optim ization. However, they also introduce many unexpected phenomena during training, for example, the Fast Equilibrium conjecture proposed by (Li et al.,2020), whic h states that the scale-invariant normalized network, when trained by SGD with η learning rate and λ weight decay, mixes to an equilibrium in $\tilde{O}(1/\eta\lambda)$ steps, as opposed to classical $e^{O(\eta^{-1})}$ mixing t ime. Recent works by Wang & Wang (2022); Li et al. (2022c) proved this conjectu re under different sets of assumptions. This paper aims to answer the fast equil ibrium conjecture in full generality by removing the non-generic assumptions of Wang & Wang (2022); Li et al. (2022c) that the minima are isolated, that the reg ion near minima forms a unique basin, and that the set of minima is an analytic set. Our main technical contribution is to show that with probability close to 1 , in exponential time trajectories will not escape the attracting basin contain ing its initial position.
****
Raj Ghugare,Matthieu Geist,Glen Berseth,Benjamin Eysenbach
Closing the Gap between TD Learning and Supervised Learning - A Generalisation P oint of View.
Some reinforcement learning (RL) algorithms have the capability of recombining t ogether pieces of previously seen experience to solve a task never seen before d uring training. This oft-sought property is one of the few ways in which dynamic programming based RL algorithms are considered different from supervised learni ng (SL) based RL algorithms. Yet, recent RL methods based on off-the-shelf SL al gorithms achieve excellent results without an explicit mechanism for stitching; it remains unclear whether those methods forgo this important stitching property . This paper studies this question in the setting of goal-reaching problems. We show that the desirable stitching property corresponds to a form of generalizati on: after training on a distribution of (state, goal) pairs, one would like to e valuate on (state, goal) pairs not seen together in the training data. Our analy sis shows that this sort of generalization is different from i.i.d. generalizati

on. This connection between stitching and generalization reveals why we should not expect existing RL methods based on SL to perform stitching, even in the limit of large datasets and models. We experimentally validate this result on carefully constructed datasets.
This connection suggests a simple remedy, the same remedy for improving generalization in supervised learning: data augmentation. We propose a naive temporal data augmentation approach and demonstrate that adding it to RL methods based on SL enables them to successfully stitch together experience, so that they succeed in navigating between states and goals unseen together during training.
****
Huaijin Wu,Wei Liu,Yatao Bian,Jiaxiang Wu,Nianzu Yang,Junchi Yan
EBMDock: Neural Probabilistic Protein-Protein Docking via a Differentiable Energy Model
Protein complex formation, a pivotal challenge in contemporary biology, has recently gained interest from the machine learning community, particularly concerning protein-ligand docking tasks. In this paper, we delve into the equally crucial but comparatively under-investigated domain of protein-protein docking. Specifically, we propose a geometric deep learning framework, termed EBMDock, which employs statistical potential as its energy function. This approach produces a probability distribution over docking poses, such that the identified docking pose aligns with a minimum point in the energy landscape. We employ a differential algorithm grounded in Langevin dynamics to efficiently sample from the docking pose distribution. Additionally, we incorporate energy-based training using contrastive divergence, enhancing both performance and stability. Empirical results demonstrate that our approach achieves superior performance on two benchmark datasets DIPS and DB5.5. Furthermore, the results suggest EBMDock can serve as an orthogonal enhancement to existing methods.
****
Moritz Akiya Zanger,Wendelin Boehmer,Matthijs T. J. Spaan
Diverse Projection Ensembles for Distributional Reinforcement Learning
In contrast to classical reinforcement learning, distributional RL algorithms aim to learn the distribution of returns rather than their expected value. Since the nature of the return distribution is generally unknown a priori or arbitrarily complex, a common approach finds approximations within a set of representable, parametric distributions. Typically, this involves a projection of the unconstrained distribution onto the set of simplified distributions. We argue that this projection step entails a strong inductive bias when coupled with neural networks and gradient descent, thereby profoundly impacting the generalization behavior of learned models. In order to facilitate reliable uncertainty estimation through diversity, this work studies the combination of several different projections and representations in a distributional ensemble. We establish theoretical properties of such projection ensembles and derive an algorithm that uses ensemble disagreement, measured by the average 1-Wasserstein distance, as a bonus for deep exploration. We evaluate our algorithm on the behavior suite benchmark and find that diverse projection ensembles lead to significant performance improvements over existing methods on a wide variety of tasks with the most pronounced gains in directed exploration problems.
****
Yuxin Li,Wenchao Chen,Xinyue Hu,Bo Chen,baolin sun,Mingyuan Zhou
Transformer-Modulated Diffusion Models for Probabilistic Multivariate Time Series Forecasting
Transformers have gained widespread usage in multivariate time series (MTS) forecasting, delivering impressive performance. Nonetheless, these existing transformer-based methods often neglect an essential aspect: the incorporation of uncertainty into the predicted series, which holds significant value in decision-making. In this paper, we introduce a Transformer-Modulated Diffusion Model (TMDM), uniting conditional diffusion generative process with transformers into a unified framework to enable precise distribution forecasting for MTS. TMDM harnesses the power of transformers to extract essential insights from historical time series data. This information is then utilized as prior knowledge, capturing covariat

e-dependence in both the forward and reverse processes within the diffusion mode l. Furthermore, we seamlessly integrate well-designed transformer-based forecast ing methods into TMDM to enhance its overall performance. Additionally, we intro duce two novel metrics for evaluating uncertainty estimation performance. Throug h extensive experiments on six datasets using four evaluation metrics, we establ ish the effectiveness of TMDM in probabilistic MTS forecasting.
****
Jiahao Zhang,Tao Lin,Weiqiang Zheng,Zhe Feng,Yifeng Teng,Xiaotie Deng
Learning Thresholds with Latent Values and Censored Feedback
In this paper, we investigate a problem of *actively* learning threshold in late nt space, where the *unknown* reward $g(\gamma, v)$ depends on the proposed thre shold γ and latent value v and it can be $only$ achieved if the thresho ld is lower than or equal to the *unknown* latent value. This problem has broad applications in practical scenarios, e.g., reserve price optimization in online auctions, online task assignments in crowdsourcing, setting recruiting bars in h iring, etc. We first characterize the query complexity of learning a threshold w ith the expected reward at most ϵ smaller than the optimum and prove th at the number of queries needed can be infinitely large even when $g(\gamma, v)$ is monotone with respect to both γ and v. On the positive side, we pro vide a tight query complexity $\tilde{\Theta}(1/\epsilon^3)$ when g is monoton e and the CDF of value distribution is Lipschitz. Moreover, we show a tight $\ti lde{\Theta}(1/\epsilon^3)$ query complexity can be achieved as long as g satis fies one-sided Lipschitzness, which provides a complete characterization for thi s problem. Finally, we extend this model to an online learning setting and demon strate a tight $\Theta(T^{2/3})$ regret bound using continuous-arm bandit techni ques and the aforementioned query complexity results.
****
Charilaos Kanatsoulis,Alejandro Ribeiro
Counting Graph Substructures with Graph Neural Networks
Graph Neural Networks (GNNs) are powerful representation learning tools that hav e achieved remarkable performance in various important tasks. However, their abi lity to count substructures, which play a crucial role in biological and social networks, remains uncertain. In this work, we fill this gap and characterize the representation power of GNNs in terms of their ability to produce powerful repr esentations that count graph substructures. In particular, we study the message- passing operations of GNNs with random stationary input and show that they can p roduce permutation equivariant representations that are associated with high-ord er statistical moments. Using these representations, we prove that GNNs can lear n how to count cycles, quasi-cliques, and the number of connected components in a graph. We also provide new insights into the generalization capacity of GNNs. Our analysis is constructive and enables the design of a generic GNN architectur e that shows remarkable performance in four distinct tasks: cycle detection, cyc le counting, graph classification, and molecular property prediction.
****
Tianle Cai,Xuezhi Wang,Tengyu Ma,Xinyun Chen,Denny Zhou
Large Language Models as Tool Makers
Recent research has highlighted the potential of large language models (LLMs) to improve their problem-solving capabilities with the aid of suitable external tools. In our work, we further advance this concept by introducing a closed- loop framework, referred to as LLMs A s Tool Makers (LATM), where LLMs create their own reusable tools for problem-solving. Our approach consists of tw o
phases: 1) tool making: an LLM acts as the tool maker that crafts tools for a se t
of tasks, where a tool is implemented as a Python utility function. 2) tool usin g:
another LLM acts as the tool user, which applies the tool built by the tool make r
for problem-solving. The tool user can be either the same or a different LLM from the tool maker. On the problem-solving server side, tool-making enables

continual tool generation and caching as new requests emerge. This framework
enables subsequent requests to access cached tools via their corresponding APIs,
enhancing the efficiency of task resolution. Beyond enabling LLMs to create thei
r
own tools, our framework also uncovers intriguing opportunities to optimize the
serving cost of LLMs: Recognizing that tool-making requires more sophisticated
capabilities, we assign this task to a powerful, albeit resource-intensive, mode
l.
Conversely, the simpler tool-using phase is delegated to a lightweight model. Th
is
strategic division of labor allows the once-off cost of tool-making to be spread
over multiple instances of tool-using, significantly reducing average costs whil
e
maintaining strong performance. Furthermore, our method offers a functional
cache through the caching and reuse of tools, which stores the functionality of
a class of requests instead of the natural language responses from LLMs, thus
extending the applicability of the conventional cache mechanism. We evaluate
our approach across various complex reasoning tasks, including Big-Bench tasks.
With GPT-4 as the tool maker and GPT-3.5 as the tool user, LATM demonstrates
performance equivalent to using GPT-4 for both roles, but with a significantly
reduced inference cost.
****
Jiayan Teng,Wendi Zheng,Ming Ding,Wenyi Hong,Jianqiao Wangni,Zhuoyi Yang,Jie Tan
g
Relay Diffusion: Unifying diffusion process across resolutions for image synthes
is
Diffusion models achieved great success in image synthesis, but still face chall
enges in high-resolution generation. Through the lens of discrete cosine transfo
rmation, we find the main reason is that *the same noise level on a higher resol
ution results in a higher Signal-to-Noise Ratio in the frequency domain*. In thi
s work, we present Relay Diffusion Model (RDM), which transfers a low-resolution
 image or noise into an equivalent high-resolution one for diffusion model via b
lurring diffusion and block noise. Therefore, the diffusion process can continue
 seamlessly in any new resolution or model without restarting from pure noise or
 low-resolution conditioning. RDM achieves state-of-the-art FID on CelebA-HQ and
 sFID on ImageNet 256$\times$256, surpassing previous works such as ADM, LDM and
 DiT by a large margin. All the codes and checkpoints are open-sourced at \url{h
ttps://github.com/THUDM/RelayDiffusion}.
****
Chunshu Wu,Ruibing Song,Chuan Liu,Yunan Yang,Ang Li,Michael Huang,Tong Geng
NP-GL: Extending Power of Nature from Binary Problems to Real-World Graph Learni
ng
Nature performs complex computations constantly at clearly lower cost and higher
 performance than digital computers. It is crucial to understand how to harness
the unique computational power of nature in Machine Learning (ML). In the past d
ecade, besides the development of Neural Networks (NNs), the community has also
relentlessly explored nature-powered ML paradigms. Although most of them are sti
ll predominantly theoretical, a new practical paradigm enabled by the recent adv
ent of CMOS-compatible room-temperature nature-based computers has emerged. By h
arnessing a dynamical system's intrinsic behavior of chasing the lowest energy s
tate, this paradigm can solve some simple binary problems delivering considerabl
e speedup and energy savings compared with NNs, while maintaining comparable acc
uracy. Regrettably, its values to the real world are highly constrained by its b
inary nature. A clear pathway to its extension to real-valued problems remains e
lusive. This paper aims to unleash this pathway by proposing a novel end-to-end
Nature-Powered Graph Learning (NP-GL) framework. Specifically, through a three-d
imensional co-design, NP-GL can leverage the spontaneous energy decrease in natu
re to efficiently solve real-valued graph learning problems. Experimental result
s across 4 real-world applications with 6 datasets demonstrate that NP-GL delive
rs, on average, 6.97×10^3 speedup and 10^5 energy consumption reductio

n with comparable or even higher accuracy than Graph Neural Networks (GNNs).
****
Zhenfeng He,Yao Shu,Zhongxiang Dai,Bryan Kian Hsiang Low
Robustifying and Boosting Training-Free Neural Architecture Search
Neural architecture search (NAS) has become a key component of AutoML and a standard tool to automate the design of deep neural networks. Recently, training-free NAS as an emerging paradigm has successfully reduced the search costs of standard training-based NAS by estimating the true architecture performance with only training-free metrics. Nevertheless, the estimation ability of these metrics typically varies across different tasks, making it challenging to achieve robust and consistently good search performance on diverse tasks with only a single training-free metric. Meanwhile, the estimation gap between training-free metrics and the true architecture performances limits training-free NAS to achieve superior performance. To address these challenges, we propose the robustifying and boosting training-free NAS (RoBoT) algorithm which (a) employs the optimized combination of existing training-free metrics explored from Bayesian optimization to develop a robust and consistently better-performing metric on diverse tasks, and (b) applies greedy search, i.e., the exploitation, on the newly developed metric to bridge the aforementioned gap and consequently to boost the search performance of standard training-free NAS further. Remarkably, the expected performance of our RoBoT can be theoretically guaranteed, which improves over the existing training-free NAS under mild conditions with additional interesting insights. Our extensive experiments on various NAS benchmark tasks yield substantial empirical evidence to support our theoretical results.
****
Celine Lee,Abdulrahman Mahmoud,Michal Kurek,Simone Campanoni,David Brooks,Stephen Chong,Gu-Yeon Wei,Alexander M Rush
Guess & Sketch: Language Model Guided Transpilation
Maintaining legacy software requires many software and systems engineering hours. Assembly code programs, which demand low-level control over the computer machine state and have no variable names, are particularly difficult for humans to analyze.
Existing conventional program translators guarantee correctness, but are hand-engineered for the source and target programming languages in question. Learned transpilation, i.e. automatic translation of code, offers an alternative to manual re-writing and engineering efforts. Automated symbolic program translation approaches guarantee correctness but struggle to scale to longer programs due to the exponentially large search space. Their rigid rule-based systems also limit their expressivity, so they can only reason about a reduced space of programs. Probabilistic neural language models (LMs) produce plausible outputs for every input, but do so at the cost of guaranteed correctness. In this work, we leverage the strengths of LMs and symbolic solvers in a neurosymbolic approach to learned transpilation for assembly code. Assembly code is an appropriate setting for a neurosymbolic approach, since assembly code can be divided into shorter non-branching basic blocks amenable to the use of symbolic methods. Guess & Sketch extracts alignment and confidence information from features of the LM then passes it to a symbolic solver to resolve semantic equivalence of the transpilation input and output. We test Guess & Sketch on three different test sets of assembly transpilation tasks, varying in difficulty, and show that it successfully transpiles 57.6% more examples than GPT-4 and 39.6% more examples than an engineered transpiler. We also share a training and evaluation dataset for this task.
****
Advait Harshal Gadhikar,Rebekka Burkholz
Masks, Signs, And Learning Rate Rewinding
Learning Rate Rewinding (LRR) has been established as a strong variant of Iterative Magnitude Pruning (IMP) to find lottery tickets in deep overparameterized neural networks. While both iterative pruning schemes couple structure and parameter learning, understanding how LRR excels in both aspects can bring us closer to the design of more flexible deep learning algorithms that can optimize diverse sets of sparse architectures. To this end, we conduct experiments that disentang

le the effect of mask learning and parameter optimization and how both benefit f
rom overparameterization. The ability of LRR to flip parameter signs early and s
tay robust to sign perturbations seems to make it not only more effective in mas
k identification but also in optimizing diverse sets of masks, including random
 ones. In support of this hypothesis, we prove in a simplified single hidden neu
ron setting that LRR succeeds in more cases than IMP, as it can escape initially
 problematic sign configurations.
****
Yu Tian,Min Shi,Yan Luo,Ava Kouhana,Tobias Elze,Mengyu Wang
FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning
Using Segment Anything Model with Fair Error-Bound Scaling
Fairness in artificial intelligence models has gained significantly more attenti
on in recent years, especially in the area of medicine, as fairness in medical m
odels is critical to people's well-being and lives. High-quality medical fairnes
s datasets are needed to promote fairness learning research. Existing medical fa
irness datasets are all for classification tasks, and no fairness datasets are a
vailable for medical segmentation, while medical segmentation is an equally impo
rtant clinical task as classifications, which can provide detailed spatial infor
mation on organ abnormalities ready to be assessed by clinicians. In this paper,
 we propose the first fairness dataset for medical segmentation named Harvard-Fa
irSeg with 10,000 subject samples. In addition, we propose a fair error-bound sc
aling approach to reweight the loss function with the upper error-bound in each
identity group, using the segment anything model (SAM). We anticipate that the s
egmentation performance equity can be improved by explicitly tackling the hard c
ases with high training errors in each identity group. To facilitate fair compar
isons, we utilize a novel equity-scaled segmentation performance metric to compa
re segmentation metrics in the context of fairness, such as the equity-scaled Di
ce coefficient. Through comprehensive experiments, we demonstrate that our fair
error-bound scaling approach either has superior or comparable fairness performa
nce to the state-of-the-art fairness learning models. The dataset and code are p
ublicly accessible via https://ophai.hms.harvard.edu/datasets/harvard-fairseg10k
.
****
Elias Stengel-Eskin,Kyle Rawlins,Benjamin Van Durme
Zero and Few-shot Semantic Parsing with Ambiguous Inputs
Despite the frequent challenges posed by ambiguity when representing meaning via
 natural language, it is often ignored or deliberately removed in tasks mapping
language to formally-designed representations, which generally assume a one-to-o
ne mapping between linguistic and formal representations.
We attempt to address this shortcoming by introducing AmP, a framework, dataset,
 and challenge for translating ambiguous natural language to formal representati
ons like logic and code.
We define templates and generate data for five well-documented linguistic ambigu
ities.
Using AmP, we investigate how several few-shot text-to-code systems handle ambig
uity, introducing three new metrics.
We find that large pre-trained models perform poorly at capturing the distributi
on of possible meanings without deliberate instruction.
However, models are able to capture the distribution well when ambiguity is atte
sted in their inputs.
These results motivate a call for including ambiguity explicitly in datasets and
 promote considering the distribution of possible outputs when evaluating system
s. We release our data and code.
****
Mucong Ding,Bang An,Yuancheng Xu,Anirudh Satheesh,Furong Huang
SAFLEX: Self-Adaptive Augmentation via Feature Label Extrapolation
Data augmentation, a cornerstone technique in deep learning, is crucial in enhan
cing model performance, especially with scarce labeled data. While traditional t
echniques are effective, their reliance on hand-crafted methods limits their app
licability across diverse data types and tasks. Although modern learnable augmen

tation methods offer increased adaptability, they are computationally expensive and challenging to incorporate within prevalent augmentation workflows. In this work, we present a novel, efficient method for data augmentation, effectively bridging the gap between existing augmentation strategies and emerging datasets and learning tasks. We introduce SAFLEX (Self-Adaptive Augmentation via Feature Label EXtrapolation), which learns the sample weights and soft labels of augmented samples provided by any given upstream augmentation pipeline, using a specifically designed efficient bilevel optimization algorithm. Remarkably, SAFLEX effectively reduces the noise and label errors of the upstream augmentation pipeline with a marginal computational cost. As a versatile module, SAFLEX excels across diverse datasets, including natural and medical images and tabular data, showcasing its prowess in few-shot learning and out-of-distribution generalization. SAFLEX seamlessly integrates with common augmentation strategies like RandAug, CutMix, and those from large pre-trained generative models like stable diffusion and is also compatible with frameworks such as CLIP's fine-tuning. Our findings highlight the potential to adapt existing augmentation pipelines for new data types and tasks, signaling a move towards more adaptable and resilient training frameworks.
****
Lionel Wong,Jiayuan Mao,Pratyusha Sharma,Zachary S Siegel,Jiahai Feng,Noa Korneev,Joshua B. Tenenbaum,Jacob Andreas
Learning Grounded Action Abstractions from Language
Effective planning in the real world requires not only world knowledge, but the ability to leverage that knowledge to build the right representation of the task at hand. Decades of hierarchical planning techniques have used domain-specific temporal action abstractions to support efficient and accurate planning, almost always relying on human priors and domain knowledge to decompose hard tasks into smaller subproblems appropriate for a goal or set of goals. This paper describes Ada (Action Domain Acquisition), a framework for automatically constructing task-specific planning representations using task-general background knowledge from language models (LMs). Starting with a general-purpose hierarchical planner and a low-level goal-conditioned policy, Ada interactively learns a library of planner-compatible high-level action abstractions and low-level controllers adapted to a particular domain of planning tasks. On two language-guided interactive planning benchmarks (Mini Minecraft and ALFRED Household Tasks), Ada strongly outperforms other approaches that use LMs for sequential decision-making, offering more accurate plans and better generalization to complex tasks.
****
Xianfan Gu,Chuan Wen,Weirui Ye,Jiaming Song,Yang Gao
Seer: Language Instructed Video Prediction with Latent Diffusion Models
Imagining the future trajectory is the key for robots to make sound planning and successfully reach their goals. Therefore, text-conditioned video prediction (TVP) is an essential task to facilitate general robot policy learning.
To tackle this task and empower robots with the ability to foresee the future, we propose a sample and computation-efficient model, named Seer, by inflating the pretrained text-to-image (T2I) stable diffusion models along the temporal axis. We enhance the U-Net and language conditioning model by incorporating computation-efficient spatial-temporal attention. Furthermore, we introduce a novel Frame Sequential Text Decomposer module that dissects a sentence's global instruction into temporally aligned sub-instructions, ensuring precise integration into each frame of generation. Our framework allows us to effectively leverage the extensive prior knowledge embedded in pretrained T2I models across the frames.
With the adaptable-designed architecture, Seer makes it possible to generate high-fidelity, coherent, and instruction-aligned video frames by fine-tuning a few layers on a small amount of data. The experimental results on Something Something V2 (SSv2), Bridgedata and EpicKitchens-100 datasets demonstrate our superior video prediction performance with around 480-GPU hours versus CogVideo with over 12,480-GPU hours: achieving the 31\% FVD improvement compared to the current SOTA model on SSv2 and 83.7\% average preference in the human evaluation. Our project is available at https://seervideodiffusion.github.io/

****
Zhilin Huang,Ling Yang,Xiangxin Zhou,Zhilong Zhang,Wentao Zhang,Xiawu Zheng,Jie Chen,Yu Wang,Bin CUI,Wenming Yang
Protein-Ligand Interaction Prior for Binding-aware 3D Molecule Diffusion Models
Generating 3D ligand molecules that bind to specific protein targets via diffusion models has shown great promise for structure-based drug design. The key idea is to disrupt molecules into noise through a fixed forward process and learn its reverse process to generate molecules from noise in a denoising way. However, existing diffusion models primarily focus on incorporating protein-ligand interaction information solely in the reverse process, and neglect the interactions in the forward process. The inconsistency between forward and reverse processes may impair the binding affinity of generated molecules towards target protein. In this paper, we propose a novel Interaction Prior-guided Diffusion model (IPDiff) for the protein-specific 3D molecular generation by introducing geometric protein-ligand interactions into both diffusion and sampling process. Specifically, we begin by pretraining a protein-ligand interaction prior network (IPNet) by utilizing the binding affinity signals as supervision. Subsequently, we leverage the pretrained prior network to (1) integrate interactions between the target protein and the molecular ligand into the forward process for adapting the molecule diffusion trajectories (prior-shifting), and (2) enhance the binding-aware molecule sampling process (prior-conditioning). Empirical studies on CrossDocked2020 dataset show IPDiff can generate molecules with more realistic 3D structures and state-of-the-art binding affinities towards the protein targets, with up to -6.42 Avg. Vina Score, while maintaining proper molecular properties. https://github.com/YangLing0818/IPDiff
****
Debo Cheng,Ziqi Xu,Jiuyong Li,Lin Liu,Jixue Liu,Thuc Duy Le
Conditional Instrumental Variable Regression with Representation Learning for Causal Inference
This paper studies the challenging problem of estimating causal effects from observational data, in the presence of unobserved confounders. The two-stage least square (TSLS) method and its variants with a standard instrumental variable (IV) are commonly used to eliminate confounding bias, including the bias caused by unobserved confounders, but they rely on the linearity assumption. Besides, the strict condition of unconfounded instruments posed on a standard IV is too strong to be practical. To address these challenging and practical problems of the standard IV method (linearity assumption and the strict condition), in this paper, we use a conditional IV (CIV) to relax the unconfounded instrument condition of standard IV and propose a non-linear \underline{CIV} regression with \underline{C}onfounding \underline{B}alancing \underline{R}epresentation \underline{L}earning, CBRL.CIV, for jointly eliminating the confounding bias from unobserved confounders and balancing the observed confounders, without the linearity assumption. We theoretically demonstrate the soundness of CBRL.CIV. Extensive experiments on synthetic and two real-world datasets show the competitive performance of CBRL.CIV against state-of-the-art IV-based estimators and superiority in dealing with the non-linear situation.
****
Zichuan Liu,Yingying ZHANG,Tianchun Wang,Zefan Wang,Dongsheng Luo,Mengnan Du,Min Wu,Yi Wang,Chunlin Chen,Lunting Fan,Qingsong Wen
Explaining Time Series via Contrastive and Locally Sparse Perturbations
Explaining multivariate time series is a compound challenge, as it requires identifying important locations in the time series and matching complex temporal patterns.
Although previous saliency-based methods addressed the challenges,
their perturbation may not alleviate the distribution shift issue, which is inevitable especially in heterogeneous samples.
We present ContraLSP, a locally sparse model that introduces counterfactual samples to build uninformative perturbations but keeps distribution using contrastive learning.
Furthermore, we incorporate sample-specific sparse gates to generate more binary

-skewed and smooth masks, which easily integrate temporal trends and select the salient features parsimoniously.
Empirical studies on both synthetic and real-world datasets show that ContraLSP outperforms state-of-the-art models, demonstrating a substantial improvement in explanation quality for time series data.
The source code is available at \url{https://github.com/zichuan-liu/ContraLSP}.
****
Haonan Yu,Wei Xu
VONet: Unsupervised Video Object Learning With Parallel U-Net Attention and Object-wise Sequential VAE
Unsupervised video object learning seeks to decompose video scenes into structural object representations without any supervision from depth, optical flow, or segmentation. We present VONet, an innovative approach that is inspired by MONet. While utilizing a U-Net architecture, VONet employs an efficient and effective parallel attention inference process, generating attention masks for all slots simultaneously. Additionally, to enhance the temporal consistency of each mask across consecutive video frames, VONet develops an object-wise sequential VAE framework. The integration of these innovative encoder-side techniques, in conjunction with an expressive transformer-based decoder, establishes VONet as the leading unsupervised method for object learning across five MOVI datasets, encompassing videos of diverse complexities. Code is available at https://github.com/hnyu/vonet.
****
Zecheng Tang,Chenfei Wu,Juntao Li,Nan Duan
LayoutNUWA: Revealing the Hidden Layout Expertise of Large Language Models
Graphic layout generation, a growing research field, plays a significant role in user engagement and information perception.
Existing methods primarily treat layout generation as a numerical optimization task, focusing on quantitative aspects while overlooking the semantic information of layout, such as the relationship between each layout element.
In this paper, we propose LayoutNUWA, the first model that treats layout generation as a code generation task to enhance semantic information and harness the hidden layout expertise of large language models~(LLMs).
Concretely, we develop a Code Instruct Tuning (CIT) approach comprising three interconnected modules: 1) the Code Initialization (CI) module quantifies the numerical conditions and initializes them as HTML code with strategically placed masks; 2) the Code Completion (CC) module employs the formatting knowledge of LLMs to fill in the masked portions within the HTML code; 3) the Code Rendering (CR) module transforms the completed code into the final layout output, ensuring a highly interpretable and transparent layout generation procedure that directly maps code to a visualized layout. We attain significant state-of-the-art performance (even over 50\% improvements compared to previous works) on multiple datasets, showcasing the strong capabilities of LayoutNUWA.
****
Aochuan Chen,Yimeng Zhang,Jinghan Jia,James Diffenderfer,Konstantinos Parasyris,Jiancheng Liu,Yihua Zhang,Zheng Zhang,Bhavya Kailkhura,Sijia Liu
DeepZero: Scaling Up Zeroth-Order Optimization for Deep Model Training
Zeroth-order (ZO) optimization has become a popular technique for solving machine learning (ML) problems when first-order (FO) information is difficult or impossible to obtain. However, the scalability of ZO optimization remains an open problem: Its use has primarily been limited to relatively small-scale ML problems, such as sample-wise adversarial attack generation. To our best knowledge, no prior work has demonstrated the effectiveness of ZO optimization in training deep neural networks (DNNs) without a significant decrease in performance. To overcome this roadblock, we develop DeepZero, a principled and practical ZO deep learning (DL) framework that can scale ZO optimization to DNN training from scratch through three primary innovations. First, we demonstrate the advantages of coordinate-wise gradient estimation (CGE) over randomized vector-wise gradient estimation in training accuracy and computational efficiency. Second, we propose a sparsity-induced ZO training protocol that extends the model pruning methodology using

only finite differences to explore and exploit the sparse DL prior in CGE. Third, we develop the methods of feature reuse and forward parallelization to advance the practical implementations of ZO training. Our extensive experiments show that DeepZero achieves state-of-the-art (SOTA) accuracy on ResNet-20 trained on CIFAR-10, approaching FO training performance for the first time. Furthermore, we show the practical utility of DeepZero in applications of certified adversarial defense and DL-based partial differential equation error correction, achieving 10-20% improvement over SOTA. We believe our results will inspire future research on scalable ZO optimization and contribute to advancing deep learning.
****
Sijia Chen,Baochun Li,Di Niu
Boosting of Thoughts: Trial-and-Error Problem Solving with Large Language Models
The reasoning performance of Large Language Models (LLMs) on a wide range of problems critically relies on chain-of-thought prompting, which involves providing a few chain of thought demonstrations as exemplars in prompts. Recent work, e.g., Tree of Thoughts, has pointed out the importance of exploration and self-evaluation in reasoning step selection for complex problem solving. In this paper, we present Boosting of Thoughts (BoT), an automated prompting framework for problem solving with LLMs by iteratively exploring and self-evaluating many trees of thoughts in order to acquire an ensemble of trial-and-error reasoning experiences, which will serve as a new form of prompting to solve the complex problem. Starting from a simple prompt without requiring examples, BoT iteratively explores and evaluates a large collection of reasoning steps, and more importantly, uses error analysis obtained from the LLM on them to explicitly revise prompting, which in turn enhances reasoning step generation, until a final answer is attained. Our experiments with GPT-4 and Llama2 across extensive complex mathematical problems demonstrate that BoT consistently achieves higher or comparable problem-solving rates than other advanced prompting approaches.
****
Jing Xiong,Zixuan Li,Chuanyang Zheng,Zhijiang Guo,Yichun Yin,Enze Xie,Zhicheng YANG,Qingxing Cao,Haiming Wang,Xiongwei Han,Jing Tang,Chengming Li,Xiaodan Liang
DQ-LoRe: Dual Queries with Low Rank Approximation Re-ranking for In-Context Learning
Recent advances in natural language processing, primarily propelled by Large Language Models (LLMs), have showcased their remarkable capabilities grounded in in-context learning. A promising avenue for guiding LLMs in intricate reasoning tasks involves the utilization of intermediate reasoning steps within the Chain-of-Thought (CoT) paradigm. Nevertheless, the central challenge lies in the effective selection of exemplars for facilitating in-context learning. In this study, we introduce a framework that leverages Dual Queries and Low-rank approximation Re-ranking (DQ-LoRe) to automatically select exemplars for in-context learning. Dual Queries first query LLM to obtain LLM-generated knowledge such as CoT, then query the retriever to obtain the final exemplars via both question and the knowledge. Moreover, for the second query, LoRe employs dimensionality reduction techniques to refine exemplar selection, ensuring close alignment with the input question's knowledge. Through extensive experiments, we demonstrate that DQ-LoRe significantly outperforms prior state-of-the-art methods in the automatic selection of exemplars for GPT-4, enhancing performance from 92.5\% to 94.2\%. Our comprehensive analysis further reveals that DQ-LoRe consistently outperforms retrieval-based approaches in terms of both performance and adaptability, especially in scenarios characterized by distribution shifts. DQ-LoRe pushes the boundaries of in-context learning and opens up new avenues for addressing complex reasoning challenges.
****
Kushagra Pandey,Maja Rudolph,Stephan Mandt
Efficient Integrators for Diffusion Generative Models
Diffusion models suffer from slow sample generation at inference time. Therefore, developing a principled framework for fast deterministic/stochastic sampling for a broader class of diffusion models is a promising direction. We propose two complementary frameworks for accelerating sample generation in pre-trained model

s: Conjugate Integrators and Splitting Integrators. Conjugate integrators generalize DDIM, mapping the reverse diffusion dynamics to a more amenable space for sampling. In contrast, splitting-based integrators, commonly used in molecular dynamics, reduce the numerical simulation error by cleverly alternating between numerical updates involving the data and auxiliary variables. After extensively studying these methods empirically and theoretically, we present a hybrid method that leads to the best-reported performance for diffusion models in augmented spaces. Applied to Phase Space Langevin Diffusion [Pandey \& Mandt, 2023] on CIFAR-10, our deterministic and stochastic samplers achieve FID scores of 2.11 and 2.36 in only 100 network function evaluations (NFE) as compared to 2.57 and 2.63 for the best-performing baselines, respectively. Our code and model checkpoints will be made publicly available at https://github.com/mandt-lab/PSLD
****
Heng Dong,Junyu Zhang,Chongjie Zhang
Leveraging Hyperbolic Embeddings for Coarse-to-Fine Robot Design
Multi-cellular robot design aims to create robots comprised of numerous cells that can be efficiently controlled to perform diverse tasks. Previous research has demonstrated the ability to generate robots for various tasks, but these approaches often optimize robots directly in the vast design space, resulting in robots with complicated morphologies that are hard to control. In response, this paper presents a novel coarse-to-fine method for designing multi-cellular robots. Initially, this strategy seeks optimal coarse-grained robots and progressively refines them. To mitigate the challenge of determining the precise refinement juncture during the coarse-to-fine transition, we introduce the Hyperbolic Embeddings for Robot Design (HERD) framework. HERD unifies robots of various granularity within a shared hyperbolic space and leverages a refined Cross-Entropy Method for optimization. This framework enables our method to autonomously identify areas of exploration in hyperbolic space and concentrate on regions demonstrating promise. Finally, the extensive empirical studies on various challenging tasks sourced from EvoGym show our approach's superior efficiency and generalization capability.
****
Ziang Cao,Fangzhou Hong,Tong Wu,Liang Pan,Ziwei Liu
Large-Vocabulary 3D Diffusion Model with Transformer
Creating diverse and high-quality 3D assets with an automatic generative model is highly desirable. Despite extensive efforts on 3D generation, most existing works focus on the generation of a single category or a few categories. In this paper, we introduce a diffusion-based feed-forward framework for synthesizing massive categories of real-world 3D objects \textit{with a single generative model}. Notably, there are three major challenges for this large-vocabulary 3D generation: \textbf{a}) the need for expressive yet efficient 3D representation; \textbf{b}) large diversity in geometry and texture across categories; \textbf{c}) complexity in the appearances of real-world objects. To this end, we propose a novel triplane-based 3D-aware \textbf{Diff}usion model with \textbf{T}rans\textbf{F}ormer, \textbf{DiffTF}, for handling challenges via three aspects. \textbf{1}) Considering efficiency and robustness, we adopt a revised triplane representation and improve the fitting speed and accuracy. \textbf{2}) To handle the drastic variations in geometry and texture, we regard the features of all 3D objects as a combination of generalized 3D knowledge and specialized 3D features. To extract generalized 3D knowledge from diverse categories, we propose a novel 3D-aware transformer with shared cross-plane attention. It learns the cross-plane relations across different planes and aggregates the generalized 3D knowledge with specialized 3D features. \textbf{3}) In addition, we devise the 3D-aware encoder/decoder to enhance the generalized 3D knowledge in the encoded triplanes for handling categories with complex appearances. Extensive experiments on ShapeNet and OmniObject3D (over 200 diverse real-world categories) convincingly demonstrate that a single DiffTF model achieves state-of-the-art large-vocabulary 3D object generation performance with large diversity, rich semantics, and high quality. More results are available at https://difftf.github.io/
****

Subha Maity,Mayank Agarwal,Mikhail Yurochkin,Yuekai Sun
An Investigation of Representation and Allocation Harms in Contrastive Learning
The effect of underrepresentation on the performance of minority groups is known to be a serious problem in supervised learning settings; however, it has been underexplored so far in the context of self-supervised learning (SSL). In this paper, we demonstrate that contrastive learning (CL), a popular variant of SSL, tends to collapse representations of minority groups with certain majority groups. We refer to this phenomenon as representation harm and demonstrate it on image and text datasets using the corresponding popular CL methods. Furthermore, our causal mediation analysis of allocation harm on a downstream classification task reveals that representation harm is partly responsible for it, thus emphasizing the importance of studying and mitigating representation harm. Finally, we provide a theoretical explanation for representation harm using a stochastic block model that leads to a representational neural collapse in a contrastive learning setting.
****
Shikai Fang,Madison Cooley,Da Long,Shibo Li,Mike Kirby,Shandian Zhe
Solving High Frequency and Multi-Scale PDEs with Gaussian Processes
Machine learning based solvers have garnered much attention in physical simulation and scientific computing, with a prominent example, physics-informed neural networks (PINNs). However, PINNs often struggle to solve high-frequency and multi-scale PDEs, which can be due to spectral bias during neural network training. To address this problem, we resort to the Gaussian process (GP) framework. To flexibly capture the dominant frequencies, we model the power spectrum of the PDE solution with a student t mixture or Gaussian mixture. We apply the inverse Fourier transform to obtain the covariance function (by Wiener-Khinchin theorem). The covariance derived from the Gaussian mixture spectrum corresponds to the known spectral mixture kernel. Next, we estimate the mixture weights in the log domain, which we show is equivalent to placing a Jeffreys prior. It automatically induces sparsity, prunes excessive frequencies, and adjusts the remaining toward the ground truth. Third, to enable efficient and scalable computation on massive collocation points, which are critical to capture high frequencies, we place the collocation points on a grid, and multiply our covariance function at each input dimension. We use the GP conditional mean to predict the solution and its derivatives so as to fit the boundary condition and the equation itself.
As a result, we can derive a Kronecker product structure in the covariance matrix. We use Kronecker product properties and multilinear algebra to promote computational efficiency and scalability, without low-rank approximations. We show the advantage of our method in systematic experiments. The code is released at {https://github.com/xuangu-fang/Gaussian-Process-Slover-for-High-Freq-PDE}.
****
Binchi Zhang,Yushun Dong,Chen Chen,Yada Zhu,Minnan Luo,Jundong Li
Adversarial Attacks on Fairness of Graph Neural Networks
Fairness-aware graph neural networks (GNNs) have gained a surge of attention as they can reduce the bias of predictions on any demographic group (e.g., female) in graph-based applications. Although these methods greatly improve the algorithmic fairness of GNNs, the fairness can be easily corrupted by carefully designed adversarial attacks. In this paper, we investigate the problem of adversarial attacks on fairness of GNNs and propose G-FairAttack, a general framework for attacking various types of fairness-aware GNNs in terms of fairness with an unnoticeable effect on prediction utility. In addition, we propose a fast computation technique to reduce the time complexity of G-FairAttack. The experimental study demonstrates that G-FairAttack successfully corrupts the fairness of different types of GNNs while keeping the attack unnoticeable. Our study on fairness attacks sheds light on potential vulnerabilities in fairness-aware GNNs and guides further research on the robustness of GNNs in terms of fairness.
****
Arpit Bansal,Hong-Min Chu,Avi Schwarzschild,Soumyadip Sengupta,Micah Goldblum,Jonas Geiping,Tom Goldstein
Universal Guidance for Diffusion Models

Typical diffusion models are trained to accept a particular form of conditioning
, most commonly text, and cannot be conditioned on other modalities without retr
aining. In this work, we propose a universal guidance algorithm that enables dif
fusion models to be controlled by arbitrary guidance modalities without the need
 to retrain any use-specific components. We show that our algorithm successfully
 generates quality images with guidance functions including segmentation, face r
ecognition, object detection, style guidance and classifier signals.
****
Victor Quach,Adam Fisch,Tal Schuster,Adam Yala,Jae Ho Sohn,Tommi S. Jaakkola,Reg
ina Barzilay
Conformal Language Modeling
In this paper, we propose a novel approach to conformal prediction for language
 models (LMs) in which we produce prediction sets with performance guarantees. L
M responses are typically sampled from a predicted distribution over the large,
combinatorial output space of language. Translating this to conformal predictio
n, we calibrate a stopping rule for sampling LM outputs that get added to a gro
wing set of candidates until we are confident that the set covers at least one a
cceptable response. Since some samples may be low-quality, we also simultaneousl
y calibrate a rejection rule for removing candidates from the output set to redu
ce noise. Similar to conformal prediction, we can prove that the final output s
et obeys certain desirable distribution-free guarantees. Within these sets of ca
ndidate responses, we also show that we can also identify subsets of individual
components---such as phrases or sentences---that are each independently correct
(e.g., that are not ``hallucinations''), again with guarantees. Our method can b
e applied to any LM API that supports sampling. Furthermore, we empirically demo
nstrate that we can achieve many desired coverage levels within a limited number
 of total samples when applying our method to multiple tasks in open-domain que
stion answering, text summarization, and radiology report generation using diffe
rent LM variants.
****
Pascal Chang,Jingwei Tang,Markus Gross,Vinicius C. Azevedo
How I Warped Your Noise: a Temporally-Correlated Noise Prior for Diffusion Model
s
Video editing and generation methods often rely on pre-trained image-based diffu
sion models. During the diffusion process, however, the reliance on rudimentary
noise sampling techniques that do not preserve correlations present in subsequen
t frames of a video is detrimental to the quality of the results. This either pr
oduces high-frequency flickering, or texture-sticking artifacts that are not ame
nable to post-processing. With this in mind, we propose a novel method for prese
rving temporal correlations in a sequence of noise samples. This approach is mat
erialized by a novel noise representation, dubbed $\int$-noise (integral noise),
 that reinterprets individual noise samples as a continuously integrated noise f
ield: pixel values do not represent discrete values, but are rather the integral
 of an underlying infinite-resolution noise over the pixel area. Additionally, w
e propose a carefully tailored transport method that uses $\int$-noise to accura
tely advect noise samples over a sequence of frames, maximizing the correlation
between different frames while also preserving the noise properties. Our results
 demonstrate that the proposed $\int$-noise can be used for a variety of tasks,
such as video restoration, surrogate rendering, and conditional video generation
.
****
Joar Max Viktor Skalse,Alessandro Abate
Quantifying the Sensitivity of Inverse Reinforcement Learning to Misspecificatio
n
Inverse reinforcement learning (IRL) aims to infer an agent's *preferences* (rep
resented as a reward function R) from their *behaviour* (represented as a poli
cy π). To do this, we need a *behavioural model* of how π relates to R
. In the current literature, the most common behavioural models are *optimality*
, *Boltzmann-rationality*, and *causal entropy maximisation*. However, the true
relationship between a human's preferences and their behaviour is much more comp

lex than any of these behavioural models. This means that the behavioural models
 are *misspecified*, which raises the concern that they may lead to systematic e
rrors if applied to real data. In this paper, we analyse how sensitive the IRL p
roblem is to misspecification of the behavioural model. Specifically, we provide
 necessary and sufficient conditions that completely characterise how the observ
ed data may differ from the assumed behavioural model without incurring an error
 above a given threshold. In addition to this, we also characterise the conditio
ns under which a behavioural model is robust to small perturbations of the obser
ved policy, and we analyse how robust many behavioural models are to misspecific
ation of their parameter values (such as e.g. the discount rate). Our analysis s
uggests that the IRL problem is highly sensitive to misspecification, in the sen
se that very mild misspecification can lead to very large errors in the inferred
 reward function.
****
Qinbin Li,Chulin Xie,Xiaojun Xu,Xiaoyuan Liu,Ce Zhang,Bo Li,Bingsheng He,Dawn So
ng
Effective and Efficient Federated Tree Learning on Hybrid Data
Federated learning has emerged as a promising distributed learning paradigm that
 facilitates collaborative learning among multiple parties without transferring
raw data. However, most existing federated learning studies focus on either hori
zontal or vertical data settings, where the data of different parties are assume
d to be from the same feature or sample space. In practice, a common scenario is
 the hybrid data setting, where data from different parties may differ both in t
he features and samples. To address this, we propose HybridTree, a novel federat
ed learning approach that enables federated tree learning on hybrid data. We obs
erve the existence of consistent split rules in trees. With the help of these sp
lit rules, we theoretically show that the knowledge of parties can be incorporat
ed into the lower layers of a tree. Based on our theoretical analysis, we propos
e a layer-level solution that does not need frequent communication traffic to tr
ain a tree. Our experiments demonstrate that HybridTree can achieve comparable a
ccuracy to the centralized setting with low computational and communication over
head. HybridTree can achieve up to 8 times speedup compared with the other basel
ines.
****
Krzysztof Kacprzyk,Samuel Holt,Jeroen Berrevoets,Zhaozhi Qian,Mihaela van der Sc
haar
ODE Discovery for Longitudinal Heterogeneous Treatment Effects Inference
Inferring unbiased treatment effects has received widespread attention in the ma
chine learning community. In recent years, our community has proposed numerous s
olutions in standard settings, high-dimensional treatment settings, and even lon
gitudinal settings. While very diverse, the solution has mostly relied on neural
 networks for inference and simultaneous correction of assignment bias. New appr
oaches typically build on top of previous approaches by proposing new (or refine
d) architectures and learning algorithms. However, the end result—a neural-netwo
rk-based inference machine—remains unchallenged. In this paper, we introduce a d
ifferent type of solution in the longitudinal setting: a closed-form ordinary di
fferential equation (ODE). While we still rely on continuous optimization to lea
rn an ODE, the resulting inference machine is no longer a neural network. Doing
so yields several advantages such as interpretability, irregular sampling, and a
 different set of identification assumptions. Above all, we consider the introdu
ction of a completely new type of solution to be our most important contribution
 as it may spark entirely new innovations in treatment effects in general. We fa
cilitate this by formulating our contribution as a framework that can transform
any ODE discovery method into a treatment effects method.
****
Jiawei Sun,Kailai Li,Ruoxin Chen,Jie LI,Chentao Wu,Yue Ding,Junchi Yan
InterpGNN: Understand and Improve Generalization Ability of Transdutive GNNs thr
ough the Lens of Interplay between Train and Test Nodes
Transductive node prediction has been a popular learning setting in Graph Neural
 Networks (GNNs). It has been widely observed that the shortage of information f

low between the distant nodes and intra-batch nodes (for large-scale graphs) oft
en hurt the generalization of GNNs which overwhelmingly adopt message-passing. Y
et there is still no formal and direct theoretical results to quantitatively cap
ture the underlying mechanism, despite the recent advance in both theoretical an
d empirical studies for GNN's generalization ability. In this paper, the L-hop
 interplay (i.e., message passing capability with training nodes) for a L-lay
er GNN is successfully incorporated in our derived PAC-Bayesian bound for GNNs i
n the semi-supervised transductive setting. In other words, we quantitatively sh
ow how the interplay between training and testing sets influence the generalizat
ion ability which also partly explains the effectiveness of some existing empiri
cal methods for enhancing generalization. Based on this result, we further desig
n a plug-and-play ***Graph** **G**lobal **W**orkspace* module for GNNs (InterpGN
N-GW) to enhance the interplay, utilizing the key-value attention mechanism to s
ummarize crucial nodes' embeddings into memory and broadcast the memory to all n
odes, in contrast to the pairwise attention scheme in previous graph transformer
s. Extensive experiments on both small-scale and large-scale graph datasets vali
date the effectiveness of our theory and approaches.
****
Xiaoxiao Sun,Xingjian Leng,Zijian Wang,Yang Yang,Zi Huang,Liang Zheng
CIFAR-10-Warehouse: Broad and More Realistic Testbeds in Model Generalization An
alysis
Analyzing model performance in various unseen environments is a critical researc
h problem in the machine learning community. To study this problem, it is import
ant to construct a testbed with out-of-distribution test sets that have broad co
verage of environmental discrepancies. However, existing testbeds typically eith
er have a small number of domains or are synthesized by image corruptions, hinde
ring algorithm design that demonstrates real-world effectiveness. In this paper,
 we introduce CIFAR-10-Warehouse, consisting of 180 datasets collected by prompt
ing image search engines and diffusion models in various ways. Generally sized b
etween 300 and 8,000 images, the datasets contain natural images, cartoons, cert
ain colors, or objects that do not naturally appear. With CIFAR-10-W, we aim to
enhance the evaluation and deepen the understanding of two generalization tasks:
 domain generalization and model accuracy prediction in various out-of-distribut
ion environments. We conduct extensive benchmarking and comparison experiments a
nd show that CIFAR-10-W offers new and interesting insights inherent to these ta
sks. We also discuss other fields that would benefit from CIFAR-10-W. Data and c
ode are available at https://sites.google.com/view/CIFAR-10-warehouse/.
****
Cristian Meo,Louis Mahon,Anirudh Goyal,Justin Dauwels
αTC-VAE: On the relationship between Disentanglement and Diversity
Understanding and developing optimal representations has long been foundational
in machine learning (ML). While disentangled representations have shown promise
in generative modeling and representation learning, their downstream usefulness
remains debated. Recent studies re-defined disentanglement through a formal conn
ection to symmetries, emphasizing the ability to reduce latent domains (i.e., ML
 problem spaces) and consequently enhance data efficiency and generative capabil
ities. However, from an information theory viewpoint, assigning a complex attrib
ute (i.e., features) to a specific latent variable may be infeasible, limiting t
he applicability of disentangled representations to simple datasets. In this wor
k, we introduce α-TCVAE, a variational autoencoder optimized using a nove
l total correlation (TC) lower bound that maximizes disentanglement and latent v
ariables informativeness. The proposed TC bound is grounded in information theor
y constructs, generalizes the β-VAE lower bound, and can be reduced to a c
onvex combination of the known variational information bottleneck (VIB) and cond
itional entropy bottleneck (CEB) terms. Moreover, we present quantitative analys
es and correlation studies that support the idea that smaller latent domains (i.
e., disentangled representations) lead to better generative capabilities and div
ersity. Additionally, we perform downstream task experiments from both represent
ation and RL domains to assess our questions from a broader ML perspective. Our
results demonstrate that α-TCVAE consistently learns more disentangled re

presentations than baselines and generates more diverse observations without sac rificing visual fidelity. Notably, α-TCVAE exhibits marked improvements o n MPI3D-Real, the most realistic disentangled dataset in our study, confirming its ability to represent complex datasets when maximizing the informativeness of individual variables. Finally, testing the proposed model off-the-shelf on a st ate-of-the-art model-based RL agent, Director, significantly shows α-TCVA E downstream usefulness on the loconav Ant Maze task. Implementation available a t https://github.com/Cmeo97/Alpha-TCVAE
****
Tianzhe Chu,Shengbang Tong,Tianjiao Ding,Xili Dai,Benjamin David Haeffele,Rene V idal,Yi Ma
Image Clustering via the Principle of Rate Reduction in the Age of Pretrained Mo dels
The advent of large pre-trained models has brought about a paradigm shift in bot h visual representation learning and natural language processing. However, clust ering unlabeled images, as a fundamental and classic machine learning problem, s till lacks an effective solution, particularly for large-scale datasets. In this paper, we propose a novel image clustering pipeline that leverages the powerful feature representation of large pre-trained models such as CLIP and cluster ima ges effectively and efficiently at scale. We first developed a novel algorithm t o estimate the number of clusters in a given dataset. We then show that the pre- trained features are significantly more structured by further optimizing the rat e reduction objective. The resulting features may significantly improve the clus tering accuracy, e.g., from 57\% to 66\% on ImageNet-1k. Furthermore, by levera ging CLIP's multimodality bridge between image and text, we develop a simple yet effective self-labeling algorithm that produces meaningful text labels for the clusters. Through extensive experiments, we show that our pipeline works well on standard datasets such as CIFAR-10, CIFAR-100, and ImageNet-1k. It also extends to datasets without predefined labels, such as LAION-Aesthetics and WikiArts.
****
Keming Lu,Hongyi Yuan,Zheng Yuan,Runji Lin,Junyang Lin,Chuanqi Tan,Chang Zhou,Ji ngren Zhou
#InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Langu age Models
Pre-trained large language models (LLMs) can understand and align with human ins tructions by supervised fine-tuning (SFT).
It is commonly believed that diverse and complex SFT data are of the essence to enable good instruction-following abilities.
However, such diversity and complexity are obscure and lack quantitative analyse s.
In this work, we propose InsTag, an open-set instruction tagging method, to iden tify semantics and intentions of human instructions by tags that provide access to definitions and quantified analyses of instruction diversity and complexity.
We obtain 6.6K fine-grained tags to describe instructions from popular open-sour ced SFT datasets comprehensively.
We find that the abilities of aligned LLMs benefit from more diverse and complex instructions in SFT data.
Based on this observation, we propose a data sampling procedure based on InsTag, and select 6K diverse and complex samples from open-source datasets for SFT.
The resulting models, TagLM, outperform open-source models based on considerably larger SFT data evaluated by MT-Bench, echoing the importance of instruction di versity and complexity and the effectiveness of InsTag.
InsTag has robust potential to be extended to more applications beyond the data selection as it provides an effective way to analyze the distribution of instruc tions.
****
Hang Yu,Cong Liao,Ruolan Liu,Jianguo Li,Hu Yun,Xinzhe Wang
AmortizedPeriod: Attention-based Amortized Inference for Periodicity Identificat ion
Periodic patterns are a fundamental characteristic of time series in natural wor

ld, with significant implications for a range of disciplines, from economics to cloud systems. However, the current literature on periodicity detection faces two key challenges: limited robustness in real-world scenarios and a lack of memory to leverage previously observed time series to accelerate and improve inference on new data. To overcome these obstacles, this paper presents AmortizedPeriod, an innovative approach to periodicity identification based on amortized variational inference that integrates Bayesian statistics and deep learning. Through the Bayesian generative process, our method flexibly captures the dependencies of the periods, trends, noise, and outliers in time series, while also considering missing data and irregular periods in a robust manner. In addition, it utilizes the evidence lower bound of the log-likelihood of the observed time series as the loss function to train a deep attention inference network, facilitating knowledge transfer from the seen time series (and their labels) to unseen ones. Experimental results show that AmortizedPeriod surpasses the state-of-the-art methods by a large margin of 28.5% on average in terms of micro F_1-score, with at least 55% less inference time.
****
Yingyu Lin,Yian Ma,Yu-Xiang Wang,Rachel Emily Redberg,Zhiqi Bu
Tractable MCMC for Private Learning with Pure and Gaussian Differential Privacy
Posterior sampling, i.e., exponential mechanism to sample from the posterior distribution, provides ε-pure differential privacy (DP) guarantees and does not suffer from potentially unbounded privacy breach introduced by (ε,δ)-approximate DP. In practice, however, one needs to apply approximate sampling methods such as Markov chain Monte Carlo (MCMC), thus re-introducing the unappealing δ-approximation error into the privacy guarantees. To bridge this gap, we propose the Approximate SAample Perturbation (abbr. ASAP) algorithm which perturbs an MCMC sample with noise proportional to its Wasserstein-infinity (W_∞) distance from a reference distribution that satisfies pure DP or pure Gaussian DP (i.e., $\delta=0$). We then leverage a Metropolis-Hastings algorithm to generate the sample and prove that the algorithm converges in W$_\infty$ distance. We show that by combining our new techniques with a localization step, we obtain the first nearly linear-time algorithm that achieves the optimal rates in the DP-ERM problem with strongly convex and smooth losses.
****
Erfan Shayegani,Yue Dong,Nael Abu-Ghazaleh
Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models
We introduce new jailbreak attacks on vision language models (VLMs), which use aligned LLMs and are resilient to text-only jailbreak attacks. Specifically, we develop cross-modality attacks on alignment where we pair adversarial images going through the vision encoder with textual prompts to break the alignment of the language model. Our attacks employ a novel compositional strategy that combines an image, adversarially targeted towards toxic embeddings, with generic prompts to accomplish the jailbreak. Thus, the LLM draws the context to answer the generic prompt from the adversarial image. The generation of benign-appearing adversarial images leverages a novel embedding-space-based methodology, operating with no access to the LLM model. Instead, the attacks require access only to the vision encoder and utilize one of our four embedding space targeting strategies. By not requiring access to the LLM, the attacks lower the entry barrier for attackers, particularly when vision encoders such as CLIP are embedded in closed-source LLMs. The attacks achieve a high success rate across different VLMs, highlighting the risk of cross-modality alignment vulnerabilities, and the need for new alignment approaches for multi-modal models.
****
Haanvid Lee,Tri Wahyu Guntara,Jongmin Lee,Yung-Kyun Noh,Kee-Eung Kim
Kernel Metric Learning for In-Sample Off-Policy Evaluation of Deterministic RL Policies
We consider off-policy evaluation (OPE) of deterministic target policies for reinforcement learning (RL) in environments with continuous action spaces. While it is common to use importance sampling for OPE, it suffers from high variance whe

n the behavior policy deviates significantly from the target policy. In order to address this issue, some recent works on OPE proposed in-sample learning with importance resampling. Yet, these approaches are not applicable to deterministic target policies for continuous action spaces. To address this limitation, we propose to relax the deterministic target policy using a kernel and learn the kernel metrics that minimize the overall mean squared error of the estimated temporal difference update vector of an action value function, where the action value function is used for policy evaluation. We derive the bias and variance of the estimation error due to this relaxation and provide analytic solutions for the optimal kernel metric. In empirical studies using various test domains, we show that the OPE with in-sample learning using the kernel with optimized metric achieves significantly improved accuracy than other baselines.
****
Sherry Yang,Yilun Du,Bo Dai,Dale Schuurmans,Joshua B. Tenenbaum,Pieter Abbeel
Probabilistic Adaptation of Black-Box Text-to-Video Models
Large text-to-video models trained on internet-scale data have demonstrated exceptional capabilities in generating high-fidelity videos from arbitrary textual descriptions. However, similar to proprietary language models, large text-to-video models are often black boxes whose weight parameters are not publicly available, posing a significant challenge to adapting these models to specific domains such as robotics, animation, and personalized stylization. Inspired by how a large language model can be prompted to perform new tasks without access to the model weights, we investigate how to adapt a black-box pretrained text-to-video model to a variety of downstream domains without weight access to the pretrained model. In answering this question, we propose \emph{\methodname}, which leverages the score function of a large pretrained video diffusion model as a probabilistic prior to guide the generation of a task-specific small video model. Our experiments show that, by incorporating broad knowledge and fidelity of the pretrained model probabilistically, a small model with as few as 1.25% parameters of the pretrained model can generate high-quality yet domain-specific videos for a variety of downstream domains such as animation, egocentric modeling, and modeling of simulated and real-world robotics data. As large text-to-video models starting to become available as a service similar to large language models, we advocate for private institutions to expose scores of video diffusion models as outputs in addition to generated videos to allow flexible adaptation of large pretrained text-to-video models by the general public.
****
Xiang Lisa Li,Vaishnavi Shrivastava,Siyan Li,Tatsunori Hashimoto,Percy Liang
Benchmarking and Improving Generator-Validator Consistency of Language Models
As of September 2023, ChatGPT correctly answers “what is 7+8” with 15, but when asked “7+8=15, True or False” it responds with “False”. This inconsistency between generating and validating an answer is prevalent in language models (LMs) and erodes trust. In this paper, we propose a framework for measuring the consistency between generation and validation (which we call generator-validator consistency, or GV-consistency), finding that even GPT-4 (0613), a state-of-the-art LM, is GV-consistent only 76% of the time. To improve the consistency of LMs, we propose to finetune on the filtered generator and validator responses that are GV-consistent, and call this approach consistency fine-tuning. We find that this approach improves GV-consistency of Alpaca-30B from 60% to 93%, and the improvement extrapolates to unseen tasks and domains (e.g., GV-consistency for positive style transfers extrapolates to unseen styles like humor). In addition to improving consistency, consistency fine-tuning improves both generator quality and validator accuracy without using any labeled data. Evaluated across 6 tasks, including math questions, knowledge-intensive QA, and instruction following, our method improves generator quality by an average of 16% and validator accuracy by an average of 6.3% across all tasks.
****
Sachin Goyal,Ziwei Ji,Ankit Singh Rawat,Aditya Krishna Menon,Sanjiv Kumar,Vaishnavh Nagarajan
Think before you speak: Training Language Models With Pause Tokens

Language models generate responses by producing a series of tokens in immediate succession: the $(K+1)^{\rm th}$ token is an outcome of manipulating K hidden vectors per layer, one vector per preceding token. What if instead we were to let the model manipulate say, $K+10$ hidden vectors, before it outputs the $(K+1)^{\rm th}$ token? We operationalize this idea by performing
training and inference on language models with a (learnable) $\textit{pause}$ token, a sequence of which is appended to the input prefix. We then delay extracting the model's outputs until the last pause token is seen, thereby allowing the model to process extra computation before committing to an answer. We empirically evaluate $\textit{pause-training}$ on decoder-only models of 1B and 130M parameters with causal pretraining on C4, and on downstream tasks covering reasoning, question-answering, general understanding and fact recall. Our main finding is that inference-time delays show gains when the model is both pre-trained and finetuned with delays. For the 1B model, we witness gains on 8 of 9 tasks, most prominently, a gain of $18\\%$ EM score on the QA task of SQuAD, $8\\%$ on CommonSenseQA and $1\\%$ accuracy on the reasoning task of GSM8k. Our work raises a range of conceptual and practical future research questions on making delayed next-token prediction a widely applicable new paradigm.
****
Raphael Poulain,Rahmatollah Beheshti
Graph Transformers on EHRs: Better Representation Improves Downstream Performance
Following the success of transformer-based methods across various machine learning applications, their adoption to healthcare predictive tasks using electronic health records (EHR) has also expanded extensively. Similarly, graph-based methods have been shown to be very effective in capturing inherent graph-type relationships in EHRs, leading to improved downstream performance. Although integrating these two families of approaches seems like a natural next step, in practice, creating such a design is challenging and has not been done. This is partly due to known EHR problems, such as high sparsity, making extracting meaningful temporal representations of medical visits challenging. In this study, we propose GT-BEHRT, a new approach that leverages temporal visit embeddings extracted from a graph transformer and uses a BERT-based model to obtain more robust patient representations, especially on longer EHR sequences. The graph-based approach allows GT-BEHRT to implicitly capture the intrinsic graphical relationships between medical observations, while the BERT model extracts the temporal relationships between visits, loosely mimicking the clinicians' decision-making process. As part of our method, we also present a two-step pre-training strategy for learning better graphical and temporal representations. Our proposed method achieves state-of-the-art performance in a variety of standard medical predictive tasks, demonstrating the versatility of our approach.
****
Wu Ran,Peirong Ma,Zhiquan He,Hao Ren,Hong Lu
Harnessing Joint Rain-/Detail-aware Representations to Eliminate Intricate Rains
Recent advances in image deraining have focused on training powerful models on mixed multiple datasets comprising diverse rain types and backgrounds. However, this approach tends to overlook the inherent differences among rainy images, leading to suboptimal results. To overcome this limitation, we focus on addressing various rainy images by delving into meaningful representations that encapsulate both the rain and background components. Leveraging these representations as instructive guidance, we put forth a Context-based Instance-level Modulation (CoI-M) mechanism adept at efficiently modulating CNN- or Transformer-based models. Furthermore, we devise a rain-/detail-aware contrastive learning strategy to help extract joint rain-/detail-aware representations. By integrating CoI-M with the rain-/detail-aware Contrastive learning, we develop [CoIC](https://github.com/Schizophreni/CoIC), an innovative and potent algorithm tailored for training models on mixed datasets. Moreover, CoIC offers insight into modeling relationships of datasets, quantitatively assessing the impact of rain and details on restoration, and unveiling distinct behaviors of models given diverse inputs. Extensive experiments validate the efficacy of CoIC in boosting the deraining ability of CN

N and Transformer models. CoIC also enhances the deraining prowess remarkably wh
en real-world dataset is included.
****
LI Yang,RUIZHENG WU,Jiyong Li,Ying-Cong Chen
GNeRP: Gaussian-guided Neural Reconstruction of Reflective Objects with Noisy Po
larization Priors
Learning surfaces from neural radiance field (NeRF) became a rising topic in Mul
ti-View Stereo (MVS). Recent Signed Distance Function (SDF)-based methods demon
strated their ability to reconstruct exact 3D shapes of Lambertian scenes. Howev
er, their results on reflective scenes are unsatisfactory due to the entanglemen
t of specular radiance and complicated geometry. To address the challenges, we p
ropose a Gaussian-based representation of normals in SDF fields. Supervised by p
olarization priors, this representation guides the learning of geometry behind t
he specular reflection and capture more details than existing methods. Moreover
, we propose a reweighting strategy in optimization process to alleviate the noi
se issue of polarization priors. To validate the effectiveness of our design, we
 capture polarimetric information and ground truth meshes in additional reflecti
ve scenes with various geometry. We also evaluated our framework on PANDORA data
set. Both qualitative and quantitative comparisons prove our method outperforms
existing neural 3D reconstruction methods in reflective scenes by a large margin
.
****
Ning Miao,Yee Whye Teh,Tom Rainforth
SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning
The recent progress in large language models (LLMs), especially the invention of
 chain-of-thought prompting, has made it possible to automatically answer questi
ons by stepwise reasoning. However, when faced with more complicated problems th
at require non-linear thinking, even the strongest LLMs make mistakes. To addres
s this, we explore whether LLMs are able to recognize errors in their own step-b
y-step reasoning, without resorting to external resources. To this end, we propo
se SelfCheck, a general-purpose zero-shot verification schema for recognizing su
ch errors. We then use the results of these checks to improve question-answering
 performance by conducting weighted voting on multiple solutions to the question
. We test SelfCheck on math- and logic-based datasets and find that it successfu
lly recognizes errors and, in turn, increases final answer accuracies.
****
Tianyang Liu,Canwen Xu,Julian McAuley
RepoBench: Benchmarking Repository-Level Code Auto-Completion Systems
Large Language Models (LLMs) have greatly advanced code auto-completion systems,
 with a potential for substantial productivity enhancements for developers. Howe
ver, current benchmarks mainly focus on single-file tasks, leaving an assessment
 gap for more complex, real-world, multi-file programming scenarios. To fill thi
s gap, we introduce RepoBench, a new benchmark specifically designed for evaluat
ing repository-level code auto-completion systems. RepoBench consists of three i
nterconnected evaluation tasks: RepoBench-R (Retrieval), RepoBench-C (Code Compl
etion), and RepoBench-P (Pipeline). Each task respectively measures the system's
 ability to retrieve the most relevant code snippets from other files as cross-f
ile context, predict the next line of code with cross-file and in-file context,
and handle complex tasks that require a combination of both retrieval and next-l
ine prediction. RepoBench aims to facilitate a more complete comparison of perfo
rmance and encouraging continuous improvement in auto-completion systems. RepoBe
nch is actively maintained with the latest code, serving as a live benchmark pub
licly available at https://github.com/Leolty/repobench.
****
Zhongpai Gao,Huayi Zhou,Abhishek Sharma,Meng Zheng,Benjamin Planche,Terrence Che
n,Ziyan Wu
PBADet: A One-Stage Anchor-Free Approach for Part-Body Association
The detection of human parts (e.g., hands, face) and their correct association w
ith individuals is an essential task, e.g., for ubiquitous human-machine interfa
ces and action recognition. Traditional methods often employ multi-stage process

es, rely on cumbersome anchor-based systems, or do not scale well to larger part
 sets. This paper presents PBADet, a novel one-stage, anchor-free approach for p
art-body association detection. Building upon the anchor-free object representat
ion across multi-scale feature maps, we introduce a singular part-to-body center
 offset that effectively encapsulates the relationship between parts and their p
arent bodies. Our design is inherently versatile and capable of managing multipl
e parts-to-body associations without compromising on detection accuracy or robus
tness. Comprehensive experiments on various datasets underscore the efficacy of
our approach, which not only outperforms existing state-of-the-art techniques bu
t also offers a more streamlined and efficient solution to the part-body associa
tion challenge.
****
Yuyang Liu,Weijun Dong,Yingdong Hu,Chuan Wen,Zhao-Heng Yin,Chongjie Zhang,Yang G
ao
Imitation Learning from Observation with Automatic Discount Scheduling
Humans often acquire new skills through observation and imitation. For robotic a
gents, learning from the plethora of unlabeled video demonstration data availabl
e on the Internet necessitates imitating the expert without access to its action
, presenting a challenge known as Imitation Learning from Observation (ILfO). A
common approach to tackle ILfO problems is to convert them into inverse reinforc
ement learning problems, utilizing a proxy reward computed from the agent's and
the expert's observations. Nonetheless, we identify that tasks characterized by
a progress dependency property pose significant challenges for such approaches;
in these tasks, the agent needs to initially learn the expert's preceding behavi
ors before mastering the subsequent ones. Our investigation reveals that the mai
n cause is that the reward signals assigned to later steps hinder the learning o
f initial behaviors. To address this challenge, we present a novel ILfO framewor
k that enables the agent to master earlier behaviors before advancing to later o
nes. We introduce an Automatic Discount Scheduling (ADS) mechanism that adaptive
ly alters the discount factor in reinforcement learning during the training phas
e, prioritizing earlier rewards initially and gradually engaging later rewards o
nly when the earlier behaviors have been mastered. Our experiments, conducted on
 nine Meta-World tasks, demonstrate that our method significantly outperforms st
ate-of-the-art methods across all tasks, including those that are unsolvable by
them. Our code is available at https://il-ads.github.io.
****
Jonathan Richens,Tom Everitt
Robust agents learn causal world models
It has long been hypothesised that causal reasoning plays a fundamental role in
robust and general intelligence. However, it is not known if agents must learn c
ausal models in order to generalise to new domains, or if other inductive biases
 are sufficient. We answer this question, showing that any agent capable of sati
sfying a regret bound for a large set of distributional shifts must have learned
 an approximate causal model of the data generating process, which converges to
the true causal model for optimal agents. We discuss the implications of this re
sult for several research areas including transfer learning and causal inference
.
****
Aryaman Reddi,Maximilian Tölle,Jan Peters,Georgia Chalvatzaki,Carlo D'Eramo
Robust Adversarial Reinforcement Learning via Bounded Rationality Curricula
Robustness against adversarial attacks and distribution shifts is a long-standin
g goal of Reinforcement Learning (RL). To this end, Robust Adversarial Reinforce
ment Learning (RARL) trains a protagonist against destabilizing forces exercised
 by an adversary in a competitive zero-sum Markov game, whose optimal solution,
i.e., rational strategy, corresponds to a Nash equilibrium. However, finding Nas
h equilibria requires facing complex saddle point optimization problems, which c
an be prohibitive to solve, especially for high-dimensional control. In this pap
er, we propose a novel approach for adversarial RL based on entropy regularizati
on to ease the complexity of the saddle point optimization problem. We show that
 the solution of this entropy-regularized problem corresponds to a Quantal Respo

nse Equilibrium (QRE), a generalization of Nash equilibria that accounts for bou
nded rationality, i.e., agents sometimes play random actions instead of optimal
ones. Crucially, the connection between the entropy-regularized objective and QR
E enables free modulation of the rationality of the agents by simply tuning the
temperature coefficient. We leverage this insight to propose our novel algorithm
, Quantal Adversarial RL (QARL), which gradually increases the rationality of th
e adversary in a curriculum fashion until it is fully rational, easing the compl
exity of the optimization problem while retaining robustness. We provide extensi
ve evidence of QARL outperforming RARL and recent baselines across several MuJoC
o locomotion and navigation problems in overall performance and robustness.
****
Yuhui Li,Fangyun Wei,Jinjing Zhao,Chao Zhang,Hongyang Zhang
RAIN: Your Language Models Can Align Themselves without Finetuning
Large language models (LLMs) often demonstrate inconsistencies with human prefer
ences. Previous research typically gathered human preference data and then align
ed the pre-trained models using reinforcement learning or instruction tuning, a.
k.a. the finetuning step. In contrast, aligning frozen LLMs without requiring al
ignment data is more appealing. This work explores the potential of the latter s
etting. We discover that by integrating self-evaluation and rewind mechanisms, u
naligned LLMs can directly produce responses consistent with human preferences v
ia self-boosting. We introduce a novel inference method, Rewindable Auto-regress
ive INference (RAIN), that allows pre-trained LLMs to evaluate their own generat
ion and use the evaluation results to guide rewind and generation for AI safety.
 Notably, RAIN operates without the need of extra data for model alignment and a
bstains from any training, gradient computation, or parameter updates. Experimen
tal results evaluated by GPT-4 and humans demonstrate the effectiveness of RAIN:
 on the HH dataset, RAIN improves the harmlessness rate of LLaMA 30B from 82% of
 vanilla inference to 97%, while maintaining the helpfulness rate. On the Truthf
ulQA dataset, RAIN improves the truthfulness of the already-well-aligned LLaMA-2
-chat 13B model by 5%.
****
Minyang Hu,Hong Chang,Bingpeng Ma,Shiguang Shan,Xilin CHEN
Scalable Modular Network: A Framework for Adaptive Learning via Agreement Routin
g
In this paper, we propose a novel modular network framework, called Scalable Mod
ular Network (SMN), which enables adaptive learning capability and supports inte
gration of new modules after pre-training for better adaptation.
This adaptive capability comes from a novel design of router within SMN, named a
greement router, which selects and composes different specialist modules through
 an iterative message passing process.
The agreement router iteratively computes the agreements among a set of input an
d outputs of all modules to allocate inputs to specific module.
During the iterative routing, messages of modules are passed to each other, whic
h improves the module selection process with consideration of both local interac
tions (between a single module and input) and global interactions involving mult
iple other modules.
To validate our contributions, we conduct experiments on two problems: a toy min
-max game and few-shot image classification task.
Our experimental results demonstrate that SMN can generalize to new distribution
s and exhibit sample-efficient adaptation to new tasks.
Furthermore, SMN can achieve a better adaptation capability when new modules are
 introduced after pre-training.
Our code is available at https://github.com/hu-my/ScalableModularNetwork.
****
Keller Jordan
On the Variance of Neural Network Training with respect to Test Sets and Distrib
utions
Neural network trainings are stochastic, causing the performance of trained netw
orks to vary across repeated runs of training.
We contribute the following results towards understanding this variation.

(1) Despite having significant variance on their test-sets, we demonstrate that standard CIFAR-10 and ImageNet trainings have little variance in their performan ce on the test-distributions from which their test-sets are sampled.
(2) We introduce the independent errors assumption and show that it suffices to recover the structure and variance of the empirical accuracy distribution across repeated runs of training.
(3) We prove that test-set variance is unavoidable given the observation that en sembles of identically trained networks are calibrated (Jiang et al., 2021), and demonstrate that the variance of binary classification trainings closely follow s a simple formula based on the error rate and number of test examples.
(4) We conduct preliminary studies of data augmentation, learning rate, finetuni ng instability and distribution-shift through the lens of variance between runs.
****
Max Losch,Mohamed Omran,David Stutz,Mario Fritz,Bernt Schiele
On Adversarial Training without Perturbing all Examples
Adversarial training is the de-facto standard for improving robustness against a dversarial examples. This usually involves a multi-step adversarial attack appli ed on each example during training. In this paper, we explore only constructing adversarial examples (AE) on a subset of the training examples. That is, we spli t the training set in two subsets A and B, train models on both ($A\cup B$) but construct AEs only for examples in A. Starting with A containing only a single class, we systematically increase the size of A and consider splitting by class and by examples. We observe that: (i) adv. robustness transfers by diff iculty and to classes in B that have never been adv. attacked during training, (ii) we observe a tendency for hard examples to provide better robustness trans fer than easy examples, yet find this tendency to diminish with increasing compl exity of datasets (iii) generating AEs on only 50% of training data is suffici ent to recover most of the baseline AT performance even on ImageNet. We observe similar transfer properties across tasks, where generating AEs on only 30% of data can recover baseline robustness on the target task. We evaluate our subset analysis on a wide variety of image datasets like CIFAR-10, CIFAR-100, ImageNet-200 and show transfer to SVHN, Oxford-Flowers-102 and Caltech-256. In contrast t o conventional practice, our experiments indicate that the utility of computing AEs varies by class and examples and that weighting examples from A higher tha n B provides high transfer performance. Code is available at [http://github.co m/mlosch/SAT](http://github.com/mlosch/SAT).
****
Xiangyu Liu,Souradip Chakraborty,Yanchao Sun,Furong Huang
Rethinking Adversarial Policies: A Generalized Attack Formulation and Provable D efense in RL
Most existing works focus on direct perturbations to the victim's state/action o r the underlying transition dynamics to demonstrate the vulnerability of reinfor cement learning agents to adversarial attacks.
However, such direct manipulations may not be always realizable.
In this paper, we consider a multi-agent setting where a well-trained victim age nt ν is exploited by an attacker controlling another
agent α with an \textit{adversarial policy}. Previous models do not accou nt for the possibility that the attacker may only have partial control over
α or that the attack may produce easily detectable ``abnormal'' behaviors . Furthermore, there is a lack of provably efficient defenses against these adve rsarial policies.
To address these limitations, we introduce a generalized attack framework that h as the flexibility to model to what extent the adversary is able to control the agent, and allows the attacker to regulate the state distribution shift and prod uce stealthier adversarial policies. Moreover, we offer a provably efficient def ense with polynomial convergence to the most robust victim policy through advers arial training with timescale separation.
This stands in sharp contrast to supervised learning, where adversarial training typically provides only \textit{empirical} defenses.
Using the Robosumo competition experiments, we show that our generalized attack

formulation results in much stealthier adversarial policies when maintaining the same winning rate as baselines.
Additionally, our adversarial training approach yields stable learning dynamics and less exploitable victim policies.
****
Yihan Wang,Si Si,Daliang Li,Michal Lukasik,Felix Yu,Cho-Jui Hsieh,Inderjit S Dhillon,Sanjiv Kumar
Two-stage LLM Fine-tuning with Less Specialization and More Generalization
Pretrained large language models (LLMs) are general purpose problem solvers applicable to a diverse set of tasks with prompts. They can be further improved towards a specific task by fine-tuning on a specialized dataset. However, fine-tuning usually makes the model narrowly specialized on this dataset with reduced general in-context learning performances, which is undesirable whenever the fine-tuned model needs to handle additional tasks where no fine-tuning data is available.
In this work, we first demonstrate that fine-tuning on a single task indeed decreases LLMs' general in-context learning performance. We discover one important cause of such forgetting, format specialization, where the model overfits to the format of the fine-tuned task.We further show that format specialization happens at the very beginning of fine-tuning. To solve this problem, we propose Prompt Tuning with MOdel Tuning (ProMoT), a simple yet effective two-stage fine-tuning framework that reduces format specialization and improves generalization.ProMoT offloads task-specific format learning into additional and removable parameters by first doing prompt tuning and then fine-tuning the model itself with this soft prompt attached.
With experiments on several fine-tuning tasks and 8 in-context evaluation tasks, we show that ProMoT achieves comparable performance on fine-tuned tasks to standard fine-tuning, but with much less loss of in-context learning performances across a board range of out-of-domain evaluation tasks. More importantly, ProMoT can even enhance generalization on in-context learning tasks that are semantically related to the fine-tuned task, e.g. ProMoT on En-Fr translation significantly improves performance on other language pairs, and ProMoT on NLI improves performance on summarization.
Experiments also show that ProMoT can improve the generalization performance of multi-task training.
****
Salar Abbaspourazad,Oussama Elachqar,Andrew Miller,Saba Emrani,Udhyakumar Nallasamy,Ian Shapiro
Large-scale Training of Foundation Models for Wearable Biosignals
Tracking biosignals is crucial for monitoring wellness and preempting the development of severe medical conditions. Today, wearable devices can conveniently record various biosignals, creating the opportunity to monitor health status without disruption to one's daily routine. Despite widespread use of wearable devices and existing digital biomarkers, the absence of curated data with annotated medical labels hinders the development of new biomarkers to measure common health conditions. In fact, medical datasets are usually small in comparison to other domains, which is an obstacle for developing neural network models for biosignals.
To address this challenge, we have employed self-supervised learning using the unlabeled sensor data collected under informed consent from the large longitudinal Apple Heart and Movement Study (AHMS) to train foundation models for two common biosignals: photoplethysmography (PPG) and electrocardiogram (ECG) recorded on Apple Watch. We curated PPG and ECG datasets from AHMS that include data from ${\sim} 141$K participants spanning ${\sim} 3$ years. Our self-supervised learning framework includes participant level positive pair selection, stochastic augmentation module and a regularized contrastive loss optimized with momentum training, and generalizes well to both PPG and ECG modalities. We show that the pre-trained foundation models readily encode information regarding participants' demographics and health conditions. To the best of our knowledge, this is the first study that builds foundation models using large-scale PPG and ECG data collected via wearable consumer devices $\textendash$ prior works have commonly used small

er-size datasets collected in clinical and experimental settings. We believe PPG and ECG foundation models can enhance future wearable devices by reducing the reliance on labeled data and hold the potential to help the users improve their health.
****
Avinash Kori,Francesco Locatello,Fabio De Sousa Ribeiro,Francesca Toni,Ben Glocker
Grounded Object-Centric Learning
The extraction of object-centric representations for downstream tasks is an emerging area of research. Learning grounded representations of objects that are guaranteed to be stable and invariant promises robust performance across different tasks and environments. Slot Attention (SA) learns object-centric representations by assigning objects to *slots*, but presupposes a *single* distribution from which all slots are randomly initialised. This results in an inability to learn *specialized* slots which bind to specific object types and remain invariant to identity-preserving changes in object appearance. To address this, we present *Conditional Slot Attention* (CoSA) using a novel concept of *Grounded Slot Dictionary* (GSD) inspired by vector quantization. Our proposed GSD comprises (i) canonical object-level property vectors and (ii) parametric Gaussian distributions, which define a prior over the slots. We demonstrate the benefits of our method in multiple downstream tasks such as scene generation, composition, and task adaptation, whilst remaining competitive with SA in object discovery.
****
Hao Wang,Chenyi Zhang,Tongyang Li
Near-Optimal Quantum Algorithm for Minimizing the Maximal Loss
The problem of minimizing the maximum of N convex, Lipschitz functions plays significant roles in optimization and machine learning. It has a series of results, with the most recent one requiring $O(N\epsilon^{-2/3} + \epsilon^{-8/3})$ queries to a first-order oracle to compute an ϵ-suboptimal point. On the other hand, quantum algorithms for optimization are rapidly advancing with speedups shown on many important optimization problems. In this paper, we conduct a systematic study of quantum algorithms and lower bounds for minimizing the maximum of N convex, Lipschitz functions. On one hand, we develop quantum algorithms with an improved complexity bound of $\tilde{O}(\sqrt{N}\epsilon^{-5/3} + \epsilon^{-8/3})$. On the other hand, we prove that quantum algorithms must take $\tilde{\Omega}(\sqrt{N}\epsilon^{-2/3})$ queries to a first-order quantum oracle, showing that our dependence on N is optimal up to poly-logarithmic factors.
****
Seunghan Lee,Taeyoung Park,Kibok Lee
Soft Contrastive Learning for Time Series
Contrastive learning has shown to be effective to learn representations from time series in a self-supervised way.
However, contrasting similar time series instances or values from adjacent timestamps within a time series leads to ignore their inherent correlations, which results in deteriorating the quality of learned representations.
To address this issue, we propose \textit{SoftCLT}, a simple yet effective soft contrastive learning strategy for time series.
This is achieved by introducing instance-wise and temporal contrastive loss with soft assignments ranging from zero to one.
Specifically, we define soft assignments for 1) instance-wise contrastive loss by distance between time series on the data space, warping and 2) temporal contrastive loss by the difference of timestamps.
SoftCLT is a plug-and-play method for time series contrastive learning that improves the quality of learned representations without bells and whistles.
In experiments, we demonstrate that SoftCLT consistently improves the performance in various downstream tasks including classification, semi-supervised learning, transfer learning, and anomaly detection, showing state-of-the-art performance.
Code is available at this repository: https://github.com/seunghan96/softclt.
****

Ahmed Abdulaal,adamos hadjivasiliou,Nina Montana-Brown,Tiantian He,Ayodeji Ijishakin,Ivana Drobnjak,Daniel C. Castro,Daniel C. Alexander
Causal Modelling Agents: Causal Graph Discovery through Synergising Metadata- and Data-driven Reasoning
Scientific discovery hinges on the effective integration of metadata, which refers to a set of 'cognitive' operations such as determining what information is relevant for inquiry, and data, which encompasses physical operations such as observation and experimentation. This paper introduces the Causal Modelling Agent (CMA), a novel framework that synergizes the metadata-based reasoning capabilities of Large Language Models (LLMs) with the data-driven modelling of Deep Structural Causal Models (DSCMs) for the task of causal discovery. We evaluate the CMA's performance on a number of benchmarks, as well as on the real-world task of modelling the clinical and radiological phenotype of Alzheimer's Disease (AD). Our experimental results indicate that the CMA can outperform previous data-driven or metadata-driven approaches to causal discovery. In our real-world application, we use the CMA to derive new insights into the causal relationships among biomarkers of AD.
****
Changwoo Lee,Hun-Seok Kim
Differentiable Learning of Generalized Structured Matrices for Efficient Deep Neural Networks
This paper investigates efficient deep neural networks (DNNs) to replace dense unstructured weight matrices with structured ones that possess desired properties. The challenge arises because the optimal weight matrix structure in popular neural network models is obscure in most cases and may vary from layer to layer even in the same network. Prior structured matrices proposed for efficient DNNs were mostly hand-crafted without a generalized framework to systematically learn them. To address this issue, we propose a generalized and differentiable framework to learn efficient structures of weight matrices by gradient descent. We first define a new class of structured matrices that covers a wide range of structured matrices in the literature by adjusting the structural parameters. Then, the frequency-domain differentiable parameterization scheme based on the Gaussian-Dirichlet kernel is adopted to learn the structural parameters by proximal gradient descent. On the image and language tasks, our method learns efficient DNNs with structured matrices, achieving lower complexity and/or higher performance than prior approaches that employ low-rank, block-sparse, or block-low-rank matrices.
****
Jiechao Guan,Hui Xiong
Improved Regret Bounds for Non-Convex Online-Within-Online Meta Learning
Online-Within-Online (OWO) meta learning stands for the online multi-task learning paradigm in which both tasks and data within each task become available in a sequential order. In this work, we study the OWO meta learning of the initialization and step size of within-task online algorithms in the non-convex setting, and provide improved regret bounds under mild assumptions of loss functions. Previous work analyzing this scenario has obtained for bounded and piecewise Lipschitz functions an averaged regret bound $O((\frac{\sqrt{m}}{T^{1/4}}+\frac{(\log{m})\log{T}}{\sqrt{T}}+V)\sqrt{m})$ across T tasks, with m iterations per task and V the task similarity. Our first contribution is to modify the existing non-convex OWO meta learning algorithm and improve the regret bound to $O((\frac{1}{T^{1/2-\alpha}}+\frac{(\log{T})^{9/2}}{T}+V)\sqrt{m})$, for any $\alpha \in (0,1/2)$. The derived bound has a faster convergence rate with respect to T, and guarantees a vanishing task-averaged regret with respect to m (for any fixed T). Then, we propose a new algorithm of regret $O((\frac{\log{T}}{T}+V)\sqrt{m})$ for non-convex OWO meta learning. This regret bound exhibits a better asymptotic performance than previous ones, and holds for any bounded (not necessarily Lipschitz) loss functions. Besides the improved regret bounds, our contributions include investigating how to attain generalization bounds for statistical meta learning via regret analysis. Specifically, by online-to-batch arguments, we achieve a transfer risk bound for batch meta learning that assumes all tasks are drawn from a distribution. Moreover, by connecting multi-task generalization erro

r with task-averaged regret, we develop for statistical multi-task learning a no vel PAC-Bayes generalization error bound that involves our regret bound for OWO meta learning.
****
Aadirupa Saha,Branislav Kveton
Only Pay for What Is Uncertain: Variance-Adaptive Thompson Sampling
Most bandit algorithms assume that the reward variances or their upper bounds ar e known, and that they are the same for all arms. This naturally leads to subopt imal performance and higher regret due to variance overestimation. On the other hand, underestimated reward variances may lead to linear regret due to committin g early to a suboptimal arm. This motivated prior works on variance-adaptive fre quentist algorithms, which have strong instance-dependent regret bounds but cann ot incorporate prior knowledge on reward variances. We lay foundations for the B ayesian setting, which incorporates prior knowledge. This results in lower regre t in practice, due to using the prior in the algorithm design, and also improved regret guarantees. Specifically, we study Gaussian bandits with \emph{unknown h eterogeneous reward variances}, and develop a Thompson sampling algorithm with p rior-dependent Bayes regret bounds. We achieve lower regret with lower reward va riances and more informative priors on them, which is precisely why we pay only for what is uncertain. This is the first result of its kind. Finally, we corrobo rate our theory with extensive experiments, which show the superiority of our va riance-adaptive Bayesian algorithm over prior frequentist approaches. We also sh ow that our approach is robust to model misspecification and can be applied with estimated priors.
****
Jonghyun Lee,Hansam Cho,YoungJoon Yoo,Seoung Bum Kim,Yonghyun Jeong
Compose and Conquer: Diffusion-Based 3D Depth Aware Composable Image Synthesis
Addressing the limitations of text as a source of accurate layout representation in text-conditional diffusion models, many works incorporate additional signals to condition certain attributes within a generated image. Although successful, previous works do not account for the specific localization of said attributes e xtended into the three dimensional plane. In this context, we present a conditio nal diffusion model that integrates control over three-dimensional object placem ent with disentangled representations of global stylistic semantics from multipl e exemplar images. Specifically, we first introduce depth disentanglement traini ng to leverage the relative depth of objects as an estimator, allowing the model to identify the absolute positions of unseen objects through the use of synthet ic image triplets. We also introduce soft guidance, a method for imposing global semantics onto targeted regions without the use of any additional localization cues. Our integrated framework, Compose and Conquer (CnC), unifies these techniq ues to localize multiple conditions in a disentangled manner. We demonstrate tha t our approach allows perception of objects at varying depths while offering a v ersatile framework for composing localized objects with different global semanti cs.
****
T. Anderson Keller,Lyle Muller,Terrence Sejnowski,Max Welling
Traveling Waves Encode The Recent Past and Enhance Sequence Learning
Traveling waves of neural activity have been observed throughout the brain at a diversity of regions and scales; however, their precise computational role is st ill debated. One physically inspired hypothesis suggests that the cortical sheet may act like a wave-propagating system capable of invertibly storing a short-te rm memory of sequential stimuli through induced waves traveling across the corti cal surface, and indeed many experimental results from neuroscience correlate wa ve activity with memory tasks. To date, however, the computational implications of this idea have remained hypothetical due to the lack of a simple recurrent ne ural network architecture capable of exhibiting such waves. In this work, we int roduce a model to fill this gap, which we denote the Wave-RNN (wRNN), and demons trate how such an architecture indeed efficiently encodes the recent past throug h a suite of synthetic memory tasks where wRNNs learn faster and reach significa ntly lower error than wave-free counterparts. We further explore the implication

s of this memory storage system on more complex sequence modeling tasks such as sequential image classification and find that wave-based models not only again outperform comparable wave-free RNNs while using significantly fewer parameters, but additionally perform comparably to more complex gated architectures such as LSTMs and GRUs.
****
Mircea Mironenco,Patrick Forré
Lie Group Decompositions for Equivariant Neural Networks
Invariance and equivariance to geometrical transformations have proven to be very useful inductive biases when training (convolutional) neural network models, especially in the low-data regime.
Much work has focused on the case where the symmetry group employed is compact or abelian, or both.
Recent work has explored enlarging the class of transformations used to the case of Lie groups, principally through the use of their Lie algebra, as well as the group exponential and logarithm maps.
The applicability of such methods to larger transformation groups is limited by the fact that depending on the group of interest G, the exponential map may not be surjective.
Further limitations are encountered when G is neither compact nor abelian.
Using the structure and geometry of Lie groups and their homogeneous spaces, we present a framework by which it is possible to work with such groups primarily focusing on the Lie groups $G = \textnormal{GL}^{+}(n, \mathbb{R})$ and $G = \textnormal{SL}(n, \mathbb{R})$, as well as their representation as affine transformations $\mathbb{R}^{n} \rtimes G$.
Invariant integration as well as a global parametrization is realized by decomposing the "larger" groups into subgroups and submanifolds which can be handled individually.
Under this framework, we show how convolution kernels can be parametrized to build models equivariant with respect to affine transformations.
We evaluate the robustness and out-of-distribution generalisation capability of our model on the standard affine-invariant benchmark classification task, where we outperform all previous equivariant models as well as all Capsule Network proposals.
****
Pratyusha Sharma,Jordan T. Ash,Dipendra Misra
The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction
Transformer-based Large Language Models (LLMs) have become a fixture in modern machine learning. Correspondingly, significant resources are allocated towards research that aims to further advance this technology, typically resulting in models of increasing size that are trained on increasing amounts of data. This work, however, demonstrates the surprising result that it is often possible to significantly improve the performance of LLMs by selectively removing higher-order components of their weight matrices. This simple intervention, which we call LAyer-SElective Rank reduction (LASER), can be done on a model after training has completed, and requires minimal additional parameters and data. We show extensive experiments demonstrating the generality of this finding across language models and datasets, and provide in-depth analyses offering insights into both when LASER is effective and the mechanism by which it operates
****
Weiyu Sun,Xinyu Zhang,Hao LU,Ying-Cong Chen,Ting Wang,Jinghui Chen,Lu Lin
Backdoor Contrastive Learning via Bi-level Trigger Optimization
Contrastive Learning (CL) has attracted enormous attention due to its remarkable capability in unsupervised representation learning. However, recent works have revealed the vulnerability of CL to backdoor attacks: the feature extractor could be misled to embed backdoored data close to an attack target class, thus fooling the downstream predictor to misclassify it as the target. Existing attacks usually adopt a fixed trigger pattern and poison the training set with trigger-injected data, hoping for the feature extractor to learn the association between tr

igger and target class. However, we find that such fixed trigger design fails to effectively associate trigger-injected data with target class in the embedding space due to special CL mechanisms, leading to a limited attack success rate (ASR). This phenomenon motivates us to find a better backdoor trigger design tailored for CL framework. In this paper, we propose a bi-level optimization approach to achieve this goal, where the inner optimization simulates the CL dynamics of a surrogate victim, and the outer optimization enforces the backdoor trigger to stay close to the target throughout the surrogate CL procedure. Extensive experiments show that our attack can achieve a higher attack success rate (e.g., 99\% ASR on ImageNet-100) with a very low poisoning rate (1\%). Besides, our attack can effectively evade existing state-of-the-art defenses.
****

Robin van de Water,Hendrik Nils Aurel Schmidt,Paul Elbers,Patrick Thoral,Bert Arnrich,Patrick Rockenschaub

Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML

Medical applications of machine learning (ML) have experienced a surge in popularity in recent years. Given the abundance of available data from electronic health records, the intensive care unit (ICU) is a natural habitat for ML. Models have been proposed to address numerous ICU prediction tasks like the early detection of complications. While authors frequently report state-of-the-art performance, it is challenging to verify claims of superiority. Datasets and code are not always published, and cohort definitions, preprocessing pipelines, and training setups are difficult to reproduce. This work introduces Yet Another ICU Benchmark (YAIB), a modular framework that allows researchers to define reproducible and comparable clinical ML experiments; we offer an end-to-end solution from cohort definition to model evaluation. The framework natively supports most open-access ICU datasets (MIMIC III/IV, eICU, HiRID, AUMCdb) and is easily adaptable to future ICU datasets. Combined with a transparent preprocessing pipeline and extensible training code for multiple ML and deep learning models, YAIB enables unified model development, transfer, and evaluation. Our benchmark comes with five predefined established prediction tasks (mortality, acute kidney injury, sepsis, kidney function, and length of stay) developed in collaboration with clinicians. Adding further tasks is straightforward by design. Using YAIB, we demonstrate that the choice of dataset, cohort definition, and preprocessing have a major impact on the prediction performance — often more so than model class — indicating an urgent need for YAIB as a holistic benchmarking tool. We provide our work to the clinical ML community to accelerate method development and enable real-world clinical implementations.
****

Zheng Chen,Yulun Zhang,Jinjin Gu,Linghe Kong,Xiaokang Yang

Recursive Generalization Transformer for Image Super-Resolution

Transformer architectures have exhibited remarkable performance in image super-resolution (SR). Since the quadratic computational complexity of the self-attention (SA) in Transformer, existing methods tend to adopt SA in a local region to reduce overheads. However, the local design restricts the global context exploitation, which is crucial for accurate image reconstruction. In this work, we propose the Recursive Generalization Transformer (RGT) for image SR, which can capture global spatial information and is suitable for high-resolution images. Specifically, we propose the recursive-generalization self-attention (RG-SA). It recursively aggregates input features into representative feature maps, and then utilizes cross-attention to extract global information. Meanwhile, the channel dimensions of attention matrices ($query$, key, and $value$) are further scaled to mitigate the redundancy in the channel domain. Furthermore, we combine the RG-SA with local self-attention to enhance the exploitation of the global context, and propose the hybrid adaptive integration (HAI) for module integration. The HAI allows the direct and effective fusion between features at different levels (local or global). Extensive experiments demonstrate that our RGT outperforms recent state-of-the-art methods quantitatively and qualitatively. Code and pre-trained models are available at https://github.com/zhengchen1999/RGT.
****

Herbie Bradley,Andrew Dai,Hannah Benita Teufel,Jenny Zhang,Koen Oostermeijer,Mar
co Bellagente,Jeff Clune,Kenneth Stanley,Gregory Schott,Joel Lehman
Quality-Diversity through AI Feedback
In many text-generation problems, users may prefer not only a single response, b
ut a diverse range of high-quality outputs from which to choose. Quality-diversi
ty (QD) search algorithms aim at such outcomes, by continually improving and div
ersifying a population of candidates. However, the applicability of QD to qualit
ative domains, like creative writing, has been limited by the difficulty of algo
rithmically specifying measures of quality and diversity. Interestingly, recent
developments in language models (LMs) have enabled guiding search through \emph{
AI feedback}, wherein LMs are prompted in natural language to evaluate qualitati
ve aspects of text. Leveraging this development, we introduce Quality-Diversity
through AI Feedback (QDAIF), wherein an evolutionary algorithm applies LMs to bo
th generate variation and evaluate the quality and diversity of candidate text.
When assessed on creative writing domains, QDAIF covers more of a specified sear
ch space with high-quality samples than do non-QD controls. Further, human evalu
ation of QDAIF-generated creative texts validates reasonable agreement between A
I and human evaluation. Our results thus highlight the potential of AI feedback
to guide open-ended search for creative and original solutions, providing a reci
pe that seemingly generalizes to many domains and modalities. In this way, QDAIF
 is a step towards AI systems that can independently search, diversify, evaluate
, and improve, which are among the core skills underlying human society's capaci
ty for innovation.
****
Artem Agafonov,Dmitry Kamzolov,Alexander Gasnikov,Ali Kavis,Kimon Antonakopoulos
,Volkan Cevher,Martin Taká■
Advancing the Lower Bounds: an Accelerated, Stochastic, Second-order Method with
 Optimal Adaptation to Inexactness
We present a new accelerated stochastic second-order method that is robust to bo
th gradient and Hessian inexactness, typical in machine learning. We establish t
heoretical lower bounds and prove that our algorithm achieves optimal convergenc
e in both gradient and Hessian inexactness in this key setting. We further intro
duce a tensor generalization for stochastic higher-order derivatives. When the o
racles are non-stochastic, the proposed tensor algorithm matches the global conv
ergence of Nesterov Accelerated Tensor method. Both algorithms allow for approxi
mate solutions of their auxiliary subproblems with verifiable conditions on the
accuracy of the solution.
****
Yilan Zhang,Yingxue Xu,Jianqi Chen,Fengying Xie,Hao Chen
Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer S
urvival Prediction
Multimodal learning significantly benefits cancer survival prediction, especiall
y the integration of pathological images and genomic data. Despite advantages of
 multimodal learning for cancer survival prediction, massive redundancy in multi
modal data prevents it from extracting discriminative and compact information: (
1) An extensive amount of intra-modal task-unrelated information blurs discrimin
ability, especially for gigapixel whole slide images (WSIs) with many patches in
 pathology and thousands of pathways in genomic data, leading to an "intra-modal
 redundancy" issue. (2) Duplicated information among modalities dominates the re
presentation of multimodal data, which makes modality-specific information prone
 to being ignored, resulting in an "inter-modal redundancy" issue. To address th
ese, we propose a new framework, Prototypical Information Bottlenecking and Dise
ntangling (PIBD), consisting of Prototypical Information Bottleneck (PIB) module
 for intra-modal redundancy and Prototypical Information Disentanglement (PID) m
odule for inter-modal redundancy. Specifically, a variant of information bottlen
eck, PIB, is proposed to model prototypes approximating a bunch of instances for
 different risk levels, which can be used for selection of discriminative instan
ces within modality. PID module decouples entangled multimodal data into compact
 distinct components: modality-common and modality-specific knowledge, under the
 guidance of the joint prototypical distribution. Extensive experiments on five

cancer benchmark datasets demonstrated our superiority over other methods. The code is released.
****
Seyed Iman Mirzadeh,Keivan Alizadeh-Vahid,Sachin Mehta,Carlo C del Mundo,Oncel Tuzel,Golnoosh Samei,Mohammad Rastegari,Mehrdad Farajtabar
ReLU Strikes Back: Exploiting Activation Sparsity in Large Language Models
Large Language Models (LLMs) with billions of parameters have drastically transformed AI applications. However, their demanding computation during inference has raised significant challenges for deployment on resource-constrained devices. Despite recent trends favoring alternative activation functions such as GELU or SiLU, known for increased computation, this study strongly advocates for reinstating ReLU activation in LLMs. We demonstrate that using the ReLU activation function has a negligible impact on convergence and performance while significantly reducing computation and weight transfer. This reduction is particularly valuable during the memory-bound inference step, where efficiency is paramount. Exploring sparsity patterns in ReLU-based LLMs, we unveil the reutilization of activated neurons for generating new tokens and leveraging these insights, we propose practical strategies to substantially reduce LLM inference computation up to three times, using ReLU activations with minimal performance trade-offs.
****
Ziheng Chen,Yue Song,Yunmei Liu,Nicu Sebe
A Lie Group Approach to Riemannian Batch Normalization
Manifold-valued measurements exist in numerous applications within computer vision and machine learning. Recent studies have extended Deep Neural Networks (DNNs) to manifolds, and concomitantly, normalization techniques have also been adapted to several manifolds, referred to as Riemannian normalization. Nonetheless, most of the existing Riemannian normalization methods have been derived in an ad hoc manner and only apply to specific manifolds. This paper establishes a unified framework for Riemannian Batch Normalization (RBN) techniques on Lie groups. Our framework offers the theoretical guarantee of controlling both the Riemannian mean and variance. Empirically, we focus on Symmetric Positive Definite (SPD) manifolds, which possess three distinct types of Lie group structures. Using the deformation concept, we generalize the existing Lie groups on SPD manifolds into three families of parameterized Lie groups. Specific normalization layers induced by these Lie groups are then proposed for SPD neural networks. We demonstrate the effectiveness of our approach through three sets of experiments: radar recognition, human action recognition, and electroencephalography (EEG) classification. The code is available at https://github.com/GitZH-Chen/LieBN.git.
****
Sophia Huiwen Sun,Rose Yu
Copula Conformal prediction for multi-step time series prediction
Accurate uncertainty measurement is a key step in building robust and reliable machine learning systems. Conformal prediction is a distribution-free uncertainty quantification framework popular for its ease of implementation, finite-sample coverage guarantees, and generality for underlying prediction algorithms. However, existing conformal prediction approaches for time series are limited to single-step prediction without considering the temporal dependency. In this paper, we propose the Copula Conformal Prediction algorithm for multivariate, multi-step Time Series forecasting, CopulaCPTS. We prove that CopulaCPTS has finite-sample validity guarantee. On four synthetic and real-world multivariate time series datasets, we show that CopulaCPTS produces more calibrated and efficient confidence intervals for multi-step prediction tasks than existing techniques. Our code is open-sourced at https://github.com/Rose-STL-Lab/CopulaCPTS.
****
Jaehyeon Kim,Keon Lee,Seungjun Chung,Jaewoong Cho
CLaM-TTS: Improving Neural Codec Language Model for Zero-Shot Text-to-Speech
With the emergence of neural audio codecs, which encode multiple streams of discrete tokens from audio, large language models have recently gained attention as a promising approach for zero-shot Text-to-Speech (TTS) synthesis. Despite the ongoing rush towards scaling paradigms, audio tokenization ironically amplifies t

he scalability challenge, stemming from its long sequence length and the complex ity of modelling the multiple sequences. To mitigate these issues, we present CL aM-TTS that employs a probabilistic residual vector quantization to (1) achieve superior compression in the token length, and (2) allow a language model to gene rate multiple tokens at once, thereby eliminating the need for cascaded modeling to handle the number of token streams. Our experimental results demonstrate tha t CLaM-TTS is better than or comparable to state-of-the-art neural codec-based T TS models regarding naturalness, intelligibility, speaker similarity, and infere nce speed. In addition, we examine the impact of the pretraining extent of the l anguage models and their text tokenization strategies on performances.
****
Yuexiao Ma,Huixia Li,Xiawu Zheng,Feng Ling,Xuefeng Xiao,Rui Wang,Shilei Wen,Fei Chao,Rongrong Ji
AffineQuant: Affine Transformation Quantization for Large Language Models
The significant resource requirements associated with Large-scale Language Model s (LLMs) have generated considerable interest in the development of techniques a imed at compressing and accelerating neural networks.
Among these techniques, Post-Training Quantization (PTQ) has emerged as a subjec t of considerable interest due to its noteworthy compression efficiency and cost -effectiveness in the context of training.
Existing PTQ methods for LLMs limit the optimization scope to scaling transforma tions between pre- and post-quantization weights.
This constraint results in significant errors after quantization, particularly i n low-bit configurations.
In this paper, we advocate for the direct optimization using equivalent Affine t ransformations in PTQ (AffineQuant).
This approach extends the optimization scope and thus significantly minimizing q uantization errors.
Additionally, by employing the corresponding inverse matrix, we can ensure equiv alence between the pre- and post-quantization outputs of PTQ, thereby maintainin g its efficiency and generalization capabilities.
To ensure the invertibility of the transformation during optimization, we furthe r introduce a gradual mask optimization method.
This method initially focuses on optimizing the diagonal elements and gradually extends to the other elements.
Such an approach aligns with the Levy-Desplanques theorem, theoretically ensurin g invertibility of the transformation.
As a result, significant performance improvements are evident across different L LMs on diverse datasets.
Notably, these improvements are most pronounced when using very low-bit quantiza tion, enabling the deployment of large models on edge devices.
To illustrate, we attain a C4 perplexity of 15.76 (2.26$\downarrow$ vs 18.02 in OmniQuant) on the LLaMA2-7B model of W4A4 quantization without overhea d.
On zero-shot tasks, AffineQuant achieves an average of 58.61% accuracy ($1.9 8\%\uparrow$ vs 56.63 in OmniQuant) when using 4/4-bit quantization for LL aMA-30B, which setting a new state-of-the-art benchmark for PTQ in LLMs.
Codes are available at: https://github.com/bytedance/AffineQuant.
****
Mehdi Zadem,Sergio Mover,Sao Mai Nguyen
Reconciling Spatial and Temporal Abstractions for Goal Representation
Goal representation affects the performance of Hierarchical Reinforcement Learn- ing (HRL) algorithms by decomposing the complex learning problem into easier su btasks. Recent studies show that representations that preserve temporally ab- st ract environment dynamics are successful in solving difficult problems and pro- vide theoretical guarantees for optimality. These methods however cannot scale t o tasks where environment dynamics increase in complexity i.e. the temporally ab stract transition relations depend on larger number of variables. On the other h and, other efforts have tried to use spatial abstraction to mitigate the previou s issues. Their limitations include scalability to high dimensional environments

and dependency on prior knowledge.

In this paper, we propose a novel three-layer HRL algorithm that introduces, at different levels of the hierarchy, both a spatial and a temporal goal abstraction. We provide a theoretical study of the regret bounds of the learned policies. We evaluate the approach on complex continuous control tasks, demonstrating the effectiveness of spatial and temporal abstractions learned by this approach.
****
Hanxun Huang,Ricardo J. G. B. Campello,Sarah Monazam Erfani,Xingjun Ma,Michael E. Houle,James Bailey
LDReg: Local Dimensionality Regularized Self-Supervised Learning
Representations learned via self-supervised learning (SSL) can be susceptible to dimensional collapse, where the learned representation subspace is of extremely low dimensionality and thus fails to represent the full data distribution and modalities.
Dimensional collapse --- also known as the "underfilling" phenomenon --- is one of the major causes of degraded performance on downstream tasks. Previous work has investigated the dimensional collapse problem of SSL at a global level. In this paper, we demonstrate that representations can span over high dimensional space globally, but collapse locally. To address this, we propose a method called *local dimensionality regularization (LDReg)*. Our formulation is based on the derivation of the Fisher-Rao metric to compare and optimize local distance distributions at an asymptotically small radius for each data point. By increasing the local intrinsic dimensionality, we demonstrate through a range of experiments that LDReg improves the representation quality of SSL. The results also show that LDReg can regularize dimensionality at both local and global levels.
****
Xinyu Tang,Richard Shin,Huseyin A Inan,Andre Manoel,Fatemehsadat Mireshghallah,Zinan Lin,Sivakanth Gopi,Janardhan Kulkarni,Robert Sim
Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation
We study the problem of in-context learning (ICL) with large language models (LLMs) on private datasets.
This scenario poses privacy risks, as LLMs may leak or regurgitate the private examples demonstrated in the prompt.
We propose a novel algorithm that generates synthetic few-shot demonstrations from the private dataset with formal differential privacy (DP) guarantees, and show empirically that it can achieve effective ICL.
We conduct extensive experiments on standard benchmarks and compare our algorithm with non-private ICL and zero-shot solutions.
Our results demonstrate that our algorithm can achieve competitive performance with strong privacy levels.
These results open up new possibilities for ICL with privacy protection for a broad range of applications.
****
Yiyang Zhou,Chenhang Cui,Jaehong Yoon,Linjun Zhang,Zhun Deng,Chelsea Finn,Mohit Bansal,Huaxiu Yao
Analyzing and Mitigating Object Hallucination in Large Vision-Language Models
Large vision-language models (LVLMs) have shown remarkable abilities in understanding visual information with human languages. However, LVLMs still suffer from object hallucination, which is the problem of generating descriptions that include objects that do not actually exist in the images. This can negatively impact many vision-language tasks, such as visual summarization and reasoning. To address this issue, we propose a simple yet powerful algorithm, LVLM Hallucination Revisor (LURE), to post-hoc rectify object hallucination in LVLMs by reconstructing less hallucinatory descriptions. LURE is grounded in a rigorous statistical analysis of the key factors underlying object hallucination, including co-occurrence (the frequent appearance of certain objects alongside others in images), uncertainty (objects with higher uncertainty during LVLM decoding), and object position (hallucination often appears in the later part of the generated text). LURE

can also be seamlessly integrated with any LVLMs. We evaluate LURE on six open-s
ource LVLMs and found it outperforms the previous best approach in both general
object hallucination evaluation metrics, GPT, and human evaluations.
****
Maciej Miku■a,Szymon Tworkowski,Szymon Antoniak,Bartosz Piotrowski,Albert Q. Jia
ng,Jin Peng Zhou,Christian Szegedy,■ukasz Kuci■ski,Piotr Mi■o■,Yuhuai Wu
Magnushammer: A Transformer-Based Approach to Premise Selection
This paper presents a novel approach to premise selection, a crucial reasoning t
ask in automated theorem proving. Traditionally, symbolic methods that rely on e
xtensive domain knowledge and engineering effort are applied to this task. In co
ntrast, this work demonstrates that contrastive training with the transformer ar
chitecture can achieve higher-quality retrieval of relevant premises, without th
e knowledge or feature engineering overhead. Our method, Magnushammer, outperfor
ms the most advanced and widely used automation tool in interactive theorem prov
ing called Sledgehammer. On the PISA and miniF2f benchmarks Magnushammer achieve
s 59.5% (against 38.3%) and 34.0% (against 20.9%) success rates, res
pectively. By combining Magnushammer with a language-model-based automated theor
em prover, we further improve the state-of-the-art proof success rate from $57.0
\%$ to 71.0% on the PISA benchmark using 4x fewer parameters. Moreover, we
develop and open source a novel dataset for premise selection, containing textua
l representations of (proof state, relevant premise) pairs. To the best of our k
nowledge, this is the largest available premise selection dataset, and the first
 dataset of this kind for the Isabelle proof assistant.
****
Harshit Sikchi,Rohan Chitnis,Ahmed Touati,Alborz Geramifard,Amy Zhang,Scott Niek
um
Score Models for Offline Goal-Conditioned Reinforcement Learning
Offline Goal-Conditioned Reinforcement Learning (GCRL) is tasked with learning t
o achieve multiple goals in an environment purely from offline datasets using sp
arse reward functions. Offline GCRL is pivotal for developing generalist agents
capable of leveraging pre-existing datasets to learn diverse and reusable skills
 without hand-engineering reward functions. However, contemporary approaches to
GCRL based on supervised learning and contrastive learning are often suboptimal
in the offline setting. An alternative perspective on GCRL optimizes for occupan
cy matching, but necessitates learning a discriminator, which subsequently serve
s as a pseudo-reward for downstream RL. Inaccuracies in the learned discriminato
r can cascade, negatively influencing the resulting policy. We present a novel a
pproach to GCRL under a new lens of mixture-distribution matching, leading to ou
r discriminator-free method: SMORe. The key insight is combining the occupancy m
atching perspective of GCRL with a convex dual formulation to derive a learning
objective that can better leverage suboptimal offline data. SMORe learns *scores
* or unnormalized densities representing the importance of taking an action at a
 state for reaching a particular goal. SMORe is principled and our extensive exp
eriments on the fully offline GCRL benchmark composed of robot manipulation and
locomotion tasks, including high-dimensional observations, show that SMORe can o
utperform state-of-the-art baselines by a significant margin.
****
Bowen Cao,Deng Cai,Leyang Cui,Xuxin Cheng,Wei Bi,Yuexian Zou,Shuming Shi
Retrieval is Accurate Generation
Standard language models generate text by selecting tokens from a fixed, finite,
 and standalone vocabulary. We introduce a novel method that selects context-awa
re phrases from a collection of supporting documents. One of the most significan
t challenges for this paradigm shift is determining the training oracles, becaus
e a string of text can be segmented in various ways and each segment can be retr
ieved from numerous possible documents. To address this, we propose to initializ
e the training oracles using linguistic heuristics and, more importantly, bootst
rap the oracles through iterative self-reinforcement. Extensive experiments show
 that our model not only outperforms standard language models on a variety of kn
owledge-intensive tasks but also demonstrates improved generation quality in ope
n-ended text generation. For instance, compared to the standard language model c

ounterpart, our model raises the accuracy from 23.47% to 36.27% on OpenbookQA, and improves the MAUVE score from 42.61% to 81.58% in open-ended text generation. Remarkably, our model also achieves the best performance and the lowest latency among several retrieval-augmented baselines. In conclusion, we assert that retrieval is more accurate generation and hope that our work will encourage further research on this new paradigm shift.
****
Kensen Shi,Joey Hong,Yinlin Deng,Pengcheng Yin,Manzil Zaheer,Charles Sutton
ExeDec: Execution Decomposition for Compositional Generalization in Neural Program Synthesis
When writing programs, people have the ability to tackle a new complex task by decomposing it into smaller and more familiar subtasks. While it is difficult to measure whether neural program synthesis methods have similar capabilities, we can measure whether they compositionally generalize, that is, whether a model that has been trained on the simpler subtasks is subsequently able to solve more complex tasks. In this paper, we characterize several different forms of compositional generalization that are desirable in program synthesis, forming a meta-benchmark which we use to create generalization tasks for two popular datasets, RobustFill and DeepCoder. We then propose ExeDec, a novel decomposition-based synthesis strategy that predicts execution subgoals to solve problems step-by-step informed by program execution at each step. When used with Transformer models trained from scratch, ExeDec has better synthesis performance and greatly improved compositional generalization ability compared to baselines. Finally, we use our benchmarks to demonstrate that LLMs struggle to compositionally generalize when asked to do programming-by-example in a few-shot setting, but an ExeDec-style prompting approach can improve the generalization ability and overall performance.
****
Federico Cacciamani,Matteo Castiglioni,Nicola Gatti
Online Information Acquisition: Hiring Multiple Agents
We investigate the mechanism design problem faced by a principal who hires \emph{multiple} agents to gather and report costly information. Then, the principal exploits the information to make an informed decision. We model this problem as a game, where the principal announces a mechanism consisting in action recommendations and a payment function, a.k.a. scoring rule. Then, each agent chooses an effort level and receives partial information about an underlying state of nature based on the effort. Finally, the agents report the information (possibly non-truthfully), the principal takes a decision based on this information, and the agents are paid according to the scoring rule. While previous work focuses on single-agent problems, we consider multi-agents settings. This poses the challenge of coordinating the agents' efforts and aggregating correlated information. Indeed, we show that optimal mechanisms must correlate agents' efforts, which introduces externalities among the agents, and hence complex incentive compatibility constraints and equilibrium selection problems. First, we design a polynomial-time algorithm to find an optimal incentive compatible mechanism. Then, we study an online problem, where the principal repeatedly interacts with a group of unknown agents. We design a no-regret algorithm that provides $\widetilde{\mathcal{O}}(T^{2/3})$ regret with respect to an optimal mechanism, matching the state-of-the-art bound for single-agent settings.
****
Yaniv Blumenfeld,Itay Hubara,Daniel Soudry
Towards Cheaper Inference in Deep Networks with Lower Bit-Width Accumulators
The majority of the research on the quantization of Deep Neural Networks (DNNs) is focused on reducing the precision of tensors visible by high-level frameworks (e.g., weights, activations, and gradients). However, current hardware still relies on high-accuracy core operations. Most significant is the operation of accumulating products. This high-precision accumulation operation is gradually becoming the main computational bottleneck. This is because, so far, the usage of low-precision accumulators led to a significant degradation in performance. In this work, we present a simple method to train and fine-tune DNNs, to allow, for the first time, utilization of cheaper, 12-bits accumulators, with no significant

degradation in accuracy. Lastly, we show that as we decrease the accumulation precision further, using fine-grained gradient approximations can improve the DNN accuracy.
****
Ruoqi Yu,Shulei Wang
Treatment Effects Estimation By Uniform Transformer
In observational studies, balancing covariates in different treatment groups is essential to estimate treatment effects. One of the most commonly used methods for such purposes is weighting. The performance of this class of methods usually depends on strong regularity conditions for the underlying model, which might not hold in practice. In this paper, we investigate weighting methods from a functional estimation perspective and argue that the weights needed for covariate balancing could differ from those needed for treatment effects estimation under low regularity conditions. Motivated by this observation, we introduce a new framework of weighting that directly targets the treatment effects estimation. Unlike existing methods, the resulting estimator for a treatment effect under this new framework is a simple kernel-based U-statistic after applying a data-driven transformation to the observed covariates. We characterize the theoretical properties of the new estimators of treatment effects under a nonparametric setting and show that they are able to work robustly under low regularity conditions. The new framework is also applied to several numerical examples to demonstrate its practical merits.
****
Miltiadis Kofinas,Boris Knyazev,Yan Zhang,Yunlu Chen,Gertjan J. Burghouts,Efstratios Gavves,Cees G. M. Snoek,David W. Zhang
Graph Neural Networks for Learning Equivariant Representations of Neural Networks
Neural networks that process the parameters of other neural networks find applications in domains as diverse as classifying implicit neural representations, generating neural network weights, and predicting generalization errors. However, existing approaches either overlook the inherent permutation symmetry in the neural network or rely on intricate weight-sharing patterns to achieve equivariance, while ignoring the impact of the network architecture itself. In this work, we propose to represent neural networks as computational graphs of parameters, which allows us to harness powerful graph neural networks and transformers that preserve permutation symmetry. Consequently, our approach enables a single model to encode neural computational graphs with diverse architectures. We showcase the effectiveness of our method on a wide range of tasks, including classification and editing of implicit neural representations, predicting generalization performance, and learning to optimize, while consistently outperforming state-of-the-art methods. The source code is open-sourced at https://github.com/mkofinas/neural-graphs.
****
Sebastian Shenghong Tay,Chuan-Sheng Foo,Daisuke Urano,Richalynn Leong,Bryan Kian Hsiang Low
A Unified Framework for Bayesian Optimization under Contextual Uncertainty
Bayesian optimization under contextual uncertainty (BOCU) is a family of BO problems in which the learner makes a decision prior to observing the context and must manage the risks involved. Distributionally robust BO (DRBO) is a subset of BOCU that affords robustness against context distribution shift, and includes the optimization of expected values and worst-case values as special cases. By considering the first derivatives of the DRBO objective, we generalize DRBO to one that includes several other uncertainty objectives studied in the BOCU literature such as worst-case sensitivity (and thus notions of risk such as variance, range, and conditional value-at-risk) and mean-risk tradeoffs. We develop a general Thompson sampling algorithm that is able to optimize any objective within the BOCU framework, analyze its theoretical properties, and compare it to suitable baselines across different experimental settings and uncertainty objectives.
****
Zhihan Zhou,Yanrong Ji,Weijian Li,Pratik Dutta,Ramana V Davuluri,Han Liu

DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes
Decoding the linguistic intricacies of the genome is a crucial problem in biology, and pre-trained foundational models such as DNABERT and Nucleotide Transformer have made significant strides in this area. Existing works have largely hinged on k-mer, fixed-length permutations of A, T, C, and G, as the token of the genome language due to its simplicity. However, we argue that the computation and sample inefficiencies introduced by k-mer tokenization are primary obstacles in developing large genome foundational models. We provide conceptual and empirical insights into genome tokenization, building on which we propose to replace k-mer tokenization with Byte Pair Encoding (BPE), a statistics-based data compression algorithm that constructs tokens by iteratively merging the most frequent co-occurring genome segment in the corpus. We demonstrate that BPE not only overcomes the limitations of k-mer tokenization but also benefits from the computational efficiency of non-overlapping tokenization.
Based on these insights, we introduce DNABERT-2, a refined genome foundation model that adapts an efficient tokenizer and employs multiple strategies to overcome input length constraints, reduce time and memory expenditure, and enhance model capability. Furthermore, we identify the absence of a comprehensive and standardized benchmark for genome understanding as another significant impediment to fair comparative analysis. In response, we propose the Genome Understanding Evaluation (GUE), a comprehensive multi-species genome classification dataset that amalgamates 36 distinct datasets across 9 tasks, with input lengths ranging from 70 to 10000. Through comprehensive experiments on the GUE benchmark, we demonstrate that DNABERT-2 achieves comparable performance to the state-of-the-art model with $21 \times$ fewer parameters and approximately $92 \times$ less GPU time in pre-training.
Compared to DNABERT, while being $3 \times$ more efficient, DNABERT-2 outperforms it on 23 out of 28 datasets, with an average improvement of 6 absolute scores on GUE.
The code, data, and pre-trained model are available at \url{https://github.com/MAGICS-LAB/DNABERT_2}.
****
Qiying Yu,Yudi Zhang,Yuyan Ni,Shikun Feng,Yanyan Lan,Hao Zhou,Jingjing Liu
Multimodal Molecular Pretraining via Modality Blending
Self-supervised learning has recently gained growing interest in molecular modeling for scientific tasks such as AI-assisted drug discovery. Current studies consider leveraging both 2D and 3D molecular structures for representation learning. However, relying on straightforward alignment strategies that treat each modality separately, these methods fail to exploit the intrinsic correlation between 2D and 3D representations that reflect the underlying structural characteristics of molecules, and only perform coarse-grained molecule-level alignment. To derive fine-grained alignment and promote structural molecule understanding, we introduce an atomic-relation level "blend-then-predict" self-supervised learning approach, MoleBLEND, which first blends atom relations represented by different modalities into one unified relation matrix for joint encoding, then recovers modality-specific information for 2D and 3D structures individually. By treating atom relationships as anchors, MoleBLEND organically aligns and integrates visually dissimilar 2D and 3D modalities of the same molecule at fine-grained atomic level, painting a more comprehensive depiction of each molecule. Extensive experiments show that MoleBLEND achieves state-of-the-art performance across major 2D/3D molecular benchmarks. We further provide theoretical insights from the perspective of mutual-information maximization, demonstrating that our method unifies contrastive, generative (cross-modality prediction) and mask-then-predict (single-modality prediction) objectives into one single cohesive framework.
****
Jianlan Luo,Perry Dong,Yuexiang Zhai,Yi Ma,Sergey Levine
RLIF: Interactive Imitation Learning as Reinforcement Learning
Although reinforcement learning methods offer a powerful framework for auto-matic skill acquisition, for practical learning-based control problems in domains

such as robotics, imitation learning often provides a more convenient and access
ible
alternative. In particular, an interactive imitation learning method such as DAg
ger,
which queries a near-optimal expert to intervene online to collect correction da
ta for
addressing the distributional shift challenges that afflict naïve behavioral clo
ning,
can enjoy good performance both in theory and practice without requiring manuall
y
specified reward functions and other components of full reinforcement learning
methods. In this paper, we explore how off-policy reinforcement learning can
enable improved performance under assumptions that are similar but potentially
even more practical than those of interactive imitation learning. Our proposed
method uses reinforcement learning with user intervention signals themselves as
rewards. This relaxes the assumption that intervening experts in interactive imi
ta-
tion learning should be near-optimal and enables the algorithm to learn behavior
s
that improve over the potential suboptimal human expert. We also provide a uni-
fied framework to analyze our RL method and DAgger; for which we present the
asymptotic analysis of the suboptimal gap for both methods as well as the non-
asymptotic sample complexity bound of our method. We then evaluate our method
on challenging high-dimensional continuous control simulation benchmarks as
well as real-world robotic vision-based manipulation tasks. The results show tha
t it
strongly outperforms DAgger-like approaches across the different tasks, especial
ly
when the intervening experts are suboptimal. Additional ablations also empirical
ly
verify the proposed theoretical justification that the performance of our method
 is
associated with the choice of intervention model and suboptimality of the expert
.
Code and videos can be found on the project website: https://rlif-page.github.io
****
Shuyan Zhou,Frank F. Xu,Hao Zhu,Xuhui Zhou,Robert Lo,Abishek Sridhar,Xianyi Chen
g,Tianyue Ou,Yonatan Bisk,Daniel Fried,Uri Alon,Graham Neubig
WebArena: A Realistic Web Environment for Building Autonomous Agents
With advances in generative AI, there is now potential for autonomous agents to
manage daily tasks via natural language commands. However, current agents are pr
imarily created and tested in simplified synthetic environments, leading to a di
sconnect with real-world scenarios. In this paper, we build an environment for l
anguage-guided agents that is highly realistic and reproducible. Specifically, w
e focus on agents that perform tasks on the web, and create an environment with
fully functional websites from four common domains: e-commerce, social forum dis
cussions, collaborative software development, and content management. Our enviro
nment is enriched with tools (e.g., a map) and external knowledge bases (e.g., u
ser manuals) to encourage human-like task-solving. Building upon our environment
, we release a set of benchmark tasks focusing on evaluating the functional corr
ectness of task completions. The tasks in our benchmark are diverse, long-horizo
n, and designed to emulate tasks that humans routinely perform on the internet.
We experiment with several baseline agents, integrating recent techniques such a
s reasoning before acting. The results demonstrate that solving complex tasks i
s challenging: our best GPT-4-based agent only achieves an end-to-end task succe
ss rate of 14.41%, significantly lower than the human performance of 78.24%. The
se results highlight the need for further development of robust agents, that cur
rent state-of-the-art large language models are far from perfect performance in
these real-life tasks, and that \ours can be used to measure such progress.\foot
note{Code, data, environment reproduction instructions, video demonstrations are

available in the supplementary.}
****
Bolian Li,Ruqi Zhang
Entropy-MCMC: Sampling from Flat Basins with Ease
Bayesian deep learning counts on the quality of posterior distribution estimation. However, the posterior of deep neural networks is highly multi-modal in nature, with local modes exhibiting varying generalization performance. Given a practical budget, targeting at the original posterior can lead to suboptimal performance, as some samples may become trapped in "bad" modes and suffer from overfitting. Leveraging the observation that "good" modes with low generalization error often reside in flat basins of the energy landscape, we propose to bias sampling on the posterior toward these flat regions. Specifically, we introduce an auxiliary guiding variable, the stationary distribution of which resembles a smoothed posterior free from sharp modes, to lead the MCMC sampler to flat basins. By integrating this guiding variable with the model parameter, we create a simple joint distribution that enables efficient sampling with minimal computational overhead. We prove the convergence of our method and further show that it converges faster than several existing flatness-aware methods in the strongly convex setting. Empirical results demonstrate that our method can successfully sample from flat basins of the posterior, and outperforms all compared baselines on multiple benchmarks including classification, calibration, and out-of-distribution detection.
****
Zeyu Liu,Gourav Datta,Anni Li,Peter Anthony Beerel
LMUFormer: Low Complexity Yet Powerful Spiking Model With Legendre Memory Units
Transformer models have demonstrated high accuracy in numerous applications but have high complexity and lack sequential processing capability making them ill-suited for many streaming applications at the edge where devices are heavily resource-constrained. Thus motivated, many researchers have proposed reformulating the transformer models as RNN modules which modify the self-attention computation with explicit states. However, these approaches often incur significant performance degradation.
The ultimate goal is to develop a model that has the following properties: parallel training, streaming and low-cost inference, and state-of-the-art (SOTA) performance. In this paper, we propose a new direction to achieve this goal. We show how architectural modifications to a fully-sequential recurrent model can help push its performance toward Transformer models while retaining its sequential processing capability. Specifically, inspired by the recent success of Legendre Memory Units (LMU) in sequence learning tasks, we propose LMUFormer, which augments the LMU with convolutional patch embedding and convolutional channel mixer.
Moreover, we present a spiking version of this architecture, which introduces the benefit of states within the patch embedding and channel mixer modules while simultaneously reducing the computing complexity.
We evaluated our architectures on multiple sequence datasets. Of particular note is our performance on the Speech Commands V2 dataset (35 classes). In comparison to SOTA transformer-based models within the ANN domain, our LMUFormer demonstrates comparable performance while necessitating a remarkable $70\times$ reduction in parameters and a substantial $140\times$ decrement in FLOPs. Furthermore, when benchmarked against extant low-complexity SNN variants, our model establishes a new SOTA with an accuracy of 96.12\%.
Additionally, owing to our model's proficiency in real-time data processing, we are able to achieve a 32.03\% reduction in sequence length, all while incurring an inconsequential decline in performance.
****
Lucas Dax Lingle
Transformer-VQ: Linear-Time Transformers via Vector Quantization
We introduce Transformer-VQ, a decoder-only transformer computing softmax-based dense self-attention in linear time. Transformer-VQ's efficient attention is enabled by vector-quantized keys and a novel caching mechanism.
In our large-scale experiments, Transformer-VQ is shown highly competitive in qu

ality, obtaining 0.99 bpb on Enwik8, 26.6 ppl on PG-19, and 3.16 bpb on ImageNet 64. In addition, the optimized implementation of Transformer-VQ is over 3x faster than a comparable quadratic-time transformer at sequence length 8k, is over 12x faster at 32k, and can scale to 131k with similar throughput. Code available: \url{https://github.com/transformer-vq/transformer_vq}
****

Frederic Koehler,Thuy-Duong Vuong
Sampling Multimodal Distributions with the Vanilla Score: Benefits of Data-Based Initialization
There is a long history, as well as a recent explosion of interest, in statistical and generative modeling approaches based on \emph{score functions} --- derivatives of the log-likelihood of a distribution. In seminal works, Hyv\"arinen proposed vanilla score matching as a way to learn distributions from data by computing an estimate of the score function of the underlying ground truth, and established connections between this method and established techniques like Contrastive Divergence and Pseudolikelihood estimation. It is by now well-known that vanilla score matching has significant difficulties learning multimodal distributions. Although there are various ways to overcome this difficulty, the following question has remained unanswered --- is there a natural way to sample multimodal distributions using just the vanilla score? Inspired by a long line of related experimental works, we prove that the Langevin diffusion with early stopping, initialized at the empirical distribution, and run on a score function estimated from data successfully generates natural multimodal distributions (mixtures of log-concave distributions).
****

Huafeng Qin,Xin Jin,Yun Jiang,Mounîm El-Yacoubi,Xinbo Gao
Adversarial AutoMixup
Data mixing augmentation has been widely applied to improve the generalization ability of deep neural networks. Recently, offline data mixing augmentation, e.g. handcrafted and saliency information-based mixup, has been gradually replaced by automatic mixing approaches. Through minimizing two sub-tasks, namely, mixed sample generation and mixup classification in an end-to-end way, AutoMix significantly improves accuracy on image classification tasks. However, as the optimization objective is consistent for the two sub-tasks, this approach is prone to generating consistent instead of diverse mixed samples, which results in overfitting for target task training. In this paper, we propose AdAutomixup, an adversarial automatic mixup augmentation approach that generates challenging samples to train a robust classifier for image classification, by alternatively optimizing the classifier and the mixup sample generator. AdAutomixup comprises two modules, a mixed example generator, and a target classifier. The mixed sample generator aims to produce hard mixed examples to challenge the target classifier, while the target classifier's aim is to learn robust features from hard mixed examples to improve generalization. To prevent the collapse of the inherent meanings of images, we further introduce an exponential moving average (EMA) teacher and cosine similarity to train AdAutomixup in an end-to-end way. Extensive experiments on seven image benchmarks consistently prove that our approach outperforms the state of the art in various classification scenarios. The source code is available at
https://github.com/JinXins/Adversarial-AutoMixup.
****

Zhipeng Xie,Yahe Li
Information Retention via Learning Supplemental Features
The information bottleneck principle provides an information-theoretic method for learning a good representation as a trade-off between conciseness and predictive ability, which can reduce information redundancy, eliminate irrelevant and superfluous features, and thus enhance the in-domain generalizability. However, in low-resource or out-of-domain scenarios where the assumption of i.i.d does not necessarily hold true, superfluous (or redundant) relevant features may be supplemental to the mainline features of the model, and be beneficial in making prediction for test dataset with distribution shift. Therefore, instead of squeezing

the input information by information bottleneck, we propose to keep as much rele
vant information as possible in use for making predictions. A three-stage superv
ised learning framework is designed and implemented to jointly learn the mainlin
e and supplemental features, relieving supplemental features from the suppressio
n of mainline features. Extensive experiments have shown that the learned repres
entations of our method have good in-domain and out-of-domain generalization abi
lities, especially in low-resource cases.
****
Hao Wang,Yongsheng Yu,Tiejian Luo,Heng Fan,Libo Zhang
MaGIC: Multi-modality Guided Image Completion
Vanilla image completion approaches exhibit sensitivity to large missing regions
, attributed to the limited availability of reference information for plausible
generation. To mitigate this, existing methods incorporate the extra cue as guid
ance for image completion. Despite improvements, these approaches are often rest
ricted to employing a *single modality* (e.g., *segmentation* or *sketch* maps),
 which lacks scalability in leveraging multi-modality for more plausible complet
ion.
In this paper, we propose a novel, simple yet effective method for **M**ulti-mod
al **G**uided **I**mage **C**ompletion, dubbed **MaGIC**, which not only sup
ports a wide range of single modality as the guidance (e.g., *text*, *canny edge
*, *sketch*, *segmentation*, *depth*, and *pose*), but also adapts to arbitraril
y customized combinations of these modalities (i.e., *arbitrary multi-modality*)
 for image completion.
For building MaGIC, we first introduce a modality-specific conditional U-Net (MC
U-Net) that injects single-modal signal into a U-Net denoiser for single-modal g
uided image completion. Then, we devise a consistent modality blending (CMB) met
hod to leverage modality signals encoded in multiple learned MCU-Nets through gr
adient guidance in latent space. Our CMB is *training-free*, thereby avoiding th
e cumbersome joint re-training of different modalities, which is the secret of M
aGIC to achieve exceptional flexibility in accommodating new modalities for comp
letion.
Experiments show the superiority of MaGIC over state-of-the-art methods and its
generalization to various completion tasks.
****
Nuoya Xiong,Zhihan Liu,Zhaoran Wang,Zhuoran Yang
Sample-Efficient Multi-Agent RL: An Optimization Perspective
We study multi-agent reinforcement learning (MARL) for the general-sum Markov Ga
mes (MGs) under general function approximation.
    In order to find the minimum assumption for sample-efficient learning, we in
troduce a novel complexity measure called the Multi-Agent Decoupling Coefficient
 (MADC) for general-sum MGs. Using this measure, we propose the first unified al
gorithmic framework that ensures sample efficiency in learning Nash Equilibrium,
 Coarse Correlated Equilibrium, and Correlated Equilibrium for both model-based
and model-free MARL problems with low MADC. We also show that our algorithm prov
ides comparable sublinear regret to the existing works. Moreover, our algorithm
combines an equilibrium-solving oracle with a single objective optimization subp
rocedure that solves for the regularized payoff of each deterministic joint poli
cy, which avoids solving constrained optimization problems within data-dependent
 constraints (Jin et al. 2020; Wang et al. 2023) or executing sampling procedure
s with complex multi-objective optimization problems (Foster et al. 2023), thus
being more amenable to empirical implementation.
****
Gianluca Scarpellini,Ksenia Konyushkova,Claudio Fantacci,Thomas Paine,Yutian Che
n,Misha Denil
π2vec: Policy Representation with Successor Features
This paper introduces π2vec, a method for representing black box policies as
 comparable feature vectors.
Our method combines the strengths of foundation models that serve as generic and
 powerful state representations and successor features that can model the future
 occurrence of the states for a policy.

π2vec represents the behavior of policies by capturing the statistics of the features from a pretrained model with the help of successor feature framework.
We focus on the offline setting where policies and their representations are trained on a fixed dataset of trajectories.
Finally, we employ linear regression on π2vec vector representations to predict the performance of held out policies.
The synergy of these techniques results in a method for efficient policy evaluation in resource constrained environments.
****
WANG Jiaxu,Ziyi Zhang,Renjing Xu
Learning Robust Generalizable Radiance Field with Visibility and Feature Augmented Point Representation
This paper introduces a novel paradigm for the generalizable neural radiance field (NeRF). Previous generic NeRFs combine multiview stereo techniques with image-based neural rendering, yielding impressive results, while suffering from three issues. First, occlusions often result in inconsistent feature matching. Then, they deliver distortions and artifacts in geometric discontinuities and locally sharp shapes due to their individual process of sampled points and rough feature aggregation. Third, their image-based representations experience severe degradations
when source views are not near enough to the target view. To address challenges, we propose the first paradigm that constructs the generalizable neural field based on point-based rather than image-based rendering, which we call the Generalizable neural Point Field (GPF). Our approach explicitly models visibilities by geometric priors and augments them with neural features. We propose a novel nonuniform log sampling strategy to improve rendering speed and reconstruction quality. Moreover, we present a learnable kernel spatially augmented
with features for feature aggregations, mitigating distortions at places with drastically varying geometries. Besides, our representation can be easily manipulated. Experiments show that our model can deliver better geometries, view consistencies, and rendering quality than all counterparts and benchmarks on three datasets in both generalization and finetuning settings, preliminarily proving the potential of the new paradigm for generalizable NeRF
****
Yutong He,Naoki Murata,Chieh-Hsin Lai,Yuhta Takida,Toshimitsu Uesaka,Dongjun Kim,Wei-Hsiang Liao,Yuki Mitsufuji,J Zico Kolter,Ruslan Salakhutdinov,Stefano Ermon
Manifold Preserving Guided Diffusion
Despite the recent advancements, conditional image generation still faces challenges of cost, generalizability, and the need for task-specific training. In this paper, we propose Manifold Preserving Guided Diffusion (MPGD), a training-free conditional generation framework that leverages pretrained diffusion models and off-the-shelf neural networks with minimal additional inference cost for a broad range of tasks. Specifically, we leverage the manifold hypothesis to refine the guided diffusion steps and introduce a shortcut algorithm in the process. We then propose two methods for on-manifold training-free guidance using pre-trained autoencoders and demonstrate that our shortcut inherently preserves the manifolds when applied to latent diffusion models. Our experiments show that MPGD is efficient and effective for solving a variety of conditional generation applications in low-compute settings, and can consistently offer up to 3.8× speed-ups with the same number of diffusion steps while maintaining high sample quality compared to the baselines.
****
Haoqi Yuan,Zhancun Mu,Feiyang Xie,Zongqing Lu
Pre-Training Goal-based Models for Sample-Efficient Reinforcement Learning
Pre-training on task-agnostic large datasets is a promising approach for enhancing the sample efficiency of reinforcement learning (RL) in solving complex tasks. We present PTGM, a novel method that pre-trains goal-based models to augment RL by providing temporal abstractions and behavior regularization. PTGM involves pre-training a low-level, goal-conditioned policy and training a high-level policy to generate goals for subsequent RL tasks. To address the challenges posed by

the high-dimensional goal space, while simultaneously maintaining the agent's capability to accomplish various skills, we propose clustering goals in the dataset to form a discrete high-level action space. Additionally, we introduce a pre-trained goal prior model to regularize the behavior of the high-level policy in RL, enhancing sample efficiency and learning stability. Experimental results in a robotic simulation environment and the challenging open-world environment of Minecraft demonstrate PTGM's superiority in sample efficiency and task performance compared to baselines. Moreover, PTGM exemplifies enhanced interpretability and generalization of the acquired low-level skills.
****
Simon Segert
Flat Minima in Linear Estimation and an Extended Gauss Markov Theorem
We consider the problem of linear estimation, and establish an extension of the Gauss-Markov theorem, in which the bias operator is allowed to be non-zero but bounded with respect to a matrix norm of Schatten type. We derive simple and explicit formulas for the optimal estimator in the cases of Nuclear and Spectral norms (with the Frobenius case recovering ridge regression). Additionally, we analytically derive the generalization error in multiple random matrix ensembles, and compare with Ridge regression. Finally, we conduct an extensive simulation study, in which we show that the cross-validated Nuclear and Spectral regressors can outperform Ridge in several circumstances.
****
Gundeep Arora,Srujana Merugu,Anoop Saladi,Rajeev Rastogi
Leveraging Uncertainty Estimates To Improve Classifier Performance
Binary classification typically involves predicting the label of an instance based on whether the model score for the positive class exceeds a threshold chosen based on the application requirements (e.g., maximizing recall for a precision bound). However, model scores are often not aligned with true positivity rate. This is especially true when the training involves a differential sampling of classes or there is distributional drift between train and test settings. In this paper, we provide theoretical analysis and empirical evidence of the dependence of estimation bias on both uncertainty and model score. Further, we formulate the decision boundary selection using both model score and uncertainty, prove that it is NP-hard, and present algorithms based on dynamic programming and isotonic regression. Evaluation of the proposed algorithms on three real-world datasets yield 25\%-40\% improvement in recall at high precision bounds over the traditional approach of using model score alone, highlighting the benefits of leveraging uncertainty.
****
Raj Ghugare,Santiago Miret,Adriana Hugessen,Mariano Phielipp,Glen Berseth
Searching for High-Value Molecules Using Reinforcement Learning and Transformers
Reinforcement learning (RL) over text representations can be effective for finding high-value policies that can search over graphs. However, RL requires careful structuring of the search space and algorithm design to be effective in this challenge. Through extensive experiments, we explore how different design choices for text grammar and algorithmic choices for training can affect an RL policy's ability to generate molecules with desired properties. We arrive at a new RL-based molecular design algorithm (ChemRLformer) and perform a thorough analysis using 25 molecule design tasks, including computationally complex protein docking simulations. From this analysis, we discover unique insights in this problem space and show that ChemRLformer achieves state-of-the-art performance while being more straightforward than prior work by demystifying which design choices are actually helpful for text-based molecule design.
****
Peng Wang,Hao Tan,Sai Bi,Yinghao Xu,Fujun Luan,Kalyan Sunkavalli,Wenping Wang,Zexiang Xu,Kai Zhang
PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction
We propose a Pose-Free Large Reconstruction Model (PF-LRM) for reconstructing a 3D object from a few unposed images even with little visual overlap, while simultaneously estimating the relative camera poses in ~1.3 seconds on a single A100

GPU. PF-LRM is a highly scalable method utilizing self-attention blocks to exchange information between 3D object tokens and 2D image tokens; we predict a coarse point cloud for each view, and then use a differentiable Perspective-n-Point (PnP) solver to obtain camera poses. When trained on a huge amount of multi-view posed data of ~1M objects, PF-LRM shows strong cross-dataset generalization ability, and outperforms baseline methods by a large margin in terms of pose prediction accuracy and 3D reconstruction quality on various unseen evaluation datasets. We also demonstrate our model's applicability in downstream text/image-to-3D task with fast feed-forward inference. Our project website is at: https://totoro97.github.io/pf-lrm.
****

Matthieu Blanke,Marc Lelarge

Interpretable Meta-Learning of Physical Systems

Machine learning methods can be a valuable aid in the scientific process, but they need to face challenging settings where data come from inhomogeneous experimental conditions. Recent meta-learning methods have made significant progress in multi-task learning, but they rely on black-box neural networks, resulting in high computational costs and limited interpretability. We introduce CAMEL, a new meta-learning architecture capable of learning efficiently from multiple environments, with an affine structure with respect to the learning task. We prove that CAMEL can identify the physical parameters of the system, enabling interpreable learning. We demonstrate the competitive generalization performance and the low computational cost of our method by comparing it to state-of-the-art algorithms on physical systems, ranging from toy models to complex, non-analytical systems. The interpretability of our method is illustrated with original applications to parameter identification and to adaptive control and system identification.
****

Jiashuo Sun,Chengjin Xu,Lumingyuan Tang,Saizhuo Wang,Chen Lin,Yeyun Gong,Lionel Ni,Heung-Yeung Shum,Jian Guo

Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph

Although large language models (LLMs) have achieved significant success in various tasks, they often struggle with hallucination problems, especially in scenarios requiring deep and responsible reasoning. These issues could be partially addressed by introducing external knowledge graphs (KG) in LLM reasoning. In this paper, we propose a new LLM-KG integrating paradigm ``$\hbox{LLM}\otimes\hbox{KG}$'' which treats the LLM as an agent to interactively explore related entities and relations on KGs and perform reasoning based on the retrieved knowledge. We further implement this paradigm by introducing a new approach called Think-on-Graph (ToG), in which the LLM agent iteratively executes beam search on KG, discovers the most promising reasoning paths, and returns the most likely reasoning results. We use a number of well-designed experiments to examine and illustrate the following advantages of ToG: 1) compared with LLMs, ToG has better deep reasoning power; 2) ToG has the ability of knowledge traceability and knowledge correctability by leveraging LLMs reasoning and expert feedback; 3) ToG provides a flexible plug-and-play framework for different LLMs, KGs and prompting strategies without any additional training cost; 4) the performance of ToG with small LLM models could exceed large LLM such as GPT-4 in certain scenarios and this reduces the cost of LLM deployment and application. As a training-free method with lower computational cost and better generality, ToG achieves overall SOTA in 6 out of 9 datasets where most previous SOTAs rely on additional training.
****

Guibin Zhang,Kun Wang,Wei Huang,Yanwei Yue,Yang Wang,Roger Zimmermann,Aojun Zhou,Dawei Cheng,Jin Zeng,Yuxuan Liang

Graph Lottery Ticket Automated

Graph Neural Networks (GNNs) have emerged as the leading deep learning models for graph-based representation learning. However, the training and inference of GNNs on large graphs remain resource-intensive, impeding their utility in real-world scenarios and curtailing their applicability in deeper and more sophisticated GNN architectures. To address this issue, the Graph Lottery Ticket (GLT) hypot

hesis assumes that GNN with random initialization harbors a pair of core subgrap
h and sparse subnetwork, which can yield comparable performance and higher effic
iency to that of the original dense network and complete graph. Despite that GLT
 offers a new paradigm for GNN training and inference, existing GLT algorithms h
eavily rely on trial-and-error pruning rate tuning and scheduling, and adhere to
 an irreversible pruning paradigm that lacks elasticity. Worse still, current me
thods suffer scalability issues when applied to deep GNNs, as they maintain the
same topology structure across all layers. These challenges hinder the integrati
on of GLT into deeper and larger-scale GNN contexts. To bridge this critical ga
p, this paper introduces an $\textbf{A}$daptive, $\textbf{D}$ynamic, and $\textb
f{A}$utomated framework for identifying $\textbf{G}$raph $\textbf{L}$ottery $\te
xtbf{T}$ickets ($\textbf{AdaGLT}$). Our proposed method derives its key advantag
es and addresses the above limitations through the following three aspects: 1) t
ailoring layer-adaptive sparse structures for various datasets and GNNs, thus en
dowing it with the capability to facilitate deeper GNNs; 2) integrating the prun
ing and training processes, thereby achieving a dynamic workflow encompassing bo
th pruning and restoration; 3) automatically capturing graph lottery tickets acr
oss diverse sparsity levels, obviating the necessity for extensive pruning param
eter tuning. More importantly, we rigorously provide theoretical proofs to guara
ntee $\textbf{AdaGLT}$ to mitigate over-smoothing issues and obtain improved sp
arse structures in deep GNN scenarios. Extensive experiments demonstrate that $\
textbf{AdaGLT}$ outperforms state-of-the-art competitors across multiple graph d
atasets of various scales and types, particularly in scenarios involving deep GN
Ns.
****
Gaurav Shrivastava,Ser-Nam Lim,Abhinav Shrivastava
Video Decomposition Prior: Editing Videos Layer by Layer
In the evolving landscape of video editing methodologies, a majority of deep le
arning techniques are often reliant on extensive datasets of observed input and
ground truth sequence pairs for optimal performance. Such reliance often falters
 when acquiring data becomes challenging, especially in tasks like video dehazin
g and relighting, where replicating identical motions and camera angles in both
corrupted and ground truth sequences is complicated. Moreover, these conventiona
l methodologies perform best when the test distribution closely mirrors the trai
ning distribution. Recognizing these challenges, this paper introduces a novel v
ideo decomposition prior `VDP' framework which derives inspiration from professi
onal video editing practices. Our methodology does not mandate task-specific ext
ernal data corpus collection, instead pivots to utilizing the motion and appeara
nce of the input video. VDP framework decomposes a video sequence into a set of
multiple RGB layers and associated opacity levels. These set of layers are then
manipulated individually to obtain the desired results. We addresses tasks such
as video object segmentation, dehazing, and relighting. Moreover, we introduce a
 novel logarithmic video decomposition formulation for video relighting tasks, s
etting a new benchmark over the existing methodologies. We evaluate our approach
 on standard video datasets like DAVIS, REVIDE, & SDSD and show qualitative resu
lts on a diverse array of internet videos.
****
Haque Ishfaq,Qingfeng Lan,Pan Xu,A. Rupam Mahmood,Doina Precup,Anima Anandkumar,
Kamyar Azizzadenesheli
Provable and Practical: Efficient Exploration in Reinforcement Learning via Lang
evin Monte Carlo
We present a scalable and effective exploration strategy based on Thompson sampl
ing for reinforcement learning (RL). One of the key shortcomings of existing Th
ompson sampling algorithms is the need to perform a Gaussian approximation of th
e posterior distribution, which is not a good surrogate in most practical settin
gs. We instead directly sample the Q function from its posterior distribution, b
y using Langevin Monte Carlo, an efficient type of Markov Chain Monte Carlo (MC
MC) method. Our method only needs to perform noisy gradient descent updates to l
earn the exact posterior distribution of the Q function, which makes our approac
h easy to deploy in deep RL. We provide a rigorous theoretical analysis for the

proposed method and demonstrate that, in the linear Markov decision process (li
near MDP) setting, it has a regret bound of $\tilde{O}(d^{3/2}H^{3/2}\sqrt{T})$,
where d is the dimension of the feature mapping, H is the planning horizon,
and T is the total number of steps. We apply this approach to deep RL, by usi
ng Adam optimizer to perform gradient updates. Our approach achieves better or s
imilar results compared with state-of-the-art deep RL algorithms on several chal
lenging exploration tasks from the Atari57 suite.
****
Yiwei Li,Peiwen Yuan,Shaoxiong Feng,Boyuan Pan,Xinglin Wang,Bin Sun,Heda Wang,Ka
n Li
Escape Sky-high Cost: Early-stopping Self-Consistency for Multi-step Reasoning
Self-consistency (SC) has been a widely used decoding strategy for chain-of-thou
ght reasoning. Despite bringing significant performance improvements across a va
riety of multi-step reasoning tasks, it is a high-cost method that requires mult
iple sampling with the preset size. In this paper, we propose a simple and scala
ble sampling process, Early-Stopping Self-Consistency (ESC), to greatly reduce t
he cost of SC without sacrificing performance. On this basis, one control scheme
for ESC is further derivated to dynamically choose the performance-cost balance
for different tasks and models. To demonstrate ESC's effectiveness, we conducte
d extensive experiments on three popular categories of reasoning tasks: arithmet
ic, commonsense and symbolic reasoning over language models with varying scales.
The empirical results show that ESC reduces the average number of sampling of c
hain-of-thought reasoning by a significant margin on six benchmarks, including M
ATH (-33.8%), GSM8K (-80.1%), StrategyQA (-76.8%), CommonsenseQA (-78.5%), Coin
Flip (-84.2%) and Last Letters (-67.4%), while attaining comparable performances
.
****
Hugo Cui,Florent Krzakala,Eric Vanden-Eijnden,Lenka Zdeborova
Analysis of Learning a Flow-based Generative Model from Limited Sample Complexit
y
We study the problem of training a flow-based generative model, parametrized by
a two-layer autoencoder, to sample from a high-dimensional Gaussian mixture. We
provide a sharp end-to-end analysis of the problem. First, we provide a tight cl
osed-form characterization of the learnt velocity field, when parametrized by a
shallow denoising auto-encoder trained on a finite number n of samples from th
e target distribution. Building on this analysis, we provide a sharp description
of the corresponding generative flow, which pushes the base Gaussian density fo
rward to an approximation of the target density. In particular, we provide close
d-form formulae for the distance between the means of the generated mixture and
the mean of the target mixture, which we show decays as $\Theta_n(\frac{1}{n})$.
Finally, this rate is shown to be in fact Bayes-optimal.
****
Haochen Luo,Jindong Gu,Fengyuan Liu,Philip Torr
An Image Is Worth 1000 Lies: Transferability of Adversarial Images across Prompt
s on Vision-Language Models
Different from traditional task-specific vision models, recent large VLMs can re
adily adapt to different vision tasks by simply using different textual instruct
ions, i.e., prompts. However, a well-known concern about traditional task-specif
ic vision models is that they can be misled by imperceptible adversarial perturb
ations. Furthermore, the concern is exacerbated by the phenomenon that the same
adversarial perturbations can fool different task-specific models. Given that VL
Ms rely on prompts to adapt to different tasks, an intriguing question emerges:
Can a single adversarial image mislead all predictions of VLMs when a thousand d
ifferent prompts are given? This question essentially introduces a novel perspec
tive on adversarial transferability: cross-prompt adversarial transferability. I
n this work, we propose the Cross-Prompt Attack (CroPA). This proposed method up
dates the visual adversarial perturbation with learnable textual prompts, which
are designed to counteract the misleading effects of the adversarial image. By d
oing this, CroPA significantly improves the transferability of adversarial examp
les across prompts. Extensive experiments are conducted to verify the strong cro

ss-prompt adversarial transferability of CroPA with prevalent VLMs including Fla
mingo, BLIP-2, and InstructBLIP in various different tasks.
****
Haozhao Wang,Haoran Xu,Yichen Li,Yuan Xu,Ruixuan Li,Tianwei Zhang
FedCDA: Federated Learning with Cross-rounds Divergence-aware Aggregation
In Federated Learning (FL), model aggregation is pivotal. It involves a global server iteratively aggregating client local trained models in successive rounds without accessing private data. Traditional methods typically aggregate the local models from the current round alone. However, due to the statistical heterogeneity across clients, the local models from different clients may be greatly diverse, making the obtained global model incapable of maintaining the specific knowledge of each local model. In this paper, we introduce a novel method, FedCDA, which selectively aggregates cross-round local models, decreasing discrepancies between the global model and local models.
The principle behind FedCDA is that due to the different global model parameters received in different rounds and the non-convexity of deep neural networks, the local models from each client may converge to different local optima across rounds. Therefore, for each client, we select a local model from its several recent local models obtained in multiple rounds, where the local model is selected by minimizing its divergence from the local models of other clients. This ensures the aggregated global model remains close to all selected local models to maintain their data knowledge. Extensive experiments conducted on various models and datasets reveal our approach outperforms state-of-the-art aggregation methods.
****
Jun Nie,Yonggang Zhang,Zhen Fang,Tongliang Liu,Bo Han,Xinmei Tian
Out-of-Distribution Detection with Negative Prompts
Out-of-distribution (OOD) detection is indispensable for open-world machine learning models. Inspired by recent success in large pre-trained language-vision models, e.g., CLIP, advanced works have achieved impressive OOD detection results by matching the *similarity* between image features and features of learned prompts, i.e., positive prompts. However, existing works typically struggle with OOD samples having similar features with those of known classes. One straightforward approach is to introduce negative prompts to achieve a *dissimilarity* matching, which further assesses the anomaly level of image features by introducing the absence of specific features. Unfortunately, our experimental observations show that either employing a prompt like "not a photo of a" or learning a prompt to represent "not containing" fails to capture the dissimilarity for identifying OOD samples. The failure may be contributed to the diversity of negative features, i.e., tons of features could indicate features not belonging to a known class. To this end, we propose to learn a set of negative prompts for each class. The learned positive prompt (for all classes) and negative prompts (for each class) are leveraged to measure the similarity and dissimilarity in the feature space simultaneously, enabling more accurate detection of OOD samples. Extensive experiments are conducted on diverse OOD detection benchmarks, showing the effectiveness of our proposed method.
****
Enneng Yang,Zhenyi Wang,Li Shen,Shiwei Liu,Guibing Guo,Xingwei Wang,Dacheng Tao
AdaMerging: Adaptive Model Merging for Multi-Task Learning
Multi-task learning (MTL) aims to empower a model to tackle multiple tasks simultaneously. A recent development known as task arithmetic has revealed that several models, each fine-tuned for distinct tasks, can be directly merged into a single model to execute MTL without necessitating a retraining process using the initial training data. Nevertheless, this direct addition of models often leads to a significant deterioration in the overall performance of the merged model. This decline occurs due to potential conflicts and intricate correlations among the multiple tasks. Consequently, the challenge emerges of how to merge pre-trained models more effectively without using their original training data. This paper introduces an innovative technique called Adaptive Model Merging (AdaMerging). This approach aims to autonomously learn the coefficients for model merging, either in a task-wise or layer-wise manner, without relying on the original training

data. Specifically, our AdaMerging method operates as an automatic, unsupervised task arithmetic scheme. It leverages entropy minimization on unlabeled test samples from the multi-task setup as a surrogate objective function to iteratively refine the merging coefficients of the multiple models. Our experimental findings across eight tasks demonstrate the efficacy of the AdaMerging scheme we put forth. Compared to the current state-of-the-art (SOTA) task arithmetic merging scheme, AdaMerging showcases a remarkable 11\% improvement in performance. Notably, AdaMerging also exhibits superior generalization capabilities when applied to unseen downstream tasks. Furthermore, it displays a significantly enhanced robustness to data distribution shifts that may occur during the testing phase.
****
Pol Labarbarie,Adrien CHAN-HON-TONG,Stéphane Herbin,Milad Leyli-abadi
Optimal transport based adversarial patch to leverage large scale attack transferability
Adversarial patch attacks, where a small patch is placed in the scene to fool neural networks, have been studied for numerous applications. Focusing on image classification, we consider the setting of a black-box transfer attack where an attacker does not know the target model. Instead of forcing corrupted image representations to cross the nearest decision boundaries or converge to a particular point, we propose a distribution-oriented approach. We rely on optimal transport to push the feature distribution of attacked images towards an already modeled distribution. We show that this new distribution-oriented approach leads to better transferable patches. Through digital experiments conducted on ImageNet-1K, we provide evidence that our new patches are the only ones that can simultaneously influence multiple Transformer models and Convolutional Neural Networks. Physical world experiments demonstrate that our patch can affect systems in deployment without explicit knowledge.
****
Hanlei Zhang,Xin Wang,Hua Xu,Qianrui Zhou,Kai Gao,Jianhua Su,jinyue Zhao,Wenrui Li,Yanting Chen
MIntRec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations
Multimodal intent recognition poses significant challenges, requiring the incorporation of non-verbal modalities from real-world contexts to enhance the comprehension of human intentions. However, most existing multimodal intent benchmark datasets are limited in scale and suffer from difficulties in handling out-of-scope samples that arise in multi-turn conversational interactions. In this paper, we introduce MIntRec2.0, a large-scale benchmark dataset for multimodal intent recognition in multi-party conversations. It contains 1,245 high-quality dialogues with 15,040 samples, each annotated within a new intent taxonomy of 30 fine-grained classes, across text, video, and audio modalities. In addition to more than 9,300 in-scope samples, it also includes over 5,700 out-of-scope samples appearing in multi-turn contexts, which naturally occur in real-world open scenarios, enhancing its practical applicability. Furthermore, we provide comprehensive information on the speakers in each utterance, enriching its utility for multi-party conversational research. We establish a general framework supporting the organization of single-turn and multi-turn dialogue data, modality feature extraction, multimodal fusion, as well as in-scope classification and out-of-scope detection. Evaluation benchmarks are built using classic multimodal fusion methods, ChatGPT, and human evaluators. While existing methods incorporating nonverbal information yield improvements, effectively leveraging context information and detecting out-of-scope samples remains a substantial challenge. Notably, powerful large language models exhibit a significant performance gap compared to humans, highlighting the limitations of machine learning methods in the advanced cognitive intent understanding task. We believe that MIntRec2.0 will serve as a valuable resource, providing a pioneering foundation for research in human-machine conversational interactions, and significantly facilitating related applications.
The full dataset and codes are available for use at https://github.com/thuiar/MIntRec2.0.
****

Mohammadreza Mousavi Kalan,Samory Kpotufe
Tight Rates in Supervised Outlier Transfer Learning
A critical barrier to learning an accurate decision rule for outlier detection is the scarcity of outlier data. As such, practitioners often turn to the use of similar but imperfect outlier data from which they might \emph{transfer} information to the target outlier detection task. Despite the recent empirical success of transfer learning in outlier detection, a fundamental understanding of when and how knowledge can be transferred from a source to a target in outlier detection remains elusive. In this work, we adopt the traditional framework of Neyman-Pearson classification---which formalizes \emph{supervised outlier detection}, i.e., unbalanced classification---with the added assumption that we have access to both source and (some or no) target outlier data. Our main results are then as follows:

We first determine the information-theoretic limits of the problem under a measure of discrepancy that extends some existing notions from traditional balanced classification; interestingly, unlike in balanced classification, seemingly very dissimilar sources can provide much information about a target, thus resulting in fast transfer.

We then show that, in principle, these information-theoretic limits are achievable by \emph{adaptive} procedures, i.e., procedures with no a priori information on the discrepancy between source and target distributions.
****
Zehao Dong,Muhan Zhang,Philip Payne,Michael A Province,Carlos Cruchaga,Tianyu Zhao,Fuhai Li,Yixin Chen
Rethinking the Power of Graph Canonization in Graph Representation Learning with Stability
The expressivity of Graph Neural Networks (GNNs) has been studied broadly in recent years to reveal the design principles for more powerful GNNs. Graph canonization is known as a typical approach to distinguish non-isomorphic graphs, yet rarely adopted when developing expressive GNNs. This paper proposes to maximize the expressivity of GNNs by graph canonization, then the power of such GNNs is studies from the perspective of model stability. A stable GNN will map similar graphs to close graph representations in the vectorial space, and the stability of GNNs is critical to generalize their performance to unseen graphs. We theoretically reveal the trade-off of expressivity and stability in graph-canonization-enhanced GNNs. Then we introduce a notion of universal graph canonization as the general solution to address the trade-off and characterize a widely applicable sufficient condition to solve the universal graph canonization. A comprehensive set of experiments demonstrates the effectiveness of the proposed method. In many popular graph benchmark datasets, graph canonization successfully enhances GNNs and provides highly competitive performance, indicating the capability and great potential of proposed method in general graph representation learning. In graph datasets where the sufficient condition holds, GNNs enhanced by universal graph canonization consistently outperform GNN baselines and successfully improve the SOTA performance up to 31%, providing the optimal solution to numerous challenging real-world graph analytical tasks like gene network representation learning in bioinformatics.
****
Di Wu,Jun Bai,Yiliao Song,Junjun Chen,Wei Zhou,Yong Xiang,Atul Sajjanhar
FedInverse: Evaluating Privacy Leakage in Federated Learning
Federated Learning (FL) is a distributed machine learning technique where multiple devices (such as smartphones or IoT devices) train a shared global model by using their local data. FL claims that the data privacy of local participants is preserved well because local data will not be shared with either the server-side or other training participants. However, this paper discovers a pioneering finding that a model inversion (MI) attacker, who acts as a benign participant, can invert the shared global model and obtain the data belonging to other participants. This will lead to severe data-leakage risk in FL because it is difficult to

identify attackers from benign participants.
In addition, we found even the most advanced defense approaches could not effectively address this issue. Therefore, it is important to evaluate such data-leakage risks of an FL system before using it. To alleviate this issue, we propose FedInverse to evaluate whether the FL global model can be inverted by MI attackers. In particular, FedInverse can be optimized by leveraging the Hilbert-Schmidt independence criterion (HSIC) as a regularizer to adjust the diversity of the MI attack generator. We test FedInverse with three typical MI attackers, GMI, KED-MI, and VMI, and the experiments show our FedInverse method can successfully obtain the data belonging to other participants. The code of this work is available at https://github.com/Jun-B0518/FedInverse
****
Yunhui Jang,Seul Lee,Sungsoo Ahn
A Simple and Scalable Representation for Graph Generation
Recently, there has been a surge of interest in employing neural networks for graph generation, a fundamental statistical learning problem with critical applications like molecule design and community analysis. However, most approaches encounter significant limitations when generating large-scale graphs. This is due to their requirement to output the full adjacency matrices whose size grows quadratically with the number of nodes. In response to this challenge, we introduce a new, simple, and scalable graph representation named gap encoded edge list (GEEL) that has a small representation size that aligns with the number of edges. In addition, GEEL significantly reduces the vocabulary size by incorporating the gap encoding and bandwidth restriction schemes. GEEL can be autoregressively generated with the incorporation of node positional encoding, and we further extend GEEL to deal with attributed graphs by designing a new grammar. Our findings reveal that the adoption of this compact representation not only enhances scalability but also bolsters performance by simplifying the graph generation process. We conduct a comprehensive evaluation across ten non-attributed and two molecular graph generation tasks, demonstrating the effectiveness of GEEL.
****
Xun Jiang,zhuomin chai,Yuxiang Zhao,Yibo Lin,Runsheng Wang,Ru Huang
CircuitNet 2.0: An Advanced Dataset for Promoting Machine Learning Innovations in Realistic Chip Design Environment
Integrated circuits or chips are key to enable computing in modern industry. Designing a chip relies on human experts to produce chip data through professional electronic design automation (EDA) software and complicated procedures. Nowadays, prompted by the wide variety of machine learning (ML) datasets, we have witnessed great advancement of ML algorithms in computer vision, natural language processing, and other fields. However, in chip design, high human workload and data sensitivity cause the lack of public datasets, which hinders the progress of ML development for EDA. To this end, we introduce an advanced large-scale dataset, CircuitNet 2.0, which targets promoting ML innovations in a realistic chip design environment. In order to approach the realistic chip design space, we collect more than 10,000 samples with a variety of chip designs (e.g., CPU, GPU, and AI Chip). All the designs are conducted through complete commercial design flows in a widely-used technology node, 14nm FinFET. We collect comprehensive data, including routability, timing, and power, from the design flow to support versatile ML tasks in EDA. Besides, we also introduce some realistic ML tasks with CircuitNet 2.0 to verify the potential for boosting innovations.
****
Trung Trinh,Markus Heinonen,Luigi Acerbi,Samuel Kaski
Input-gradient space particle inference for neural network ensembles
Deep Ensembles (DEs) demonstrate improved accuracy, calibration and robustness to perturbations over single neural networks partly due to their functional diversity. Particle-based variational inference (ParVI) methods enhance diversity by formalizing a repulsion term based on a network similarity kernel. However, weight-space repulsion is inefficient due to over-parameterization, while direct function-space repulsion has been found to produce little improvement over DEs. To sidestep these difficulties, we propose First-order Repulsive Deep Ensemble (FoR

DE), an ensemble learning method based on ParVI, which performs repulsion in the space of first-order input gradients. As input gradients uniquely characterize a function up to translation and are much smaller in dimension than the weights, this method guarantees that ensemble members are functionally different. Intuitively, diversifying the input gradients encourages each network to learn different features, which is expected to improve the robustness of an ensemble. Experiments on image classification datasets and transfer learning tasks show that FoRDE significantly outperforms the gold-standard DEs and other ensemble methods in accuracy and calibration under covariate shift due to input perturbations.
****
Tokio Kajitsuka,Issei Sato
Are Transformers with One Layer Self-Attention Using Low-Rank Weight Matrices Universal Approximators?
Existing analyses of the expressive capacity of Transformer models have required excessively deep layers for data memorization, leading to a discrepancy with the Transformers actually used in practice.
This is primarily due to the interpretation of the softmax function as an approximation of the hardmax function.
By clarifying the connection between the softmax function and the Boltzmann operator, we prove that a single layer of self-attention with low-rank weight matrices possesses the capability to perfectly capture the context of an entire input sequence.
As a consequence, we show that one-layer and single-head Transformers have a memorization capacity for finite samples, and that Transformers consisting of one self-attention layer with two feed-forward neural networks are universal approximators for continuous functions on a compact domain.
****
Songyao Jin,Feng Xie,Guangyi Chen,Biwei Huang,Zhengming Chen,Xinshuai Dong,Kun Zhang
Structural Estimation of Partially Observed Linear Non-Gaussian Acyclic Model: A Practical Approach with Identifiability
Conventional causal discovery approaches, which seek to uncover causal relationships among measured variables, are typically fragile to the presence of latent variables. While various methods have been developed to address this confounding issue, they often rely on strong assumptions about the underlying causal structure. In this paper, we consider a general scenario where measured and latent variables collectively form a partially observed causally sufficient linear system and latent variables may be anywhere in the causal structure. We theoretically show that with the aid of high-order statistics, the causal graph is (almost) fully identifiable if, roughly speaking, each latent set has a sufficient number of pure children, which can be either latent or measured. Naturally, LiNGAM, a model without latent variables, is encompassed as a special case. Based on the identification theorem, we develop a principled algorithm to identify the causal graph by testing for statistical independence involving only measured variables in specific manners. Experimental results show that our method effectively recovers the causal structure, even when latent variables are influenced by measured variables.
****
Gabriel Cardoso,Yazid Janati el idrissi,Sylvain Le Corff,Eric Moulines
Monte Carlo guided Denoising Diffusion models for Bayesian linear inverse problems.
Ill-posed linear inverse problems arise frequently in various applications, from computational photography to medical imaging.
A recent line of research exploits Bayesian inference with informative priors to handle the ill-posedness of such problems.
Amongst such priors, score-based generative models (SGM) have recently been successfully applied to several different inverse problems.
In this study, we exploit the particular structure of the prior defined by the SGM to define a sequence of intermediate linear inverse problems. As the noise level decreases, the posteriors of these inverse problems get closer to the target

 posterior of the original inverse problem.
To sample from this sequence of posteriors, we propose the use of Sequential Mon
te Carlo (SMC) methods.
The proposed algorithm, \algo, is shown to be theoretically grounded and we prov
ide numerical simulations showing that it outperforms competing baselines when d
ealing with ill-posed inverse problems in a Bayesian setting.
****
Ling Yang,Zhilong Zhang,Zhaochen Yu,Jingwei Liu,Minkai Xu,Stefano Ermon,Bin CUI
Cross-Modal Contextualized Diffusion Models for Text-Guided Visual Generation an
d Editing
Conditional diffusion models have exhibited superior performance in high-fidelit
y text-guided visual generation and editing. Nevertheless, prevailing text-guide
d visual diffusion models primarily focus on incorporating text-visual relations
hips exclusively into the reverse process, often disregarding their relevance in
 the forward process. This inconsistency between forward and reverse processes m
ay
limit the precise conveyance of textual semantics in visual synthesis results. T
o address this issue, we propose a novel and general contextualized diffusion mo
del (ContextDiff) by incorporating the cross-modal context encompassing interact
ions and alignments between text condition and visual sample into forward and re
verse processes. We propagate this context to all timesteps in the two processes
 to adapt their trajectories, thereby facilitating cross-modal conditional model
ing. We generalize our contextualized diffusion to both DDPMs and DDIMs with the
oretical derivations, and demonstrate the effectiveness of our model in evaluati
ons with two challenging tasks: text-to-image generation, and text-to-video edit
ing. In each task, our ContextDiff achieves new state-of-the-art performance, si
gnificantly enhancing the semantic alignment between text condition and generate
d samples, as evidenced by quantitative and qualitative evaluations. Our code is
 available at https://github.com/YangLing0818/ContextDiff
****
Chuanhao Li,Chong Liu,Yu-Xiang Wang
Communication-Efficient Federated Non-Linear Bandit Optimization
Federated optimization studies the problem of collaborative function optimizatio
n among multiple clients (e.g. mobile devices or organizations) under the coordi
nation of a central server. Since the data is collected separately by each clien
t and always remains decentralized, federated optimization preserves data privac
y and allows for large-scale computing, which makes it a promising decentralized
 machine learning paradigm. Though it is often deployed for tasks that are onlin
e in nature, e.g., next-word prediction on keyboard apps, most works formulate i
t as an offline problem. The few exceptions that consider federated bandit optim
ization are limited to very simplistic function classes, e.g., linear, generaliz
ed linear, or non-parametric function class with bounded RKHS norm, which severe
ly hinders its practical usage. In this paper, we propose a new algorithm, named
 Fed-GO-UCB, for federated bandit optimization with generic non-linear objective
 function. Under some mild conditions, we rigorously prove that Fed-GO-UCB is ab
le to achieve sub-linear rate for both cumulative regret and communication cost.
 At the heart of our theoretical analysis are distributed regression oracle and
individual confidence set construction, which can be of independent interests. E
mpirical evaluations also demonstrate the effectiveness of the proposed algorith
m.
****
Yuan Gong,Hongyin Luo,Alexander H. Liu,Leonid Karlinsky,James R. Glass
Listen, Think, and Understand
The ability of artificial intelligence (AI) systems to perceive and comprehend a
udio signals is crucial for many applications. Although significant progress has
 been made in this area since the development of AudioSet, most existing models
are designed to map audio inputs to pre-defined, discrete sound label sets. In c
ontrast, humans possess the ability to not only classify sounds into general cat
egories, but also to listen to the finer details of the sounds, explain the reas
on for the predictions, think about what the sound infers, and understand the sc

ene and what action needs to be taken, if any. Such capabilities beyond perception are not yet present in existing audio models. On the other hand, modern large language models (LLMs) exhibit emerging reasoning ability but they lack audio perception capabilities. Therefore, we ask the question: can we build a model that has both audio perception and reasoning ability?

In this paper, we propose a new audio foundation model, called LTU (Listen, Think, and Understand). To train LTU, we created a new OpenAQA-5M dataset consisting of 1.9 million closed-ended and 3.7 million open-ended, diverse (audio, question, answer) tuples, and have used an autoregressive training framework with a perception-to-understanding curriculum. LTU demonstrates strong performance and generalization ability on conventional audio tasks such as classification and captioning. More importantly, it exhibits emerging audio reasoning and comprehension abilities that are absent in existing audio models. To the best of our knowledge, LTU is the first multimodal large language model that focuses on general audio (rather than just speech) understanding.
****

Junwoo Park,Daehoon Gwak,Jaegul Choo,Edward Choi

Self-Supervised Contrastive Forecasting

Long-term forecasting presents unique challenges due to the time and memory complexity of handling long sequences. Existing methods, which rely on sliding windows to process long sequences, struggle to effectively capture long-term variations that are partially caught within the short window (i.e., outer-window variations). In this paper, we introduce a novel approach that overcomes this limitation by employing contrastive learning and enhanced decomposition architecture, specifically designed to focus on long-term variations. To this end, our contrastive
loss incorporates global autocorrelation held in the whole time series, which facilitates the construction of positive and negative pairs in a self-supervised manner. When combined with our decomposition networks, our constrative learning significantly improves long-term forecasting performance. Extensive experiments demonstrate that our approach outperforms 14 baseline models on well-established nine long-term benchmarks, especially in challenging scenarios that require a significantly long output for forecasting. This paper not only presents a novel direction for long-term forecasting but also offers a more reliable method for effectively integrating long-term variations into time-series representation learning.
****

Yavuz Faruk Bakman,Duygu Nur Yaldiz,Yahya H. Ezzeldin,Salman Avestimehr

Federated Orthogonal Training: Mitigating Global Catastrophic Forgetting in Continual Federated Learning

Federated Learning (FL) has gained significant attraction due to its ability to enable privacy-preserving training over decentralized data. Current literature in FL mostly focuses on single-task learning. However, over time, new tasks may appear in the clients and the global model should learn these tasks without forgetting previous tasks. This real-world scenario is known as Continual Federated Learning (CFL). The main challenge of CFL is \textit{Global Catastrophic Forgetting}, which corresponds to the fact that when the global model is trained on new tasks, its performance on old tasks decreases. There have been a few recent works on CFL to propose methods that aim to address the global catastrophic forgetting problem. However, these works either have unrealistic assumptions on the availability of past data samples or violate the privacy principles of FL. We propose a novel method, Federated Orthogonal Training (FOT), to overcome these drawbacks and address the global catastrophic forgetting in CFL. Our algorithm extracts the global input subspace of each layer for old tasks and modifies the aggregated updates of new tasks such that they are orthogonal to the global principal subspace of old tasks for each layer. This decreases the interference between tasks, which is the main cause for forgetting. Our method is almost computation-free on the client side and has negligible communication cost. We empirically show that FOT outperforms state-of-the-art continual learning methods in the CFL setti

ng, achieving an average accuracy gain of up to 15% with 27% lower forgetting while only incurring a minimal computation and communication cost. Code can be found [here](https://github.com/duygunuryldz/Federated_Orthogonal_Training)
****
Athul Paul Jacob,Yikang Shen,Gabriele Farina,Jacob Andreas
The Consensus Game: Language Model Generation via Equilibrium Search
When applied to question answering and other text generation tasks, language models (LMs) may be queried generatively (by sampling answers from their output distribution) or discriminatively (by using them to score or rank a set of candidate answers). These procedures sometimes yield very different predictions. How do we reconcile mutually incompatible scoring procedures to obtain coherent LM predictions? We introduce a new, a training-free, game-theoretic procedure for language model decoding. Our approach casts language model decoding as a regularized imperfect-information sequential signaling game—which we term the concensus game—in which a generator seeks to communicate an abstract correctness parameter using natural language sentences to a discriminator. We develop computational procedures for finding approximate equilibria of this game, resulting in a decoding algorithm we call equilibrium-ranking. Applied to a large number of tasks (including reading comprehension, commonsense reasoning, mathematical problem-solving, and assistive dialog), equilibrium-ranking consistently improves performance over existing LM decoding procedures. These improvements are sometimes substantial—on multiple benchmarks, we observe that applying equilibrium-ranking to LLaMA-7B outperforms the much larger LLaMA-65B and PaLM-540B models.
****
Yuan Feng,Yukun Cao,Wang Hairu,Xike Xie,S Kevin Zhou
Mayfly: a Neural Data Structure for Graph Stream Summarization
A graph is a structure made up of vertices and edges used to represent complex relationships between entities, while a graph stream is a continuous flow of graph updates that convey evolving relationships between entities. The massive volume and high dynamism of graph streams promote research on data structures of graph summarization, which provides a concise and approximate view of graph streams with sub-linear space and linear construction time, enabling real-time graph analytics in various domains, such as social networking, financing, and cybersecurity.
In this work, we propose the Mayfly, the first neural data structure for summarizing graph streams. The Mayfly replaces handcrafted data structures with better accuracy and adaptivity.
To cater to practical applications, Mayfly incorporates two offline training phases.
During the larval phase, the Mayfly learns basic summarization abilities from automatically and synthetically constituted meta-tasks, and in the metamorphosis phase, it rapidly adapts to real graph streams via meta-tasks.
With specific configurations of information pathways, the Mayfly enables flexible support for miscellaneous graph queries, including edge, node, and connectivity queries.
Extensive empirical studies show that the Mayfly significantly outperforms its handcrafted competitors.
****
Qinhong Zhou,Sunli Chen,Yisong Wang,Haozhe Xu,Weihua Du,Hongxin Zhang,Yilun Du,Joshua B. Tenenbaum,Chuang Gan
HAZARD Challenge: Embodied Decision Making in Dynamically Changing Environments
Recent advances in high-fidelity virtual environments serve as one of the major driving forces for building intelligent embodied agents to perceive, reason and interact with the physical world. Typically, these environments remain unchanged unless agents interact with them. However, in real-world scenarios, agents might also face dynamically changing environments characterized by unexpected events and need to rapidly take action accordingly. To remedy this gap, we propose a new simulated embodied benchmark, called HAZARD, specifically designed to assess the decision-making abilities of embodied agents in dynamic situations. HAZARD consists of three unexpected disaster scenarios, including fire, flood, and wind,

and specifically supports the utilization of large language models (LLMs) to assist common sense reasoning and decision-making. This benchmark enables us to evaluate autonomous agents' decision-making capabilities across various pipelines, including reinforcement learning (RL), rule-based, and search-based methods in dynamically changing environments. As a first step toward addressing this challenge using large language models, we further develop an LLM-based agent and perform an in-depth analysis of its promise and challenge of solving these challenging tasks. HAZARD is available at https://vis-www.cs.umass.edu/hazard/.
****
Alessandro De Palma,Rudy R Bunel,Krishnamurthy Dj Dvijotham,M. Pawan Kumar,Robert Stanforth,Alessio Lomuscio
Expressive Losses for Verified Robustness via Convex Combinations
In order to train networks for verified adversarial robustness, it is common to over-approximate the worst-case loss over perturbation regions, resulting in networks that attain verifiability at the expense of standard performance.
As shown in recent work, better trade-offs between accuracy and robustness can be obtained by carefully coupling adversarial training with over-approximations.
We hypothesize that the expressivity of a loss function, which we formalize as the ability to span a range of trade-offs between lower and upper bounds to the worst-case loss through a single parameter (the over-approximation coefficient), is key to attaining state-of-the-art performance.
To support our hypothesis, we show that trivial expressive losses, obtained via convex combinations between adversarial attacks and IBP bounds, yield state-of-the-art results across a variety of settings in spite of their conceptual simplicity.
We provide a detailed analysis of the relationship between the over-approximation coefficient and performance profiles across different expressive losses, showing that, while expressivity is essential, better approximations of the worst-case loss are not necessarily linked to superior robustness-accuracy trade-offs.
****
Bohao PENG,Zhuotao Tian,Shu Liu,Ming-Chang Yang,Jiaya Jia
Scalable Language Model with Generalized Continual Learning
Continual learning has gained increasing importance as it facilitates the acquisition and refinement of scalable knowledge and skills in language models. However, existing methods typically encounter strict limitations and challenges in real-world scenarios, such as reliance on experience replay, optimization constraints, and inference task-ID. In this study, we introduce the Scalable Language Model (SLM) to overcome these limitations within a more challenging and generalized setting, representing a significant advancement toward practical applications for continual learning. Specifically, we propose the Joint Adaptive Re-Parameterization (JARe), integrated with Dynamic Task-related Knowledge Retrieval (DTKR), to enable adaptive adjustment of language models based on specific downstream tasks. This approach leverages the task distribution within the vector space, aiming to achieve a smooth and effortless continual learning process. Our method demonstrates state-of-the-art performance on diverse backbones and benchmarks, achieving effective continual learning in both full-set and few-shot scenarios with minimal forgetting. Moreover, while prior research primarily focused on a single task type such as classification, our study goes beyond, with the large language model, i.e., LLaMA-2, to explore the effects across diverse domains and task types, such that a single language model can be decently scaled to broader applications. The code and models will be released to the public.
****
Niklas Muennighoff,Qian Liu,Armel Randy Zebaze,Qinkai Zheng,Binyuan Hui,Terry Yue Zhuo,Swayam Singh,Xiangru Tang,Leandro Von Werra,Shayne Longpre
OctoPack: Instruction Tuning Code Large Language Models
Finetuning large language models (LLMs) on instructions leads to vast performance improvements on natural language tasks. We apply instruction tuning using code, leveraging the natural structure of Git commits, which pair code changes with human instructions. We compile CommitPack: 4 terabytes of Git commits across 350 programming languages. We benchmark CommitPack against other natural and synthe

tic code instructions (xP3x, Self-Instruct, OASST) on the 16B parameter StarCode
r model, and achieve state-of-the-art performance among models not trained on Op
enAI outputs, on the HumanEval Python benchmark (46.2% pass@1). We further intro
duce HumanEvalPack, expanding the HumanEval benchmark to a total of 3 coding tas
ks (Code Repair, Code Explanation, Code Synthesis) across 6 languages (Python, J
avaScript, Java, Go, C++, Rust). Our models, OctoCoder and OctoGeeX, achieve the
 best performance across HumanEvalPack among all permissive models, demonstratin
g CommitPack's benefits in generalizing to a wider set of languages and natural
coding tasks. Code, models and data are freely available at https://github.com/b
igcode-project/octopack.
****
Ziyue Jiang,Jinglin Liu,Yi Ren,Jinzheng He,Zhenhui Ye,Shengpeng Ji,Qian Yang,Che
n Zhang,Pengfei Wei,Chunfeng Wang,Xiang Yin,Zejun MA,Zhou Zhao
Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis
Zero-shot text-to-speech (TTS) aims to synthesize voices with unseen speech prom
pts, which significantly reduces the data and computation requirements for voice
 cloning by skipping the fine-tuning process. However, the prompting mechanisms
of zero-shot TTS still face challenges in the following aspects: 1) previous wor
ks of zero-shot TTS are typically trained with single-sentence prompts, which si
gnificantly restricts their performance when the data is relatively sufficient d
uring the inference stage. 2) The prosodic information in prompts is highly coup
led with timbre, making it untransferable to each other.
This paper introduces Mega-TTS 2, a generic prompting mechanism for zero-shot TT
S, to tackle the aforementioned challenges. Specifically, we design a powerful a
coustic autoencoder that separately encodes the prosody and timbre information i
nto the compressed latent space while providing high-quality reconstructions. Th
en, we propose a multi-reference timbre encoder and a prosody latent language mo
del (P-LLM) to extract useful information from multi-sentence prompts. We furthe
r leverage the probabilities derived from multiple P-LLM outputs to produce tran
sferable and controllable prosody.
Experimental results demonstrate that Mega-TTS 2 could not only synthesize ident
ity-preserving speech with a short prompt of an unseen speaker from arbitrary so
urces but consistently outperform the fine-tuning method when the volume of data
 ranges from 10 seconds to 5 minutes. Furthermore, our method enables to transfe
r various speaking styles to the target timbre in a fine-grained and controlled
manner. Audio samples can be found in https://boostprompt.github.io/boostprompt/
.
****
Kimia Hamidieh,Haoran Zhang,Swami Sankaranarayanan,Marzyeh Ghassemi
Views Can Be Deceiving: Improved SSL Through Feature Space Augmentation
Supervised learning methods have been found to exhibit inductive biases favoring
 simpler features. When such features are spuriously correlated with the label,
this can result in suboptimal performance on minority subgroups. Despite the gro
wing popularity of methods which learn from unlabeled data, the extent to which
these representations rely on spurious features for prediction is unclear. In th
is work, we explore the impact of spurious features on Self-Supervised Learning
(SSL) for visual representation learning. We first empirically show that commonl
y used augmentations in SSL can cause undesired invariances in the image space,
and illustrate this with a simple example. We further show that classical approa
ches in combating spurious correlations, such as dataset re-sampling during SSL,
 do not consistently lead to invariant representations. Motivated by these findi
ngs, we propose LateTVG to remove spurious information from these representation
s during pre-training, by regularizing later layers of the encoder via pruning.
We find that our method produces representations which outperform the baselines
on several benchmarks, without the need for group or label information during SS
L.
****
Zilinghan Li,Pranshu Chaturvedi,Shilan He,Han Chen,Gagandeep Singh,Volodymyr Kin
dratenko,Eliu A Huerta,Kibaek Kim,Ravi Madduri
FedCompass: Efficient Cross-Silo Federated Learning on Heterogeneous Client Devi

ces Using a Computing Power-Aware Scheduler
Cross-silo federated learning offers a promising solution to collaboratively train robust and generalized AI models without compromising the privacy of local datasets, e.g., healthcare, financial, as well as scientific projects that lack a centralized data facility. Nonetheless, because of the disparity of computing resources among different clients (i.e., device heterogeneity), synchronous federated learning algorithms suffer from degraded efficiency when waiting for straggler clients. Similarly, asynchronous federated learning algorithms experience degradation in the convergence rate and final model accuracy on non-identically and independently distributed (non-IID) heterogeneous datasets due to stale local models and client drift. To address these limitations in cross-silo federated learning with heterogeneous clients and data, we propose FedCompass, an innovative semi-asynchronous federated learning algorithm with a computing power-aware scheduler on the server side, which adaptively assigns varying amounts of training tasks to different clients using the knowledge of the computing power of individual clients. FedCompass ensures that multiple locally trained models from clients are received almost simultaneously as a group for aggregation, effectively reducing the staleness of local models. At the same time, the overall training process remains asynchronous, eliminating prolonged waiting periods from straggler clients. Using diverse non-IID heterogeneous distributed datasets, we demonstrate that FedCompass achieves faster convergence and higher accuracy than other asynchronous algorithms while remaining more efficient than synchronous algorithms when performing federated learning on heterogeneous clients. The source code for FedCompass is available at https://github.com/APPFL/FedCompass.
****
Tianwei Ni,Benjamin Eysenbach,Erfan SeyedSalehi,Michel Ma,Clement Gehring,Aditya Mahajan,Pierre-Luc Bacon
Bridging State and History Representations: Understanding Self-Predictive RL
Representations are at the core of all deep reinforcement learning (RL) methods for both Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs). Many representation learning methods and theoretical frameworks have been developed to understand what constitutes an effective representation. However, the relationships between these methods and the shared properties among them remain unclear. In this paper, we show that many of these seemingly distinct methods and frameworks for state and history abstractions are, in fact, based on a common idea of self-predictive abstraction. Furthermore, we provide theoretical insights into the widely adopted objectives and optimization, such as the stop-gradient technique, in learning self-predictive representations. These findings together yield a minimalist algorithm to learn self-predictive representations for states and histories. We validate our theories by applying our algorithm to standard MDPs, MDPs with distractors, and POMDPs with sparse rewards. These findings culminate in a set of preliminary guidelines for RL practitioners.
****
Xuefei Ning,Zinan Lin,Zixuan Zhou,Zifu Wang,Huazhong Yang,Yu Wang
Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation
This work aims at decreasing the end-to-end generation latency of large language models (LLMs). One of the major causes of the high generation latency is the sequential decoding approach adopted by almost all state-of-the-art LLMs. In this work, motivated by the thinking and writing process of humans, we propose Skeleton-of-Thought (SoT), which first guides LLMs to generate the skeleton of the answer, and then conducts parallel API calls or batched decoding to complete the contents of each skeleton point in parallel. Not only does SoT provide considerable speed-ups across 12 LLMs, but it can also potentially improve the answer quality on several question categories. SoT is an initial attempt at data-centric optimization for inference efficiency, and showcases the potential of eliciting high-quality answers by explicitly planning the answer structure in language.
****
Zhangyang Gao,Cheng Tan,Xingran Chen,Yijie Zhang,Jun Xia,Siyuan Li,Stan Z. Li
KW-Design: Pushing the Limit of Protein Design via Knowledge Refinement

Recent studies have shown competitive performance in protein inverse folding, wh
ile most of them disregard the importance of predictive confidence, fail to cove
r the vast protein space, and do not incorporate common protein knowledge. Given
 the great success of pretrained models on diverse protein-related tasks and the
 fact that recovery is highly correlated with confidence, we wonder whether this
 knowledge can push the limits of protein design further. As a solution, we prop
ose a knowledge-aware module that refines low-quality residues. We also introduc
e a memory-retrieval mechanism to save more than 50\% of the training time. We e
xtensively evaluate our proposed method on the CATH, TS50, TS500, and PDB datase
ts and our results show that our KW-Design method outperforms the previous PiFol
d method by approximately 9\% on the CATH dataset. KW-Design is the first method
 that achieves 60+\% recovery on all these benchmarks. We also provide additiona
l analysis to demonstrate the effectiveness of our proposed method. The code is
publicly available via \href{https://github.com/A4Bio/ProteinInvBench}{GitHub}.
****
Ge Li,Hongyi Zhou,Dominik Roth,Serge Thilges,Fabian Otto,Rudolf Lioutikov,Gerhar
d Neumann
Open the Black Box: Step-based Policy Updates for Temporally-Correlated Episodic
 Reinforcement Learning
Current advancements in reinforcement learning (RL) have predominantly focused o
n learning step-based policies that generate actions for each perceived state. W
hile these methods efficiently leverage step information from environmental inte
raction, they often ignore the temporal correlation between actions, resulting i
n inefficient exploration and unsmooth trajectories that are challenging to impl
ement on real hardware. Episodic RL (ERL) seeks to overcome these challenges by
exploring in parameters space that capture the correlation of actions. However,
these approaches typically compromise data efficiency, as they treat trajectorie
s as opaque black boxes. In this work, we introduce a novel ERL algorithm, Tempo
rally-Correlated Episodic RL (TCE), which effectively utilizes step information
in episodic policy updates, opening the 'black box' in existing ERL methods whil
e retaining the smooth and consistent exploration in parameter space. TCE synerg
istically combines the advantages of step-based and episodic RL, achieving compa
rable performance to recent ERL methods while maintaining data efficiency akin t
o state-of-the-art (SoTA) step-based RL.
****
Lifeng Shen,Weiyu Chen,James Kwok
Multi-Resolution Diffusion Models for Time Series Forecasting
The diffusion model has been successfully used in many computer vision applicati
ons, such as text-guided image generation and image-to-image translation. Recent
ly, there have been attempts on extending the diffusion model for time series da
ta. However, these extensions are fairly straightforward and do not utilize the
unique properties of time series data. As different patterns are usually exhibit
ed at multiple scales of a time series, we in this paper leverage this multi-res
olution temporal structure and propose the multi-resolution diffusion model (mr-
Diff). By using the seasonal-trend decomposition, we sequentially extract fine-t
o-coarse trends from the time series for forward diffusion. The denoising proces
s then proceeds in an easy-to-hard non-autoregressive manner. The coarsest trend
 is generated first. Finer details are progressively added, using the predicted
coarser trends as condition variables. Experimental results on nine real-world t
ime series datasets demonstrate that mr-Diff outperforms state-of-the-art time s
eries diffusion models. It is also better than or comparable across a wide varie
ty of advanced time series prediction models.
****
Hongwei Wen,Annika Betken,Hanyuan Hang
Class Probability Matching with Calibrated Networks for Label Shift Adaption
We consider the domain adaptation problem in the context of label shift, where t
he label distributions between source and target domain differ, but the conditi
onal distributions of features given the label are the same. To solve the label
shift adaption problem, we develop a novel matching framework named \textit{clas
s probability matching} (\textit{CPM}). It is inspired by a new understanding of

the source domain's class probability, as well as a specific relationship between class probability ratios and feature probability ratios between the source and target domains. CPM is able to maintain the same theoretical guarantee with the existing feature probability matching framework, while significantly improving the computational efficiency due to directly matching the probabilities of the label variable. Within the CPM framework, we propose an algorithm named \textit{class probability matching with calibrated networks} (\textit{CPMCN}) for target domain classification. From the theoretical perspective, we establish the generalization bound of the CPMCN method in order to explain the benefits of introducing calibrated networks. From the experimental perspective, real data comparisons show that CPMCN outperforms existing matching-based and EM-based algorithms.
****
Fangyuan Xu,Weijia Shi,Eunsol Choi
RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation
Retrieval-augmented language models improve language models (LMs) by retrieving documents and prepending them in-context.
However, these documents, often spanning hundreds of words, make inference substantially less efficient. We propose compressing the retrieved documents into textual summaries prior to in-context integration. This not only reduces the computational costs but also relieve the burden of LMs to identify relevant information in long retrieved documents. We present two compressors -- an extractive compressor which selects useful sentences from retrieved documents and an abstractive compressor which generates summary by synthesizing information from multiple documents. Both are trained to achieve performance gain in LMs when we prepend the generated summary from the compressor to LMs' input, while minimizing the summary length. When retrieved documents are irrelevant to the input or offer no additional information to LM, our compressors output an empty string, enabling selective augmentation. We evaluate our approach on the language modeling task and open domain question answering task. We achieve a compression rate of as low as 6% with minimal loss in performance for both tasks, significantly outperforming the off-the-shelf summarization models. We show that our compressors trained for one LM can transfer to other LMs on the language modeling task and provide a summary largely faithful to the retrieved documents.
****
Shengjie Luo,Tianlang Chen,Aditi S. Krishnapriyan
Enabling Efficient Equivariant Operations in the Fourier Basis via Gaunt Tensor Products
Developing equivariant neural networks for the E(3) group plays an important role in modeling 3D data across real-world applications. Enforcing this equivariance primarily involves the tensor products of irreducible representations (irreps). However, the computational complexity of such operations increases significantly as higher-order tensors are used. In this work, we propose a systematic approach to substantially accelerate the computation of the tensor products of irreps. We mathematically connect the commonly used Clebsch-Gordan coefficients to the Gaunt coefficients, which are integrals of products of three spherical harmonics. Through Gaunt coefficients, the tensor product of irreps becomes equivalent to the multiplication between spherical functions represented by spherical harmonics. This perspective further allows us to change the basis for the equivariant operations from spherical harmonics to a 2D Fourier basis. Consequently, the multiplication between spherical functions represented by a 2D Fourier basis can be efficiently computed via the convolution theorem and Fast Fourier Transforms. This transformation reduces the complexity of full tensor products of irreps from $\mathcal{O}(L^6)$ to $\mathcal{O}(L^3)$, where L is the max degree of irreps. Leveraging this approach, we introduce the Gaunt Tensor Product, which serves as a new method to construct efficient equivariant operations across different model architectures. Our experiments on the Open Catalyst Project and 3BPA datasets demonstrate both the increased efficiency and improved performance of our approach. The code and models will be made publicly available at https://github.com/lsj2408/Gaunt-Tensor-Product.

****
Hansam Cho,Jonghyun Lee,Seoung Bum Kim,Tae-Hyun Oh,Yonghyun Jeong
Noise Map Guidance: Inversion with Spatial Context for Real Image Editing
Text-guided diffusion models have become a popular tool in image synthesis, know
n for producing high-quality and diverse images. However, their application to e
diting real images often encounters hurdles primarily due to the text condition
deteriorating the reconstruction quality and subsequently affecting editing fide
lity. Null-text Inversion (NTI) has made strides in this area, but it fails to c
apture spatial context and requires computationally intensive per-timestep optim
ization. Addressing these challenges, we present Noise Map Guidance (NMG), an in
version method rich in a spatial context, tailored for real-image editing. Signi
ficantly, NMG achieves this without necessitating optimization, yet preserves th
e editing quality. Our empirical investigations highlight NMG's adaptability acr
oss various editing techniques and its robustness to variants of DDIM inversions
.
****
Naman Jain,Tianjun Zhang,Wei-Lin Chiang,Joseph E. Gonzalez,Koushik Sen,Ion Stoic
a
Improving Code Style for Accurate Code Generation
Natural language to code generation is an important application area of LLMs and
 has received wide attention from the community.
The majority of relevant studies have exclusively concentrated on increasing the
 quantity and functional correctness of training sets while disregarding other s
tylistic elements of programs. More recently, data quality has garnered a lot of
 interest and multiple works have showcased its importance for improving perform
ance. In this work, we investigate data quality for code and find that making th
e code more structured and readable leads to improved code generation performanc
e of the system. We build a novel data-cleaning pipeline that uses these princip
les to transform existing programs by 1.) renaming variables, 2.) modularizing a
nd decomposing complex code into smaller helper sub-functions, and 3.) inserting
 natural-language based planning annotations. We evaluate our approach on two ch
allenging algorithmic code generation benchmarks and find that fine-tuning CodeL
LaMa-7B on our transformed programs improves the performance by up to \textbf{30
\%} compared to fine-tuning on the original dataset. Additionally, we demonstrat
e improved performance from using a smaller amount of higher-quality data, findi
ng that a model fine-tuned on the entire original dataset is outperformed by a m
odel trained on one-eighth of our cleaned dataset. Even in comparison to closed-
source models, our models outperform the much larger AlphaCode models.
****
Tri Dao
FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning
Scaling Transformers to longer sequence lengths has been a major problem in the
last several years, promising to improve performance in language modeling and hi
gh-resolution image understanding, as well as to unlock new applications in code
, audio, and video generation. The attention layer is the main bottleneck in sca
ling to longer sequences, as its runtime and memory increase quadratically in th
e sequence length. FlashAttention [5] exploits the asymmetric GPU memory hierarc
hy to bring significant memory saving (linear instead of quadratic) and runtime
speedup (2-4× compared to optimized baselines), with no approximation. However,
FlashAttention is still not nearly as fast as optimized matrix-multiply (GEMM) o
perations, reaching only 25-40% of the theoretical maximum FLOPs/s. We observe t
hat the inefficiency is due to suboptimal work partitioning between different th
read blocks and warps on the GPU, causing either low-occupancy or unnecessary sh
ared memory reads/writes. We propose FlashAttention-2, with better work partitio
ning to address these issues. In particular, we (1) tweak the algorithm to reduc
e the number of non-matmul FLOPs (2) parallelize the attention computation, even
 for a single head, across different thread blocks to increase occupancy, and (3
) within each thread block, distribute the work between warps to reduce communic
ation through shared memory. These yield around 2× speedup compared to FlashAtte
ntion, reaching 50-73% of the theoretical maximum FLOPs/s on A100 and getting cl

ose to the efficiency of GEMM operations. We empirically validate that when used end-to-end to train GPT-style models, FlashAttention-2 reaches training speed of up to 225 TFLOPs/s per A100 GPU (72% model FLOPs utilization).
****
Quanrui Rao,Lin Wang,Wuying Liu
Rethinking CNN’s Generalization to Backdoor Attack from Frequency Domain
Convolutional neural network (CNN) is easily affected by backdoor injections, whose models perform normally on clean samples but produce specific outputs on poisoned ones. Most of the existing studies have focused on the effect of trigger feature changes of poisoned samples on model generalization in spatial domain. We focus on the mechanism of CNN memorize poisoned samples in frequency domain, and find that CNN generate generalization to poisoned samples by memorizing the frequency domain distribution of trigger changes. We also explore the influence of trigger perturbations in different frequency domain components on the generalization of poisoned models from visible and invisible backdoor attacks, and prove that high-frequency components are more susceptible to perturbations than low-frequency components. Based on the above fundings, we propose a universal invisible strategy for visible triggers, which can achieve trigger invisibility while maintaining raw attack performance. We also design a novel frequency domain backdoor attack method based on low-frequency semantic information, which can achieve 100\% attack accuracy on multiple models and multiple datasets, and can bypass multiple defenses.
****
Goro Kobayashi,Tatsuki Kuribayashi,Sho Yokoi,Kentaro Inui
Analyzing Feed-Forward Blocks in Transformers through the Lens of Attention Map
Given that Transformers are ubiquitous in wide tasks, interpreting their internals is a pivotal issue.
Still, their particular components, feed-forward (FF) blocks, have typically been less analyzed despite their substantial parameter amounts.
We analyze the input contextualization effects of FF blocks by rendering them in the attention maps as a human-friendly visualization scheme.
Our experiments with both masked- and causal-language models reveal that
FF networks modify the input contextualization to emphasize specific types of linguistic compositions.
In addition, FF and its surrounding components tend to cancel out each other's effects, suggesting potential redundancy in the processing of the Transformer layer.
****
Shengjie Zhou,Lue Tao,Yuzhou Cao,Tao Xiang,Bo An,Lei Feng
On the Vulnerability of Adversarially Trained Models Against Two-faced Attacks
Adversarial robustness is an important standard for measuring the quality of learned models, and adversarial training is an effective strategy for improving the adversarial robustness of models. In this paper, we disclose that adversarially trained models are vulnerable to two-faced attacks, where slight perturbations in input features are crafted to make the model exhibit a false sense of robustness in the verification phase. Such a threat is significantly important as it can mislead our evaluation of the adversarial robustness of models, which could cause unpredictable security issues when deploying substandard models in reality. More seriously, this threat seems to be pervasive and tricky: we find that many types of models suffer from this threat, and models with higher adversarial robustness tend to be more vulnerable. Furthermore, we provide the first attempt to formulate this threat, disclose its relationships with adversarial risk, and try to circumvent it via a simple countermeasure. These findings serve as a crucial reminder for practitioners to exercise caution in the verification phase, urging them to refrain from blindly trusting the exhibited adversarial robustness of models.
****
Andrew Luo,Margaret Marie Henderson,Michael J. Tarr,Leila Wehbe
BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity
Understanding the functional organization of higher visual cortex is a central f

ocus in neuroscience. Past studies have primarily mapped the visual and semantic selectivity of neural populations using hand-selected stimuli, which may potentially bias results towards pre-existing hypotheses of visual cortex functionality. Moving beyond conventional approaches, we introduce a data-driven method that generates natural language descriptions for images predicted to maximally activate individual voxels of interest. Our method -- Semantic Captioning Using Brain Alignments ("BrainSCUBA") -- builds upon the rich embedding space learned by a contrastive vision-language model and utilizes a pre-trained large language model to generate interpretable captions. We validate our method through fine-grained voxel-level captioning across higher-order visual regions. We further perform text-conditioned image synthesis with the captions, and show that our images are semantically coherent and yield high predicted activations. Finally, to demonstrate how our method enables scientific discovery, we perform exploratory investigations on the distribution of "person" representations in the brain, and discover fine-grained semantic selectivity in body-selective areas. Unlike earlier studies that decode text, our method derives *voxel-wise captions of semantic selectivity*. Our results show that BrainSCUBA is a promising means for understanding functional preferences in the brain, and provides motivation for further hypothesis-driven investigation of visual cortex.
****
Minyoung Kim,Timothy Hospedales
A Hierarchical Bayesian Model for Few-Shot Meta Learning
We propose a novel hierarchical Bayesian model for the few-shot meta learning problem. We consider episode-wise random variables to model episode-specific generative processes, where these local random variables are governed by a higher-level global random variable. The global variable captures information shared across episodes, while controlling how much the model needs to be adapted to new episodes in a principled Bayesian manner. Within our framework, prediction on a novel episode/task can be seen as a Bayesian inference problem. For tractable training, we need to be able to relate each local episode-specific solution to the global higher-level parameters. We propose a Normal-Inverse-Wishart model, for which establishing this local-global relationship becomes feasible due to the approximate closed-form solutions for the local posterior distributions. The resulting algorithm is more attractive than the MAML in that it does not maintain a costly computational graph for the sequence of gradient descent steps in an episode. Our approach is also different from existing Bayesian meta learning methods in that rather than modeling a single random variable for all episodes, it leverages a hierarchical structure that exploits the local-global relationships desirable for principled Bayesian learning with many related tasks.
****
Hanan Gani,Shariq Farooq Bhat,Muzammal Naseer,Salman Khan,Peter Wonka
LLM Blueprint: Enabling Text-to-Image Generation with Complex and Detailed Prompts
Diffusion-based generative models have significantly advanced text-to-image generation but encounter challenges when processing lengthy and intricate text prompts describing complex scenes with multiple objects. While excelling in generating images from short, single-object descriptions, these models often struggle to faithfully capture all the nuanced details within longer and more elaborate textual inputs. In response, we present a novel approach leveraging Large Language Models (LLMs) to extract critical components from text prompts, including bounding box coordinates for foreground objects, detailed textual descriptions for individual objects, and a succinct background context. These components form the foundation of our layout-to-image generation model, which operates in two phases. The initial Global Scene Generation utilizes object layouts and background context to create an initial scene but often falls short in faithfully representing object characteristics as specified in the prompts. To address this limitation, we introduce an Iterative Refinement Scheme that iteratively evaluates and refines box-level content to align them with their textual descriptions, recomposing objects as needed to ensure consistency. Our evaluation on complex prompts featuring multiple objects demonstrates a substantial improvement in recall compared to

 baseline diffusion models. This is further validated by a user study, underscor
ing the efficacy of our approach in generating coherent and detailed scenes from
 intricate textual inputs. Our iterative framework offers a promising solution f
or enhancing text-to-image generation models' fidelity with lengthy, multifacete
d descriptions, opening new possibilities for accurate and diverse image synthes
is from textual inputs.
****
Mustafa Shukor,Alexandre Rame,Corentin Dancette,Matthieu Cord
Beyond task performance: evaluating and reducing the flaws of large multimodal m
odels with in-context-learning
Following the success of Large Language Models (LLMs), Large Multimodal Models (
LMMs), such as the Flamingo model and its subsequent competitors, have started t
o emerge as natural steps towards generalist agents. However, interacting with r
ecent LMMs reveals major limitations that are hardly captured by the current eva
luation benchmarks. Indeed, task performances (e.g., VQA accuracy) alone do not
provide enough clues to understand their real capabilities, limitations, and to
which extent such models are aligned to human expectations. To refine our unders
tanding of those flaws, we deviate from the current evaluation paradigm, and (1)
 evaluate 10 recent open-source LMMs from 3B up to 80B parameter scale, on 5 di
fferent axes; hallucinations, abstention, compositionality, explainability and i
nstruction following. Our evaluation on these axes reveals major flaws in LMMs.
While the current go-to solution to align these models is based on training, suc
h as instruction tuning or RLHF, we rather (2) explore the training-free in-cont
ext learning (ICL) as a solution, and study how it affects these limitations. Ba
sed on our ICL study, (3) we push ICL further and propose new multimodal ICL var
iants such as; Multitask-ICL, Chain-of-Hindsight-ICL, and Self-Correcting-ICL. O
ur findings are as follows; (1) Despite their success, LMMs have flaws that rema
in unsolved with scaling alone. (2) The effect of ICL on LMMs flaws is nuanced;
despite its effectiveness for improved explainability, answer abstention, ICL on
ly slightly improves instruction following, does not improve compositional abili
ties, and actually even amplifies hallucinations. (3) The proposed ICL variants
are promising as post-hoc approaches to efficiently tackle some of those flaws.
The code is available here: https://github.com/mshukor/EvALign-ICL.
****
Xuhui Zhou,Hao Zhu,Leena Mathur,Ruohong Zhang,Haofei Yu,Zhengyang Qi,Louis-Phili
ppe Morency,Yonatan Bisk,Daniel Fried,Graham Neubig,Maarten Sap
SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents
Humans are social beings; we pursue social goals in our daily interactions, wh
ich is a crucial aspect of social intelligence. Yet, AI systems' abilities in th
is realm remain elusive. We present SOTOPIA, an open-ended environment to simula
te complex social interactions between artificial agents and evaluate their soci
al intelligence. In our environment, agents role-play and *interact* under a wid
e variety of scenarios; they coordinate, collaborate, exchange, and compete with
 each other to achieve complex social goals. We simulate the role-play interacti
on between LLM-based agents and humans within this task space and evaluate their
 performance with a holistic evaluation framework called SOTOPIA-Eval. With SOTO
PIA, we find significant differences between these models in terms of their soci
al intelligence, and we identify a subset of SOTOPIA scenarios, SOTOPIA-hard, th
at is generally challenging for all models. We find that on this subset, GPT-4 a
chieves a significantly lower goal completion rate than humans and struggles to
exhibit social commonsense reasoning and strategic communication skills. These f
indings demonstrate SOTOPIA's promise as a general platform for research on eval
uating and improving social intelligence in artificial agents.
****
Quan Sun,Qiying Yu,Yufeng Cui,Fan Zhang,Xiaosong Zhang,Yueze Wang,Hongcheng Gao,
Jingjing Liu,Tiejun Huang,Xinlong Wang
Emu: Generative Pretraining in Multimodality
We present Emu, a multimodal foundation model that seamlessly generates images a
nd text in multimodal context. This omnivore model can take in any single-modali
ty or multimodal data input indiscriminately (e.g., interleaved image, text and

video) through a one-model-for-all autoregressive training process. First, visua l signals are encoded into embeddings, and together with text tokens form an int erleaved input sequence. Emu is end-to-end trained with a unified objective of c lassifying the next text token or regressing the next visual embedding in the mu ltimodal sequence. This versatile multimodality empowers the leverage of diverse pretraining data sources at scale, such as videos with interleaved frames and t ext, webpages with interleaved images and text, as well as web-scale image-text pairs and video-text pairs. Emu can serve as a generalist multimodal interface f or both image-to-text and text-to-image tasks, supporting in-context image and t ext generation. Across a broad range of zero-shot/few-shot tasks including image captioning, visual question answering, video question answering and text-to-ima ge generation, Emu demonstrates superb performance compared to state-of-the-art large multimodal models. Extended capabilities such as multimodal assistants via instruction tuning are also demonstrated with impressive performance.
****

Ming Zhong,Chenxin An,Weizhu Chen,Jiawei Han,Pengcheng He
Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Param etric Perspective
Large Language Models (LLMs) inherently encode a wealth of knowledge within thei r parameters through pre-training on extensive corpora. While prior research has delved into operations on these parameters to manipulate the underlying implici t knowledge—encompassing detection, editing, and merging—there remains an ambigu ous understanding regarding their transferability across models with varying sca les. In this paper, we seek to empirically investigate knowledge transfer from l arger to smaller models through a parametric perspective. To achieve this, we em ploy sensitivity-based techniques to extract and align knowledge-specific parame ters between different LLMs. Moreover, the LoRA module is used as the intermedia ry mechanism for injecting the extracted knowledge into smaller models. Evaluati ons across four benchmarks validate the efficacy of our proposed method. Our fin dings highlight the critical factors contributing to the process of parametric k nowledge transfer, underscoring the transferability of model parameters across L LMs of different scales.
****

Shih-Hsin Wang,Yung-Chang Hsu,Justin Baker,Andrea L. Bertozzi,Jack Xin,Bao Wang
Rethinking the Benefits of Steerable Features in 3D Equivariant Graph Neural Net works
Theoretical and empirical comparisons have been made to assess the expressive po wer and performance of invariant and equivariant GNNs. However, there is current ly no theoretical result comparing the expressive power of k-hop invariant GNN s and equivariant GNNs. Additionally, little is understood about whether the per formance of equivariant GNNs, employing steerable features up to type-L, incre ases as L grows -- especially when the feature dimension is held constant. In this study, we introduce a key lemma that allows us to analyze steerable feature s by examining their corresponding invariant features. The lemma facilitates us in understanding the limitations of k-hop invariant GNNs, which fail to captur e the global geometric structure due to the loss of geometric information betwee n local structures. Furthermore, we investigate the invariant features associate d with different types of steerable features and demonstrate that the expressive ness of steerable features is primarily determined by their dimension -- indepen dent of their irreducible decomposition. This suggests that when the feature dim ension is constant, increasing L does not lead to essentially improved perform ance in equivariant GNNs employing steerable features up to type-L. We substan tiate our theoretical insights with numerical evidence.
****

Milan Papez,Martin Rektoris,Vaclav Smidl,Tomáš Pevný
Sum-Product-Set Networks: Deep Tractable Models for Tree-Structured Graphs
Daily internet communication relies heavily on tree-structured graphs, embodied by popular data formats such as XML and JSON. However, many recent generative (p robabilistic) models utilize neural networks to learn a probability distribution over undirected cyclic graphs. This assumption of a generic graph structure bri

ngs various computational challenges, and, more importantly, the presence of non-linearities in neural networks does not permit tractable probabilistic inference. We address these problems by proposing sum-product-set networks, an extension of probabilistic circuits from unstructured tensor data to tree-structured graph data. To this end, we use random finite sets to reflect a variable number of nodes and edges in the graph and to allow for exact and efficient inference. We demonstrate that our tractable model performs comparably to various intractable models based on neural networks.
****
Sergei Solonets,Daniil Sinitsyn,Lukas Von Stumberg,Nikita Araslanov,Daniel Cremers
An Analytical Solution to Gauss-Newton Loss for Direct Image Alignment
Direct image alignment is a widely used technique for relative 6DoF pose estimation between two images, but its accuracy strongly depends on pose initialization.
Therefore, recent end-to-end frameworks increase the convergence basin of the learned feature descriptors with special training objectives, such as the Gauss-Newton loss.
However, the training data may exhibit bias toward a specific type of motion and pose initialization,
thus limiting the generalization of these methods.
In this work, we derive a closed-form solution to the expected optimum of the Gauss-Newton loss.
The solution is agnostic to the underlying feature representation and allows us to dynamically adjust the basin of convergence according to our assumptions about the uncertainty in the current estimates. These properties allow for effective control over the convergence in the alignment process.
Despite using self-supervised feature embeddings, our solution achieves compelling accuracy w.r.t. the state-of-the-art direct image alignment methods trained end-to-end with pose supervision, and demonstrates improved robustness to pose initialization.
Our analytical solution exposes some inherent limitations of end-to-end learning with the Gauss-Newton loss, and establishes an intriguing connection between direct image alignment and feature-matching approaches.
****
Yi-Lun Liao,Brandon M Wood,Abhishek Das,Tess Smidt
EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations
Equivariant Transformers such as Equiformer have demonstrated the efficacy of applying Transformers to the domain of 3D atomistic systems. However, they are limited to small degrees of equivariant representations due to their computational complexity. In this paper, we investigate whether these architectures can scale well to higher degrees. Starting from Equiformer, we first replace $SO(3)$ convolutions
with eSCN convolutions to efficiently incorporate higher-degree tensors. Then, to better leverage the power of higher degrees, we propose three architectural improvements – attention re-normalization, separable S^2 activation and separable layer normalization. Putting these all together, we propose EquiformerV2, which outperforms previous state-of-the-art methods on large-scale OC20 dataset by up to 9% on forces, 4% on energies, offers better speed-accuracy trade-offs, and 2$\times$ reduction in DFT calculations needed for computing adsorption energies. Additionally, EquiformerV2 trained on only OC22 dataset outperforms GemNet-OC trained on both OC20 and OC22 datasets, achieving much better data efficiency. Finally, we compare EquiformerV2 with Equiformer on QM9 and OC20 S2EF-2M
datasets to better understand the performance gain brought by higher degrees.
****
Longkang Li,Ignavier Ng,Gongxu Luo,Biwei Huang,Guangyi Chen,Tongliang Liu,Bin Gu,Kun Zhang
Federated Causal Discovery from Heterogeneous Data
Conventional causal discovery methods rely on centralized data, which is inconsi

stent with the decentralized nature of data in many real-world situations. This discrepancy has motivated the development of federated causal discovery (FCD) approaches. However, existing FCD methods may be limited by their potentially restrictive assumptions of identifiable functional causal models or homogeneous data distributions, narrowing their applicability in diverse scenarios. In this paper, we propose a novel FCD method attempting to accommodate arbitrary causal models and heterogeneous data. We first utilize a surrogate variable corresponding to the client index to account for the data heterogeneity across different clients. We then develop a federated conditional independence test (FCIT) for causal skeleton discovery and establish a federated independent change principle (FICP) to determine causal directions. These approaches involve constructing summary statistics as a proxy of the raw data to protect data privacy. Owing to the nonparametric properties, FCIT and FICP make no assumption about particular functional forms, thereby facilitating the handling of arbitrary causal models. We conduct extensive experiments on synthetic and real datasets to show the efficacy of our method. The code is available at https://github.com/lokali/FedCDH.git.
****
Shoumik Saha,Wenxiao Wang,Yigitcan Kaya,Soheil Feizi,Tudor Dumitras
DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness
Machine Learning (ML) models have been utilized for malware detection for over two decades. Consequently, this ignited an ongoing arms race between malware authors and antivirus systems, compelling researchers to propose defenses for malware-detection models against evasion attacks. However, most if not all existing defenses against evasion attacks suffer from sizable performance degradation and/or can defend against only specific attacks, which makes them less practical in real-world settings. In this work, we develop a certified defense, DRSM (De-Randomized Smoothed MalConv), by redesigning the *de-randomized smoothing* technique for the domain of malware detection. Specifically, we propose a *window ablation* scheme to provably limit the impact of adversarial bytes while maximally preserving local structures of the executables. After showing how DRSM is theoretically robust against attacks with contiguous adversarial bytes, we verify its performance and certified robustness experimentally, where we observe only marginal accuracy drops as the cost of robustness. To our knowledge, we are the first to offer certified robustness in the realm of static detection of malware executables. More surprisingly, through evaluating DRSM against 9 empirical attacks of different types, we observe that the proposed defense is empirically robust to some extent against a diverse set of attacks, some of which even fall out of the scope of its original threat model. In addition, we collected $15.5K$ recent benign raw executables from diverse sources, which will be made public as a dataset called PACE (Publicly Accessible Collection(s) of Executables) to alleviate the scarcity of publicly available benign datasets for studying malware detection and provide future research with more representative data of the time. Our code and dataset are available at - https://github.com/ShoumikSaha/DRSM
****
Duanyi YAO,Songze Li,Ye XUE,Jin Liu
Constructing Adversarial Examples for Vertical Federated Learning: Optimal Client Corruption through Multi-Armed Bandit
Vertical federated learning (VFL), where each participating client holds a subset of data features, has found numerous applications in finance, healthcare, and IoT systems. However, adversarial attacks, particularly through the injection of adversarial examples (AEs), pose serious challenges to the security of VFL models. In this paper, we investigate such vulnerabilities through developing a novel attack to disrupt the VFL inference process, under a practical scenario where the adversary is able to *adaptively corrupt a subset of clients*. We formulate the problem of finding optimal attack strategies as an online optimization problem, which is decomposed into an inner problem of adversarial example generation (AEG) and an outer problem of corruption pattern selection (CPS). Specifically, we establish the equivalence between the formulated CPS problem and a multi-armed bandit (MAB) problem, and propose the Thompson sampling with Empirical maximum

reward (E-TS) algorithm for the adversary to efficiently identify the optimal subset of clients for corruption. The key idea of E-TS is to introduce an estimation of the expected maximum reward for each arm, which helps to specify a small set of *competitive arms*, on which the exploration for the optimal arm is performed. This significantly reduces the exploration space, which otherwise can quickly become prohibitively large as the number of clients increases. We analytically characterize the regret bound of E-TS, and empirically demonstrate its capability of efficiently revealing the optimal corruption pattern with the highest attack success rate, under various datasets of popular VFL tasks.
****
Karsten Roth,Lukas Thede,A. Sophia Koepke,Oriol Vinyals,Olivier J Henaff,Zeynep Akata
Fantastic Gains and Where to Find Them: On the Existence and Prospect of General Knowledge Transfer between Any Pretrained Model
Training deep networks requires various design decisions regarding for instance their architecture, data augmentation, or optimization. In this work, we find these training variations to result in networks learning unique feature sets from the data. Using public model libraries comprising thousands of models trained on canonical datasets like ImageNet, we observe that for arbitrary pairings of pretrained models, one model extracts significant data context unavailable in the other – independent of overall performance. Given any arbitrary pairing of pretrained models and no external rankings (such as separate test sets, e.g. due to data privacy), we investigate if it is possible to transfer such "complementary" knowledge from one model to another without performance degradation – a task made particularly difficult as additional knowledge can be contained in stronger, equiperformant or weaker models. Yet facilitating robust transfer in scenarios agnostic to pretrained model pairings would unlock auxiliary gains and knowledge fusion from any model repository without restrictions on model and problem specifics - including from weaker, lower-performance models. This work therefore provides an initial, in-depth exploration on the viability of such general-purpose knowledge transfer. Across large-scale experiments, we first reveal the shortcomings of standard knowledge distillation techniques, and then propose a much more general extension through data partitioning for successful transfer between nearly all pretrained models, which we show can also be done unsupervised. Finally, we assess both the scalability and impact of fundamental model properties on successful model-agnostic knowledge transfer.
****
Chengxing Jia,Chenxiao Gao,Hao Yin,Fuxiang Zhang,Xiong-Hui Chen,Tian Xu,Lei Yuan,Zongzhang Zhang,Zhi-Hua Zhou,Yang Yu
Policy Rehearsing: Training Generalizable Policies for Reinforcement Learning
Human beings can make adaptive decisions in a preparatory manner, i.e., by making preparations in advance, which offers significant advantages in scenarios where both online and offline experiences are expensive and limited. Meanwhile, current reinforcement learning methods commonly rely on numerous environment interactions but hardly obtain generalizable policies. In this paper, we introduce the idea of \textit{rehearsal} into policy optimization, where the agent plans for all possible outcomes in mind and acts adaptively according to actual responses from the environment. To effectively rehearse, we propose ReDM, an algorithm that generates a diverse and eligible set of dynamics models and then rehearse the policy via adaptive training on the generated model set. Rehearsal enables the policy to make decision plans for various hypothetical dynamics and to naturally generalize to previously unseen environments. Our experimental results demonstrate that ReDM is capable of learning a valid policy solely through rehearsal, even with \emph{zero} interaction data. We further extend ReDM to scenarios where limited or mismatched interaction data is available, and our experimental results reveal that ReDM produces high-performing policies compared to other offline RL baselines.
****
Yang bai,Xinxing Xu,Yong Liu,Salman Khan,Fahad Khan,Wangmeng Zuo,Rick Siow Mong Goh,Chun-Mei Feng

Sentence-level Prompts Benefit Composed Image Retrieval
Composed image retrieval (CIR) is the task of retrieving specific images by using a query that involves both a reference image and a relative caption. Most existing CIR models adopt the late-fusion strategy to combine visual and language features. Besides, several approaches have also been suggested to generate a pseudo-word token from the reference image, which is further integrated into the relative caption for CIR. However, these pseudo-word-based prompting methods have limitations when target image encompasses complex changes on reference image, e.g., object removal and attribute modification. In this work, we demonstrate that learning an appropriate sentence-level prompt for the relative caption (SPRC) is sufficient for achieving effective composed image retrieval. Instead of relying on pseudo- word-based prompts, we propose to leverage pretrained V-L models, e.g., BLIP-2, to generate sentence-level prompts. By concatenating the learned sentence-level prompt with the relative caption, one can readily use existing text-based image retrieval models to enhance CIR performance. Furthermore, we introduce both image-text contrastive loss and text prompt alignment loss to enforce the learning of suitable sentence-level prompts. Experiments show that our proposed method performs favorably against the state-of-the-art CIR methods on the Fashion-IQ and CIRR datasets.
****
Shitong Duan,Xiaoyuan Yi,Peng Zhang,Tun Lu,Xing Xie,Ning Gu
DENEVIL: TOWARDS DECIPHERING AND NAVIGATING THE ETHICAL VALUES OF LARGE LANGUAGE MODELS VIA INSTRUCTION LEARNING
Large Language Models (LLMs) have made unprecedented breakthroughs, yet their increasing integration into everyday life might raise societal risks due to generated unethical content. Despite extensive study on specific issues like bias, the intrinsic values of LLMs remain largely unexplored from a moral philosophy perspective. This work delves into ethical values utilizing Moral Foundation Theory. Moving beyond conventional discriminative evaluations with poor reliability, we propose DeNEVIL, a novel prompt generation algorithm tailored to dynamically exploit LLMs’ value vulnerabilities and elicit the violation of ethics in a generative manner, revealing their underlying value inclinations. On such a basis, we construct MoralPrompt, a high-quality dataset comprising 2,397 prompts covering 500+ value principles, and then benchmark the intrinsic values across a spectrum of LLMs. We discovered that most models are essentially misaligned, necessitating further ethical value alignment. In response, we develop VILMO, an in-context alignment method that substantially enhances the value compliance of LLM outputs by learning to generate appropriate value instructions, outperforming existing competitors. Our methods are suitable for black-box and open-source models, offering a promising initial step in studying the ethical values of LLMs.
****
Manley Roberts,Himanshu Thakur,Christine Herlihy,Colin White,Samuel Dooley
To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination
Recent claims about the impressive abilities of large language models (LLMs) are often supported by evaluating publicly available benchmarks.
Since LLMs train on wide swaths of the internet, this practice raises concerns of data contamination, i.e., evaluating on examples that are explicitly or implicitly included in the training data.
Data contamination remains notoriously challenging to measure and mitigate, even with partial attempts like controlled experimentation of training data, canary strings, or embedding similarities.
In this work, we conduct the first thorough longitudinal analysis of data contamination in LLMs by using the natural experiment of training cutoffs in GPT models to look at benchmarks released over time.
Specifically, we consider two code/mathematical problem-solving datasets, Codeforces and Project Euler, and find statistically significant trends among LLM pass rate vs. GitHub popularity and release date that provide strong evidence of contamination.
By open-sourcing our dataset, raw results, and evaluation framework, our work pa

ves the way for rigorous analyses of data contamination in modern models. We conclude with a discussion of best practices and future steps for publicly releasing benchmark in the age of LLMs which train on webscale data.
****

Rasool Fakoor,Jonas Mueller,Zachary Chase Lipton,Pratik Chaudhari,Alex Smola
Time-Varying Propensity Score to Bridge the Gap between the Past and Present
Real-world deployment of machine learning models is challenging because data evolves over time. While no model can work when data evolves in an arbitrary fashion, if there is some pattern to these changes, we might be able to design methods to address it. This paper addresses situations when data evolves gradually. We introduce a time-varying propensity score that can detect gradual shifts in the distribution of data which allows us to selectively sample past data to update the model---not just similar data from the past like that of a standard propensity score but also data that evolved in a similar fashion in the past. The time-varying propensity score is quite general: we demonstrate different ways of implementing it and evaluate it on a variety of problems ranging from supervised learning (e.g., image classification problems) where data undergoes a sequence of gradual shifts, to reinforcement learning tasks (e.g., robotic manipulation and continuous control) where data shifts as the policy or the task changes.
****

Weida Li,Yaoliang Yu
Faster Approximation of Probabilistic and Distributional Values via Least Squares
The family of probabilistic values, axiomatically-grounded and proposed in cooperative game theory, has recently received much attention in data valuation. However, it is often computationally expensive to compute exactly (exponential w.r.t. N, the number of data being valuated). Existing generic estimators cost $O(\frac{N^2}{\epsilon^2}\log\frac{N}{\delta})$ utility evaluations to achieve an (ϵ, δ)-approximation under the 2-norm, while faster estimators have been developed recently for special cases (e.g., empirically for the Shapley value and theoretically for the Banzhaf value). In this work, based on a connection between probabilistic values and least square regressions, we propose two generic estimators for the whole family of probabilistic values that both cost $O(\frac{N}{\epsilon^2}\log\frac{N}{\delta})$ utility evaluations, largely extending the scope of this currently best complexity bound. Moreover, we show that each distributional value, proposed by Ghorbani et al. (2020) to alleviate the inconsistency of probabilistic values when using distinct databases, can also be cast as optimizing a similar least square regression. This observation makes it the first-time theoretically-grounded to train value estimators such that the distributional value of each unseen data point can be evaluated in a single forward pass. Our experiments verify the faster convergence of our proposed estimators, and demonstrate the effectiveness at learning distributional values.
****

Hao Cheng,Qingsong Wen,Yang Liu,Liang Sun
RobustTSF: Towards Theory and Design of Robust Time Series Forecasting with Anomalies
Time series forecasting is an important and forefront task whose techniques have been applied to electricity forecasting, trajectory prediction, labor planning, etc. However, most of time series forecasting techniques assume that the training data is clean without anomalies. This assumption is unrealistic since the collected time series data can be contaminated in practice. The forecasting model will be inferior if it is directly trained by time series with anomalies. Thus it is essential to develop methods to automatically learn a robust forecasting model from the contaminated data. In this paper, we first statistically define three types of anomalies, then theoretically and experimentally analyze the loss robustness and sample robustness when these anomalies exist. Based on our analyses, we propose a simple and efficient algorithm to learn a robust forecasting model. Extensive experiments show that our method is highly robust and outperforms all existing approaches. The code is available at https://github.com/haochenglouis/RobustTSF.

****
Thomas Kleine Buening,Aadirupa Saha,Christos Dimitrakakis,Haifeng Xu
Bandits Meet Mechanism Design to Combat Clickbait in Online Recommendation
We study a strategic variant of the multi-armed bandit problem, which we coin the strategic click-bandit. This model is motivated by applications in online recommendation where the choice of recommended items depends on both the click-through rates and the post-click rewards. Like in classical bandits, rewards follow a fixed unknown distribution. However, we assume that the click-rate of each arm is chosen strategically by the arm (e.g., a host on Airbnb) in order to maximize the number of times it gets clicked. The algorithm designer does not know the post-click rewards nor the arms' actions (i.e., strategically chosen click-rates) in advance, and must learn both values over time. To solve this problem, we design an incentive-aware learning algorithm, UCB-S, which achieves two goals simultaneously: (a) incentivizing desirable arm behavior under uncertainty; (b) minimizing regret by learning unknown parameters. We approximately characterize all Nash equilibria of the arms under UCB-S and show a $\tilde{\mathcal{O}} (\sqrt{KT})$ regret bound uniformly in every equilibrium. We also show that incentive-unaware algorithms generally fail to achieve low regret in the strategic click-bandit. Finally, we support our theoretical results by simulations of strategic arm behavior which confirm the effectiveness and robustness of our proposed incentive design.
****
Yuan Gao,Rustem Islamov,Sebastian U Stich
EControl: Fast Distributed Optimization with Compression and Error Control
Modern distributed training relies heavily on communication compression to reduce the communication overhead. In this work, we study algorithms employing a popular class of contractive compressors in order to reduce communication overhead. However, the naive implementation often leads to unstable convergence or even exponential divergence due to the compression bias. Error Compensation (EC) is an extremely popular mechanism to mitigate the aforementioned issues during the training of models enhanced by contractive compression operators. Compared to the effectiveness of EC in the data homogeneous regime, the understanding of the practicality and theoretical foundations of EC in the data heterogeneous regime is limited. Existing convergence analyses typically rely on strong assumptions such as bounded gradients, bounded data heterogeneity, or large batch accesses, which are often infeasible in modern Machine Learning Applications. We resolve the majority of current issues by proposing EControl, a novel mechanism that can regulate error compensation by controlling the strength of the feedback signal. We prove fast convergence for EControl in standard strongly convex, general convex, and nonconvex settings without any additional assumptions on the problem or data heterogeneity. We conduct extensive numerical evaluations to illustrate the efficacy of our method and support our theoretical findings.
****
Mehdi Fatemi,Sindhu C. M. Gowda
A Dynamical View of the Question of Why
We address causal reasoning in multivariate time series data generated by stochastic processes. Existing approaches are largely restricted to static settings, ignoring the continuity and emission of variations across time. In contrast, we propose a learning paradigm that directly establishes causation between events in the course of time. We present two key lemmas to compute causal contributions and frame them as reinforcement learning problems. Our approach offers formal and computational tools for uncovering and quantifying causal relationships in diffusion processes, subsuming various important settings such as discrete-time Markov decision processes. Finally, in fairly intricate experiments and through sheer learning, our framework reveals and quantifies causal links, which otherwise seem inexplicable.
****
Manish Prajapat,Mojmir Mutny,Melanie Zeilinger,Andreas Krause
Submodular Reinforcement Learning
In reinforcement learning (RL), rewards of states are typically considered addit

ive, and following the Markov assumption, they are independent of states visited
 previously. In many important applications, such as coverage control, experimen
t design and informative path planning, rewards naturally have diminishing retur
ns, i.e., their value decreases in light of similar states visited previously. T
o tackle this, we propose Submodular RL (subRL), a paradigm which seeks to optim
ize more general, non-additive (and history-dependent) rewards modelled via subm
odular set functions, which capture diminishing returns. Unfortunately, in gener
al, even in tabular settings, we show that the resulting optimization problem is
 hard to approximate. On the other hand, motivated by the success of greedy algo
rithms in classical submodular optimization, we propose subPO, a simple policy g
radient-based algorithm for subRL that handles non-additive rewards by greedily
maximizing marginal gains. Indeed, under some assumptions on the underlying Mark
ov Decision Process (MDP), subPO recovers optimal constant factor approximations
 of submodular bandits. Moreover, we derive a natural policy gradient approach f
or locally optimizing subRL instances even in large state- and action- spaces. W
e showcase the versatility of our approach by applying subPO to several applicat
ions, such as biodiversity monitoring, Bayesian experiment design, informative p
ath planning, and coverage maximization. Our results demonstrate sample efficien
cy, as well as scalability to high-dimensional state-action spaces.
****
Yang Yang,Wenhai Wang,Zhe Chen,Jifeng Dai,Liang Zheng
Bounding Box Stability against Feature Dropout Reflects Detector Generalization
across Environments
Bounding boxes uniquely characterize object detection, where a good detector giv
es accurate bounding boxes of categories of interest. However, in the real-world
 where test ground truths are not provided, it is non-trivial to find out whethe
r bounding boxes are accurate, thus preventing us from assessing the detector ge
neralization ability. In this work, we find under feature map dropout, good dete
ctors tend to output bounding boxes whose locations do not change much, while bo
unding boxes of poor detectors will undergo noticeable position changes. We comp
ute the box stability score (BS score) to reflect this stability. Specifically,
given an image, we compute a normal set of bounding boxes and a second set after
 feature map dropout. To obtain BS score, we use bipartite matching to find the
corresponding boxes between the two sets and compute the average Intersection ov
er Union (IoU) across the entire test set. We contribute to finding that BS scor
e has a strong, positive correlation with detection accuracy measured by mean av
erage precision (mAP) under various test environments. This relationship allows
us to predict the accuracy of detectors on various real-world test sets without
accessing test ground truths, verified on canonical detection tasks such as vehi
cle detection and pedestrian detection.
****
Yu-Lin Tsai,Chia-Yi Hsu,Chulin Xie,Chih-Hsun Lin,Jia You Chen,Bo Li,Pin-Yu Chen,
Chia-Mu Yu,Chun-Ying Huang
Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models?
Diffusion models for text-to-image (T2I) synthesis, such as Stable Diffusion (SD
), have recently demonstrated exceptional capabilities for generating high-quali
ty content. However, this progress has raised several concerns of potential misu
se, particularly in creating copyrighted, prohibited, and restricted content, or
 NSFW (not safe for work) images. While efforts have been made to mitigate such
problems, either by implementing a safety filter at the evaluation stage or by f
ine-tuning models to eliminate undesirable concepts or styles, the effectiveness
 of these safety measures in dealing with a wide range of prompts remains largel
y unexplored. In this work, we aim to investigate these safety mechanisms by pro
posing one novel concept retrieval algorithm for evaluation. We introduce Ring-A
-Bell, a model-agnostic red-teaming scheme for T2I diffusion models, where the w
hole evaluation can be prepared in advance without prior knowledge of the target
 model.
Specifically, Ring-A-Bell first performs concept extraction to obtain holistic r
epresentations for sensitive and inappropriate concepts. Subsequently, by levera
ging the extracted concept, Ring-A-Bell automatically identifies problematic pro

mpts for diffusion models with the corresponding generation of inappropriate con tent, allowing the user to assess the reliability of deployed safety mechanisms. Finally, we empirically validate our method by testing online services such as Midjourney and various methods of concept removal. Our results show that Ring-A-Bell, by manipulating safe prompting benchmarks, can transform prompts that were originally regarded as safe to evade existing safety mechanisms, thus revealing the defects of the so-called safety mechanisms which could practically lead to the generation of harmful contents. In essence, Ring-A-Bell could serve as a red-teaming tool to understand the limitations of deployed safety mechanisms and to explore the risk under plausible attacks. Our codes are available at https://github.com/chiayi-hsu/Ring-A-Bell.
****
Katie Kang,Amrith Setlur,Claire Tomlin,Sergey Levine
Deep Neural Networks Tend To Extrapolate Predictably
Conventional wisdom suggests that neural network predictions tend to be unpredictable and overconfident when faced with out-of-distribution (OOD) inputs. Our work reassesses this assumption for neural networks with high-dimensional inputs. Rather than extrapolating in arbitrary ways, we observe that neural network predictions often tend towards a constant value as input data becomes increasingly OOD. Moreover, we find that this value often closely approximates the optimal constant solution (OCS), i.e., the prediction that minimizes the average loss over the training data without observing the input. We present results showing this phenomenon across 8 datasets with different distributional shifts (including CIFAR10-C and ImageNet-R, S), different loss functions (cross entropy, MSE, and Gaussian NLL), and different architectures (CNNs and transformers). Furthermore, we present an explanation for this behavior, which we first validate empirically and then study theoretically in a simplified setting involving deep homogeneous networks with ReLU activations. Finally, we show how one can leverage our insights in practice to enable risk-sensitive decision-making in the presence of OOD inputs.
****
Ilker Kesen,Andrea Pedrotti,Mustafa Dogan,Michele Cafagna,Emre Can Acikgoz,Letitia Parcalabescu,Iacer Calixto,Anette Frank,Albert Gatt,Aykut Erdem,Erkut Erdem
ViLMA: A Zero-Shot Benchmark for Linguistic and Temporal Grounding in Video-Language Models
With the ever-increasing popularity of pretrained Video-Language Models (VidLMs), there is a pressing need to develop robust evaluation methodologies that delve deeper into their visio-linguistic capabilities. To address this challenge, we present ViLMA (Video Language Model Assessment), a task-agnostic benchmark that places the assessment of fine-grained capabilities of these models on a firm footing. Task-based evaluations, while valuable, fail to capture the complexities and specific temporal aspects of moving images that VidLMs need to process. Through carefully curated counterfactuals, ViLMA offers a controlled evaluation suite that sheds light on the true potential of these models, as well as their performance gaps compared to human-level understanding. ViLMA also includes proficiency tests, which assess basic capabilities deemed essential to solving the main counterfactual tests. We show that current VidLMs’ grounding abilities are no better than those of vision-language models which use static images. This is especially striking once the performance on proficiency tests is factored in. Our benchmark serves as a catalyst for future research on VidLMs, helping to highlight areas that still need to be explored.
****
Yuchen Zeng,Kangwook Lee
The Expressive Power of Low-Rank Adaptation
Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning method that leverages low-rank adaptation of weight matrices, has emerged as a prevalent technique for fine-tuning pre-trained models such as large language models and diffusion models.
Despite its huge success in practice, the theoretical underpinnings of LoRA have largely remained unexplored.

This paper takes the first step to bridge this gap by theoretically analyzing the expressive power of LoRA.

We prove that, for fully connected neural networks, LoRA can adapt any model f to accurately represent any smaller target model $\bar{f}$ if LoRA-rank $\geq(\text{width of }f) \times \frac{\text{depth of }\bar{f}}{\text{depth of }f}$, under a mild assumption.

We also quantify the approximation error when the LoRA-rank is lower than the threshold.

For Transformer networks, we show any model can be adapted to a target model of the same size with rank-$(\frac{\text{embedding size}}{2})$ LoRA adapters.

All our theoretical insights are validated by numerical experiments.

****

Yuyan Ni,Shikun Feng,Wei-Ying Ma,Zhi-Ming Ma,Yanyan Lan

Sliced Denoising: A Physics-Informed Molecular Pre-Training Method

While molecular pre-training has shown great potential in enhancing drug discovery, the lack of a solid physical interpretation in current methods raises concerns about whether the learned representation truly captures the underlying explanatory factors in observed data, ultimately resulting in limited generalization and robustness. Although denoising methods offer a physical interpretation, their accuracy is often compromised by ad-hoc noise design, leading to inaccurate learned force fields. To address this limitation, this paper proposes a new method for molecular pre-training, called sliced denoising (SliDe), which is based on the classical mechanical intramolecular potential theory. SliDe utilizes a novel noise strategy that perturbs bond lengths, angles, and torsion angles to achieve better sampling over conformations. Additionally, it introduces a random slicing approach that circumvents the computationally expensive calculation of the Jacobian matrix, which is otherwise essential for estimating the force field. By aligning with physical principles, SliDe shows a 42\% improvement in the accuracy of estimated force fields compared to current state-of-the-art denoising methods, and thus outperforms traditional baselines on various molecular property prediction tasks.

****

Sharut Gupta,Joshua Robinson,Derek Lim,Soledad Villar,Stefanie Jegelka

Structuring Representation Geometry with Rotationally Equivariant Contrastive Learning

Self-supervised learning converts raw perceptual data such as images to a compact space where simple Euclidean distances measure meaningful variations in data. In this paper, we extend this formulation by adding additional geometric structure to the embedding space by enforcing transformations of input space to correspond to simple (i.e., linear) transformations of embedding space. Specifically, in the contrastive learning setting, we introduce an equivariance objective and theoretically prove that its minima force augmentations on input space to correspond to rotations on the spherical embedding space. We show that merely combining our equivariant loss with a non-collapse term results in non-trivial representations, without requiring invariance to data augmentations. Optimal performance is achieved by also encouraging approximate invariance, where input augmentations correspond to small rotations. Our method, CARE: Contrastive Augmentation-induced Rotational Equivariance, leads to improved performance on downstream tasks and ensures sensitivity in embedding space to important variations in data (e.g., color) that standard contrastive methods do not achieve. Code is available at https://github.com/Sharut/CARE

****

Tian Jin,Nolan Clement,Xin Dong,Vaishnavh Nagarajan,Michael Carbin,Jonathan Ragan-Kelley,Gintare Karolina Dziugaite

The Cost of Scaling Down Large Language Models: Reducing Model Size Affects Memory before In-context Learning

We study how down-scaling large language model (LLM) size impacts LLM capabilities. We begin by measuring the effects of weight pruning – a popular technique for reducing model size – on the two abilities of LLMs: (a) recalling facts presented during pre-training and (b) processing information presented in context. Sur

prisingly, we find that existing pruning techniques affect these two abilities of LLMs differently. For example, pruning more than 30% of weights significantly decreases an LLM’s ability to recall facts presented during pre-training. Yet pruning 60-70% of weights largely preserves an LLM’s ability to process information in-context, ranging from retrieving answers based on information presented in context to learning parameterized functions such as a linear classifier based on a few examples. Moderate pruning impairs LLM’s ability to recall facts learnt from pre-training. However, its effect on model’s ability to process information presented in context is much less pronounced. The said disparate effects similarly arise when replacing the original model with a smaller dense one with reduced width and depth. This similarity suggests that model size reduction in general underpins the said disparity.
****
Zhiyu Mei,Wei Fu,Jiaxuan Gao,Guangju Wang,Huanchen Zhang,Yi Wu
SRL: Scaling Distributed Reinforcement Learning to Over Ten Thousand Cores
The ever-growing complexity of reinforcement learning (RL) tasks demands a distributed system to efficiently generate and process a massive amount of data. However, existing open-source libraries suffer from various limitations, which impede their practical use in challenging scenarios where large-scale training is necessary. In this paper, we present a novel abstraction on the dataflows of RL training, which unifies diverse RL training applications into a general framework. Following this abstraction, we develop a scalable, efficient, and extensible distributed RL system called ReaLly Scalable RL (SRL), which allows efficient and massively parallelized training and easy development of customized algorithms. Our evaluation shows that SRL outperforms existing academic libraries, reaching at most 21x higher training throughput in a distributed setting. On learning performance, beyond performing and scaling well on common RL benchmarks with different RL algorithms, SRL can reproduce the same solution in the challenging hide-and-seek environment as reported by OpenAI with up to 5x speedup in wallclock time. Notably, SRL is the first in the academic community to perform RL experiments at a large scale with over 15k CPU cores. SRL anonymous repository is available at: https://anonymous.4open.science/r/srl-1E45/.
****
Alizée Pace,Hugo Yèche,Bernhard Schölkopf,Gunnar Ratsch,Guy Tennenholtz
Delphic Offline Reinforcement Learning under Nonidentifiable Hidden Confounding
A prominent challenge of offline reinforcement learning (RL) is the issue of hidden confounding: unobserved variables may influence both the actions taken by the agent and the observed outcomes. Hidden confounding can compromise the validity of any causal conclusion drawn from data and presents a major obstacle to effective offline RL. In the present paper, we tackle the problem of hidden confounding in the nonidentifiable setting. We propose a definition of uncertainty due to hidden confounding bias, termed delphic uncertainty, which uses variation over world models compatible with the observations, and differentiate it from the well-known epistemic and aleatoric uncertainties. We derive a practical method for estimating the three types of uncertainties, and construct a pessimistic offline RL algorithm to account for them. Our method does not assume identifiability of the unobserved confounders, and attempts to reduce the amount of confounding bias. We demonstrate through extensive experiments and ablations the efficacy of our approach on a sepsis management benchmark, as well as on electronic health records. Our results suggest that nonidentifiable hidden confounding bias can be mitigated to improve offline RL solutions in practice.
****
Hao Xiong,Yehui Tang,Yunlin He,Wei Tan,Junchi Yan
Node2ket: Efficient High-Dimensional Network Embedding in Quantum Hilbert Space
Network embedding (NE) is a prominent technique for network analysis where the nodes are represented as vectorized embeddings in a continuous space. Existing works tend to resort to the low-dimensional embedding space for efficiency and less risk of over-fitting. In this paper, we explore a new NE paradigm whose embedding dimension goes exponentially high w.r.t. the number of parameters, yet being very efficient and effective. Specifically, the node embeddings are represented

 as product states that lie in a super high-dimensional (e.g. 2^{32}-dim) quan tum Hilbert space, with a carefully designed optimization approach to guarantee the robustness to work in different scenarios. In the experiments, we show diver se virtues of our methods, including but not limited to: the overwhelming perfor mance on downstream tasks against conventional low-dimensional NE baselines with the similar amount of computing resources, the super high efficiency for a fixe d low embedding dimension (e.g. 512) with less than 1/200 memory usage, the robu stness when equipped with different objectives and sampling strategies as a fund amental tool for future NE research. As a relatively unexplored topic in literat ure, the high-dimensional NE paradigm is demonstrated effective both experimenta lly and theoretically.
****
Victor Livernoche,Vineet Jain,Yashar Hezaveh,Siamak Ravanbakhsh
On Diffusion Modeling for Anomaly Detection
Known for their impressive performance in generative modeling, diffusion models are attractive candidates for density-based anomaly detection. This paper invest igates different variations of diffusion modeling for unsupervised and semi-supe rvised anomaly detection. In particular, we find that Denoising Diffusion Probab ility Models (DDPM) are performant on anomaly detection benchmarks yet computati onally expensive. By simplifying DDPM in application to anomaly detection, we ar e naturally led to an alternative approach called Diffusion Time Estimation (DTE). DTE estimates the distribution over diffusion time for a given input and uses the mode or mean of this distribution as the anomaly score. We derive an analyt ical form for this density and leverage a deep neural network to improve inferen ce efficiency. Through empirical evaluations on the ADBench benchmark, we demons trate that all diffusion-based anomaly detection methods perform competitively f or both semi-supervised and unsupervised settings. Notably, DTE achieves orders of magnitude faster inference time than DDPM, while outperforming it on this ben chmark. These results establish diffusion-based anomaly detection as a scalable alternative to traditional methods and recent deep-learning techniques for stand ard unsupervised and semi-supervised anomaly detection settings.
****
Jiayang Liu,Yiming Bu,Daniel Tso,Qinru Qiu
Improved Efficiency Based on Learned Saccade and Continuous Scene Reconstruction From Foveated Visual Sampling
High accuracy, low latency and high energy efficiency represent a set of contrad ictory goals when searching for system solutions for image classification and d etection. While high-quality images naturally result in more precise detection a nd classification, they also result in a heavier computational workload for imag ing and processing, reduce camera refresh rates, and increase the volume of data communication between the camera and processor. Taking inspiration from the fov eal-peripheral sampling mechanism, saccade mechanism observed in the human visua l system and the filling-in phenomena of brain, we have developed an active scen e reconstruction architecture based on multiple foveal views. This model stitche s together information from foveal and peripheral vision, which are sampled from multiple glances. Assisted by a reinforcement learning-based saccade mechanism, our model reduces the required input pixels by over 90\% per frame while mainta ining the same level of performance in image recognition as with the original im ages. We evaluated the effectiveness of our model using the GTSRB dataset and th e ImageNet dataset. Using an equal number of input pixels, our study demonstrate s a 5\% higher image recognition accuracy compared to state-of-the-art foveal-pe ripheral vision systems. Furthermore, we demonstrate that our foveal sampling/sa ccadic scene reconstruction model exhibits significantly lower complexity and hi gher data efficiency during the training phase compared to existing approaches.
****
Jiaxin Yin,Yuanyuan Qiao,Zitang Zhou,Xiangchao Wang,Jie Yang
MCM: Masked Cell Modeling for Anomaly Detection in Tabular Data
This paper addresses the problem of anomaly detection in tabular data, which is usually implemented in an one-class classification setting where the training se t only contains normal samples. Inspired by the success of masked image/language

modeling in vision and natural language domains, we extend masked modeling methods to address this problem by capturing intrinsic correlations between features in training set. Thus, a sample deviate from such correlations is related to a high possibility of anomaly. To obtain multiple and diverse correlations, we propose a novel masking strategy which generates multiple masks by learning, and design a diversity loss to reduce the similarity of different masks. Extensive experiments show our method achieves state-of-the-art performance. We also discuss the interpretability from the perspective of each individual feature and correlations between features.
****
Yifei Wang,Qi Zhang,Yaoyu Guo,Yisen Wang
Non-negative Contrastive Learning
Deep representations have shown promising performance when transferred to downstream tasks in a black-box manner. Yet, their inherent lack of interpretability remains a significant challenge, as these features are often opaque to human understanding. In this paper, we propose Non-negative Contrastive Learning (NCL), a renaissance of Non-negative Matrix Factorization (NMF) aimed at deriving interpretable features. The power of NCL lies in its enforcement of non-negativity constraints on features, reminiscent of NMF's capability to extract features that align closely with sample clusters. NCL not only aligns mathematically well with an NMF objective but also preserves NMF's interpretability attributes, resulting in a more sparse and disentangled representation compared to standard contrastive learning (CL). Theoretically, we establish guarantees on the identifiability and downstream generalization of NCL. Empirically, we show that these advantages enable NCL to outperform CL significantly on feature disentanglement, feature selection, as well as downstream classification tasks. At last, we show that NCL can be easily extended to other learning scenarios and benefit supervised learning as well. Code is available at https://github.com/PKU-ML/non_neg.
****
Zhiliang Peng,Wenhui Wang,Li Dong,Yaru Hao,Shaohan Huang,Shuming Ma,Qixiang Ye,Furu Wei
Grounding Multimodal Large Language Models to the World
We introduce Kosmos-2, a Multimodal Large Language Model (MLLM), enabling new capabilities of perceiving object descriptions (e.g., bounding boxes) and grounding text to the visual world. Specifically, we represent text spans (i.e., referring expressions and noun phrases) as links in Markdown, i.e., [text span](bounding boxes), where object descriptions are sequences of location tokens. To train the model, we construct a large-scale dataset about grounded image-text pairs (GrIT) together with multimodal corpora. In addition to the existing capabilities of MLLMs (e.g., perceiving general modalities, following instructions, and performing in-context learning), Kosmos-2 integrates the grounding capability to downstream applications, while maintaining the conventional capabilities of MLLMs (e.g., perceiving general modalities, following instructions, and performing in-context learning). Kosmos-2 is evaluated on a wide range of tasks, including (i) multimodal grounding, such as referring expression comprehension and phrase grounding, (ii) multimodal referring, such as referring expression generation, (iii) perception-language tasks, and (iv) language understanding and generation. This study sheds a light on the big convergence of language, multimodal perception, and world modeling, which is a key step toward artificial general intelligence. Code can be found in https://aka.ms/kosmos-2.
****
Qiongyi Zhou,Changde Du,Shengpei Wang,Huiguang He
CLIP-MUSED: CLIP-Guided Multi-Subject Visual Neural Information Semantic Decoding
The study of decoding visual neural information faces challenges in generalizing single-subject decoding models to multiple subjects, due to individual differences. Moreover, the limited availability of data from a single subject has a constraining impact on model performance. Although prior multi-subject decoding methods have made significant progress, they still suffer from several limitations, including difficulty in extracting global neural response features, linear scali

ng of model parameters with the number of subjects, and inadequate characterizat
ion of the relationship between neural responses of different subjects to variou
s stimuli.
To overcome these limitations, we propose a CLIP-guided Multi-sUbject visual neu
ral information SEmantic Decoding (CLIP-MUSED) method. Our method consists of a
Transformer-based feature extractor to effectively model global neural represent
ations. It also incorporates learnable subject-specific tokens that facilitates
the aggregation of multi-subject data without a linear increase of parameters. A
dditionally, we employ representational similarity analysis (RSA) to guide token
 representation learning based on the topological relationship of visual stimuli
 in the representation space of CLIP, enabling full characterization of the rela
tionship between neural responses of different subjects under different stimuli.
 Finally, token representations are used for multi-subject semantic decoding. Ou
r proposed method outperforms single-subject decoding methods and achieves state
-of-the-art performance among the existing multi-subject methods on two fMRI dat
asets. Visualization results provide insights into the effectiveness of our prop
osed method. Code is available at https://github.com/CLIP-MUSED/CLIP-MUSED.
****
Michal Geyer,Omer Bar-Tal,Shai Bagon,Tali Dekel
TokenFlow: Consistent Diffusion Features for Consistent Video Editing
The generative AI revolution has recently expanded to videos. Nevertheless, curr
ent state-of-the-art video models are still lagging behind image models in terms
 of visual quality and user control over the generated content. In this work, we
 present a framework that harnesses the power of a text-to-image diffusion model
 for the task of text-driven video editing. Specifically, given a source video a
nd a target text-prompt, our method generates a high-quality video that adheres
to the target text, while preserving the spatial layout and motion of the input
video. Our method is based on a key observation that consistency in the edited v
ideo can be obtained by enforcing consistency in the diffusion feature space. We
 achieve this by explicitly propagating diffusion features based on inter-frame
correspondences, readily available in the model. Thus, our framework does not re
quire any training or fine-tuning, and can work in conjunction with any off-the-
shelf text-to-image editing method. We demonstrate state-of-the-art editing resu
lts on a variety of real-world videos.
****
Jungin Park,Jiyoung Lee,Kwanghoon Sohn
Bridging Vision and Language Spaces with Assignment Prediction
While pretrained large language models (LLMs) excel in understanding linguistic
contexts, it is still an open question: Can LLMs extend their capabilities beyon
d linguistic contexts to non-linguistic information? This paper introduces VLAP,
 a novel approach that bridges vision encoders and language models through assig
nment prediction. Since the LLMs interpret and reason linguistic information fro
m correlations between word embeddings, we harness the well-established word emb
eddings to map visual representations into language space. Specifically, we simu
ltaneously assign the visual and text representations to a set of word embedding
s within LLMs. We propose a new training objective, optimal transport-based assi
gnment prediction, to enforce the consistency of word assignments for paired mul
timodal data. This allows frozen LLMs to ground their word embedding space in vi
sual data and use their robust semantic taxonomy visually. Moreover, VLAP is mem
ory- and parameter-efficient in that it trains only a single linear layer, and w
orks without extra embedding space (e.g. learnable prototypes) for the assignmen
t prediction. Experimental results show that VLAP achieves substantial improveme
nts over the previous linear transformation-based methods across a range of visi
on-language tasks, including image captioning, visual question answering, and cr
oss-modal retrieval. We also demonstrate the learned visual representations hold
 a semantic taxonomy of LLMs, making visual semantic arithmetic possible.
****
Peng Chen,Yingying ZHANG,Yunyao Cheng,Yang Shu,Yihang Wang,Qingsong Wen,Bin Yang
,Chenjuan Guo
Pathformer: Multi-scale Transformers with Adaptive Pathways for Time Series Fore

casting
Transformers for time series forecasting mainly model time series from limited o r fixed scales, making it challenging to capture different characteristics spann ing various scales. We propose Pathformer, a multi-scale Transformer with adapti ve pathways. It integrates both temporal resolution and temporal distance for mu lti-scale modeling. Multi-scale division divides the time series into different temporal resolutions using patches of various sizes. Based on the division of ea ch scale, dual attention is performed over these patches to capture global corre lations and local details as temporal dependencies. We further enrich the multi-scale Transformer with adaptive pathways, which adaptively adjust the multi-scal e modeling process based on the varying temporal dynamics of the input, improvin g the accuracy and generalization of Pathformer. Extensive experiments on eleven real-world datasets demonstrate that Pathformer not only achieves state-of-the-art performance by surpassing all current models but also exhibits stronger gene ralization abilities under various transfer scenarios. The code is made availabl e at https://github.com/decisionintelligence/pathformer.
****
Christopher Fifty,Dennis Duan,Ronald Guenther Junkins,Ehsan Amid,Jure Leskovec,C hristopher Re,Sebastian Thrun
Context-Aware Meta-Learning
Large Language Models like ChatGPT demonstrate a remarkable capacity to learn ne w concepts during inference without any fine-tuning. However, visual models trai ned to detect new objects during inference have been unable to replicate this ab ility, and instead either perform poorly or require meta-training and/or fine-tu ning on similar objects. In this work, we propose a meta-learning algorithm that emulates Large Language Models by learning new visual concepts during inference without fine-tuning. Our approach leverages a frozen pre-trained feature extrac tor, and analogous to in-context learning, recasts meta-learning as sequence mod eling over datapoints with known labels and a test datapoint with an unknown lab el. On 8 out of 11 meta-learning benchmarks, our approach---without meta-trainin g or fine-tuning---exceeds or matches the state-of-the-art algorithm, P>M>F, whi ch is meta-trained on these benchmarks.
****
Yifan Feng,Yihe Luo,Shihui Ying,Yue Gao
LightHGNN: Distilling Hypergraph Neural Networks into MLPs for 100x Faster Infer ence
Hypergraph Neural Networks (HGNNs) have recently attracted much attention and ex hibited satisfactory performance due to their superiority in high-order correlat ion modeling.
However, it is noticed that the high-order modeling capability of hypergraph als o brings increased computation complexity, which hinders its practical industria l deployment.
In practice, we find that one key barrier to the efficient deployment of HGNNs i s the high-order structural dependencies during inference.
In this paper, we propose to bridge the gap between the HGNNs and inference-effi cient Multi-Layer Perceptron (MLPs) to eliminate the hypergraph dependency of HG NNs and thus reduce computational complexity as well as improve inference speed.

Specifically, we introduce LightHGNN and LightHGNN$^+$ for fast inference with l ow complexity. LightHGNN directly distills the knowledge from teacher HGNNs to s tudent MLPs via soft labels, and LightHGNN$^+$ further explicitly injects reliab le high-order correlations into the student MLPs to achieve topology-aware disti llation and resistance to over-smoothing.
Experiments on eight hypergraph datasets demonstrate that even without hypergrap h dependency, the proposed LightHGNNs can still achieve competitive or even bett er performance than HGNNs and outperform vanilla MLPs by 16.3 on average. Exte nsive experiments on three graph datasets further show the average best performa nce of our LightHGNNs compared with all other methods.
Experiments on synthetic hypergraphs with 5.5w vertices indicate LightHGNNs can run $100\times$ faster than HGNNs, showcasing their ability for latency-sensitiv

e deployments.
****
Kaichao You,Guo Qin,Anchang Bao,Meng Cao,Ping Huang,Jiulong Shan,Mingsheng Long
Efficient ConvBN Blocks for Transfer Learning and Beyond
Convolution-BatchNorm (ConvBN) blocks are integral components in various computer vision tasks and other domains. A ConvBN block can operate in three modes: Train, Eval, and Deploy. While the Train mode is indispensable for training models from scratch, the Eval mode is suitable for transfer learning and beyond, and the Deploy mode is designed for the deployment of models. This paper focuses on the trade-off between stability and efficiency in ConvBN blocks: Deploy mode is efficient but suffers from training instability; Eval mode is widely used in transfer learning but lacks efficiency. To solve the dilemma, we theoretically reveal the reason behind the diminished training stability observed in the Deploy mode. Subsequently, we propose a novel Tune mode to bridge the gap between Eval mode and Deploy mode. The proposed Tune mode is as stable as Eval mode for transfer learning, and its computational efficiency closely matches that of the Deploy mode. Through extensive experiments in object detection, classification, and adversarial example generation across 5 datasets and 12 model architectures, we demonstrate that the proposed Tune mode retains the performance while significantly reducing GPU memory footprint and training time, thereby contributing efficient ConvBN blocks for transfer learning and beyond. Our method has been integrated into both PyTorch (general machine learning framework) and MMCV/MMEngine (computer vision framework). Practitioners just need one line of code to enjoy our efficient ConvBN blocks thanks to PyTorch's builtin machine learning compilers.
****
Gourav Datta,Zeyu Liu,Peter Anthony Beerel
Can we get the best of both Binary Neural Networks and Spiking Neural Networks for Efficient Computer Vision?
Binary Neural networks (BNN) have emerged as an attractive computing paradigm for a wide range of low-power vision tasks. However, state-of-the-art (SOTA) BNNs do not yield any sparsity, and induce a significant number of non-binary operations. On the other hand, activation sparsity can be provided by spiking neural networks (SNN), that too have gained significant traction in recent times. Thanks to this sparsity, SNNs when implemented on neuromorphic hardware, have the potential to be significantly more power-efficient compared to traditional artifical neural networks (ANN). However, SNNs incur multiple time steps to achieve close to SOTA accuracy. Ironically, this increases latency and energy---costs that SNNs were proposed to reduce---and presents itself as a major hurdle in realizing SNNs’ theoretical gains in practice. This raises an intriguing question: *Can we obtain SNN-like sparsity and BNN-like accuracy and enjoy the energy-efficiency benefits of both?* To answer this question, in this paper, we present a training framework for sparse binary activation neural networks (BANN) using a novel variant of the Hoyer regularizer. We estimate the threshold of each BANN layer as the Hoyer extremum of a clipped version of its activation map, where the clipping value is trained using gradient descent with our Hoyer regularizer.
This approach shifts the activation values away from the threshold, thereby mitigating the effect of noise that can otherwise degrade the BANN accuracy. Our approach outperforms existing BNNs, SNNs, and adder neural networks (that also avoid energy-expensive multiplication operations similar to BNNs and SNNs) in terms of the accuracy-FLOPs trade-off for complex image recognition tasks. Downstream experiments on object detection further demonstrate the efficacy of our approach. Lastly, we demonstrate the portability of our approach to SNNs with multiple time steps. Codes are publicly available [here](https://github.com/godatta/Ultra-Low-Latency-SNN).
****
Xinghang Li,Minghuan Liu,Hanbo Zhang,Cunjun Yu,Jie Xu,Hongtao Wu,Chilam Cheang,Ya Jing,Weinan Zhang,Huaping Liu,Hang Li,Tao Kong
Vision-Language Foundation Models as Effective Robot Imitators
Recent progress in vision language foundation models has shown their ability to understand multimodal data and resolve complicated vision language tasks, includ

ing robotics manipulation. We seek a straightforward way of making use of existing vision-language models (VLMs) with simple fine-tuning on robotics data.
To this end, we derive a simple and novel vision-language manipulation framework, dubbed RoboFlamingo, built upon the open-source VLMs, OpenFlamingo. Unlike prior works, RoboFlamingo utilizes pre-trained VLMs for single-step vision-language comprehension, models sequential history information with an explicit policy head, and is slightly fine-tuned by imitation learning only on language-conditioned manipulation datasets. Such a decomposition provides RoboFlamingo the flexibility for open-loop control and deployment on low-performance platforms. By exceeding the state-of-the-art performance with a large margin on the tested benchmark, we show RoboFlamingo can be an effective and competitive alternative to adapt VLMs to robot control.
Our extensive experimental results also reveal several interesting conclusions regarding the behavior of different pre-trained VLMs on manipulation tasks. We believe RoboFlamingo has the potential to be a cost-effective and easy-to-use solution for robotics manipulation, empowering everyone with the ability to fine-tune their own robotics policy. Our code will be made public upon acceptance.
****
Daniil Tiapkin,Denis Belomestny,Daniele Calandriello,Eric Moulines,Alexey Naumov,Pierre Perrault,Michal Valko,Pierre Menard
Demonstration-Regularized RL
Incorporating expert demonstrations has empirically helped to improve the sample efficiency of reinforcement learning (RL). This paper quantifies theoretically to what extent this extra information reduces RL's sample complexity. In particular, we study the demonstration-regularized reinforcement learning framework that leverages the expert demonstrations by KL-regularization for a policy learned by behavior cloning. Our findings reveal that using N^{E} expert demonstrations enables the identification of an optimal policy at a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{Poly}(S,A,H)/(\varepsilon^2 N^{\mathrm{E}}))$ in finite and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d,H)/(\varepsilon^2 N^{\mathrm{E}}))$ in linear Markov decision processes, where $\varepsilon$is the target precision, H the horizon, A the number of action, S the number of states in the finite case and d the dimension of the feature space in the linear case. As a by-product, we provide tight convergence guarantees for the behavior cloning procedure under general assumptions on the policy classes. Additionally, we establish that demonstration-regularized methods are provably efficient for reinforcement learning from human feedback (RLHF). In this respect, we provide theoretical evidence showing the benefits of KL-regularization for RLHF in tabular and linear MDPs.
Interestingly, we avoid pessimism injection by employing computationally feasible regularization to handle reward estimation uncertainty, thus setting our approach apart from the prior works.
****
Shengding Hu,Xin Liu,Xu Han,Xinrong Zhang,Chaoqun He,Weilin Zhao,Yankai Lin,Ning Ding,Zebin Ou,Guoyang Zeng,Zhiyuan Liu,Maosong Sun
Predicting Emergent Abilities with Infinite Resolution Evaluation
The scientific scale-up of large language models (LLMs) necessitates a comprehensive understanding of their scaling properties. However, the existing literature on the scaling properties only yields an incomplete answer: optimization loss decreases predictably as the model size increases, in line with established scaling law; yet no scaling law for task has been established and the task performances are far from predictable during scaling. Task performances typically show minor gains on small models until they improve dramatically once models exceed a size threshold, exemplifying the ''emergent abilities''. In this study, we discover that small models, although they exhibit minor performance, demonstrate critical and consistent task performance improvements that are not captured by conventional evaluation strategies due to insufficient measurement resolution. To measure such improvements, we introduce PassUntil, an evaluation strategy with theoretically infinite resolution, through massive sampling in the decoding phase. With PassUntil, we conduct a quantitative investigation into the scaling law of tas

k performance. The investigation contains two parts. Firstly, a strict task scaling law that is not conventionally known to exist, is identified, enhancing the predictability of task performances. Remarkably, we are able to predict the performance of the 2.4B model on code generation with merely 0.05\% deviation before training starts, which is the first systematic attempt to verify predictable scaling proposed by GPT-4's report. Secondly, underpinned by PassUntil, we are able to study emergent abilities quantitatively. We identify a kind of accelerated emergence whose scaling curve cannot be fitted by standard scaling law function and has a increasing speed. We then examine two hypothesis and imply that the ``multiple circuits hypothesis'' might be responsible for the accelerated emergence.
****
Haotian Yan,Ming Wu,Chuang Zhang
Multi-Scale Representations by Varying Window Attention for Semantic Segmentation
Multi-scale representations are central to semantic segmentation. We visualize the
effective receptive field (ERF) of canonical multi-scale representations and point
out two risks in learning them: scale inadequacy and field inactivation. To address these issues, a novel multi-scale learner, varying window attention (VWA), is presented. VWA leverages the local window attention (LWA) and disentangles LWA into the query window and context window, allowing the context’s scale to vary for the query to learn representations at multiple scales. However, varying the context to large-scale windows (enlarging ratio R) can significantly increase
the memory footprint and computation cost (R2 times larger than LWA). We propose a simple but professional re-scaling strategy to zero the extra induced cost without compromising performance. In consequence, VWA uses the same cost as LWA but overcomes the limitation of the local window. Furthermore, building upon VWA and employing various MLPs, we introduce a multi-scale decoder (MSD), VWFormer, to improve learning multi-scale representations in semantic segmentation. VWFormer achieves efficiency competitive with the most computefriendly MSDs, like FPN and MLP decoder, but performs much better than any MSDs. For instance, at little extra overhead, ~ 10G FLOPs, VWFormer improves Mask2Former by 1.0% – 1.4% mIoU on ADE20K. Using nearly half of the computation, VWFormer outperforms the popular UperNet by 1.0% – 2.1% mIoU
****
Jean-Rémy Conti,Stephan Clémençon
Assessing Uncertainty in Similarity Scoring: Performance & Fairness in Face Recognition
The ROC curve is the major tool for assessing not only the performance but also the fairness properties of a similarity scoring function. In order to draw reliable conclusions based on empirical ROC analysis, accurately evaluating the uncertainty level related to statistical versions of the ROC curves of interest is absolutely necessary, especially for applications with considerable societal impact such as Face Recognition. In this article, we prove asymptotic guarantees for empirical ROC curves of similarity functions as well as for by-product metrics useful to assess fairness. We also explain that, because the false acceptance/rejection rates are of the form of U-statistics in the case of similarity scoring, the naive bootstrap approach may jeopardize the assessment procedure. A dedicated recentering technique must be used instead. Beyond the theoretical analysis carried out, various experiments using real face image datasets provide strong empirical evidence of the practical relevance of the methods promoted here, when applied to several ROC-based measures such as popular fairness metrics.
****
Manh Luong,Khai Nguyen,Nhat Ho,Reza Haf,Dinh Phung,Lizhen Qu
Revisiting Deep Audio-Text Retrieval Through the Lens of Transportation
The Learning-to-match (LTM) framework proves to be an effective inverse optimal transport approach for learning the underlying ground metric between two sources

of data, facilitating subsequent matching. However, the conventional LTM framework faces scalability challenges, necessitating the use of the entire dataset each time the parameters of the ground metric are updated. In adapting LTM to the deep learning context, we introduce the mini-batch Learning-to-match (m-LTM) framework for audio-text retrieval problems. This framework leverages mini-batch subsampling and Mahalanobis-enhanced family of ground metrics. Moreover, to cope with misaligned training data in practice, we propose a variant using partial optimal transport to mitigate the harm of misaligned data pairs in training data. We conduct extensive experiments on audio-text matching problems using three datasets: AudioCaps, Clotho, and ESC-50. Results demonstrate that our proposed method is capable of learning rich and expressive joint embedding space, which achieves SOTA performance. Beyond this, the proposed m-LTM framework is able to close the modality gap across audio and text embedding, which surpasses both triplet and contrastive loss in the zero-shot sound event detection task on the ESC-50 dataset. Notably, our strategy of employing partial optimal transport with m-LTM demonstrates greater noise tolerance than contrastive loss, especially under varying noise ratios in training data on the AudioCaps dataset. Our code is available at https://github.com/v-manhlt3/m-LTM-Audio-Text-Retrieval
****
Thien Le,Luana Ruiz,Stefanie Jegelka
A Poincaré Inequality and Consistency Results for Signal Sampling on Large Graphs
Large-scale graph machine learning is challenging as the complexity of learning models scales with the graph size. Subsampling the graph is a viable alternative, but sampling on graphs is nontrivial as graphs are non-Euclidean. Existing graph sampling techniques require not only computing the spectra of large matrices but also repeating these computations when the graph changes, e.g., grows. In this paper, we introduce a signal sampling theory for a type of graph limit---the graphon. We prove a Poincaré inequality for graphon signals and show that complements of node subsets satisfying this inequality are unique sampling sets for Paley-Wiener spaces of graphon signals. Exploiting connections with spectral clustering and Gaussian elimination, we prove that such sampling sets are consistent in the sense that unique sampling sets on a convergent graph sequence converge to unique sampling sets on the graphon. We then propose a related graphon signal sampling algorithm for large graphs, and demonstrate its good empirical performance on graph machine learning tasks.
****
Yueru Luo,Shuguang Cui,Zhen Li
DV-3DLane: End-to-end Multi-modal 3D Lane Detection with Dual-view Representation
Accurate 3D lane estimation is crucial for ensuring safety in autonomous driving. However, prevailing monocular techniques suffer from depth loss and lighting variations, hampering accurate 3D lane detection. In contrast, LiDAR points offer geometric cues and enable precise localization. In this paper, we present DV-3DLane, a novel end-to-end **D**ual-**V**iew multi-modal **3D Lane** detection framework that synergizes the strengths of both images and LiDAR points. We propose to learn multi-modal features in dual-view spaces, *i.e.*, *perspective view* (PV) and *bird's-eye-view* (BEV), effectively leveraging the modal-specific information. To achieve this, we introduce three designs: 1) A bidirectional feature fusion strategy that integrates multi-modal features into each view space, exploiting their unique strengths. 2) A unified query generation approach that leverages lane-aware knowledge from both PV and BEV spaces to generate queries. 3) A 3D dual-view deformable attention mechanism, which aggregates discriminative features from both PV and BEV spaces into queries for accurate 3D lane detection. Extensive experiments on the public benchmark, OpenLane, demonstrate the efficacy and efficiency of DV-3DLane. It achieves state-of-the-art performance, with a remarkable 11.2 gain in F1 score and a substantial 53.5% reduction in errors. Code is available on [github](https://github.com/JMoonr/dv-3dlane).
****
Tuan Le,Julian Cremer,Frank Noe,Djork-Arné Clevert,Kristof T Schütt

Navigating the Design Space of Equivariant Diffusion-Based Generative Models for De Novo 3D Molecule Generation
Deep generative diffusion models are a promising avenue for 3D de novo molecular design in materials science and drug discovery.
However, their utility is still limited by suboptimal performance on large molecular structures and limited training data.
To address this gap, we explore the design space of E(3)-equivariant diffusion models, focusing on previously unexplored areas.
Our extensive comparative analysis evaluates the interplay between continuous and discrete state spaces.
From this investigation, we present the EQGAT-diff model, which consistently outperforms established models for the QM9 and GEOM-Drugs datasets.
Significantly, EQGAT-diff takes continuous atom positions, while chemical elements and bond types are categorical and uses time-dependent loss weighting, substantially increasing training convergence, the quality of generated samples, and inference time. We also showcase that including chemically motivated additional features like hybridization states in the diffusion process enhances the validity of generated molecules.
To further strengthen the applicability of diffusion models to limited training data, we investigate the transferability of EQGAT-diff trained on the large PubChem3D dataset with implicit hydrogen atoms to target different data distributions. Fine-tuning EQGAT-diff for just a few iterations shows an efficient distribution shift, further improving performance throughout data sets.
Finally, we test our model on the Crossdocked data set for structure-based de novo ligand generation, underlining the importance of our findings showing state-of-the-art performance on Vina docking scores.
****
Shunyu Yao,Howard Chen,Austin W. Hanjie,Runzhe Yang,Karthik R Narasimhan
COLLIE: Systematic Construction of Constrained Text Generation Tasks
Text generation under constraints have seen increasing interests in natural language processing, especially with the rapidly improving capabilities of large language models. However, existing benchmarks for constrained generation usually focus on fixed constraint types (e.g. generate a sentence containing certain words) that have proved to be easy for state-of-the-art models like GPT-4. We present COLLIE, a grammar-based framework that allows the specification of rich, compositional constraints with diverse generation levels (word, sentence, paragraph, passage) and modeling challenges (e.g. language understanding, logical reasoning, counting, semantic planning). We also develop tools for automatic extraction of task instances given a constraint structure and a raw text corpus. Using COLLIE, we compile the COLLIE-v1 dataset with 1,132 instances comprising 13 constraint structures. We perform systematic experiments across five state-of-the-art instruction-tuned language models and analyze their performances to reveal shortcomings. COLLIE is designed to be extensible and lightweight, and we hope the community finds it useful to develop more complex constraints and evaluations in the future.
****
Tom Sherborne,Naomi Saphra,Pradeep Dasigi,Hao Peng
TRAM: Bridging Trust Regions and Sharpness Aware Minimization
Sharpness-aware minimization (SAM) reports improving domain generalization by reducing the loss surface curvature in the parameter space. However, generalization during _fine-tuning_ is often more dependent on the transferability of _representations_ in the function space. Trust-region methods (TR) target this goal by regularizing representation curvature to reduce catastrophic forgetting of pre-trained task-agnostic information while adopting task-specific skills. We consider unifying these strategies for low curvature in both parameter space and function space to improve out-of-domain (OOD) generalization. We propose **Trust Region Aware Minimization** (TRAM), a SAM algorithm fine-tuning for low parameter sharpness and smooth, informative representations preserving pre-trained structure. TRAM uses a trust region bound to inform the SAM adversarial neighborhood, introducing an awareness of function

curvature within optimization for flatter minima. We empirically validate TRAM in vision (cross-dataset adaptation) and text (OOD language modeling, zero-shot cross-lingual transfer) tasks where robust domain transfer and representation generality are critical. TRAM outperforms SAM- and TR-based optimization across all tasks, notably surpassing competing methods for hard transfer between _anticorrelated_ domains. TRAM establishes a novel standard in fine-tuning for domain-generalizable models with minimal additional computation over previous sharpness-aware methods.
****
Jiyang Zheng,Yu Yao,Bo Han,Dadong Wang,Tongliang Liu
Enhancing Contrastive Learning for Ordinal Regression via Ordinal Content Preserved Data Augmentation
Contrastive learning, while highly effective for a lot of tasks, shows limited improvement in ordinal regression. We find that the limitation comes from the predefined strong data augmentations employed in contrastive learning. Intuitively, for ordinal regression datasets, the discriminative information (ordinal content information) contained in instances is subtle. The strong augmentations can easily overshadow or diminish this ordinal content information. As a result, when contrastive learning is used to extract common features between weakly and strongly augmented images, the derived features often lack this essential ordinal content, rendering them less useful in training models for ordinal regression. To improve contrastive learning's utility for ordinal regression, we propose a novel augmentation method to replace the predefined strong argumentation based on the principle of minimal change. Our method is designed in a generative manner that can effectively generate images with different styles but contains desired ordinal content information. Extensive experiments validate the effectiveness of our proposed method, which serves as a plug-and-play solution and consistently improves the performance of existing state-of-the-art methods in ordinal regression tasks.
****
Rhys Gould,Euan Ong,George Ogden,Arthur Conmy
Successor Heads: Recurring, Interpretable Attention Heads In The Wild
In this work we describe successor heads: attention heads that increment tokens with a natural ordering, such as numbers, months, and days.
For example, successor heads increment 'Monday' into 'Tuesday'.
We explain the successor head behavior with an approach rooted in mechanistic interpretability, the field that aims to explain how models complete tasks in human-understandable terms.
Existing research in this area has struggled to find recurring, mechanistically interpretable large language model (LLM) components beyond small toy models. Further, existing results have led to very little insight to explain the internals of the larger models that are used in practice.
In this paper, we analyze the behavior of successor heads in LLMs and find that they implement abstract representations that are common to different architectures.
Successor heads form in LLMs with as few as 31 million parameters, and at least as many as 12 billion parameters, such as GPT-2, Pythia, and Llama-2.
We find a set of 'mod 10' features that underlie how successor heads increment in LLMs across different architectures and sizes.
We perform vector arithmetic with these features to edit head behavior and provide insights into numeric representations within LLMs. Additionally, we study the behavior of successor heads on natural language data, where we find that successor heads are important for achieving a low loss on examples involving succession, and also identify interpretable polysemanticity in a Pythia successor head.
****
Dean A Pospisil,Brett W. Larsen,Sarah E Harvey,Alex H Williams
Estimating Shape Distances on Neural Representations with Limited Samples
Measuring geometric similarity between high-dimensional network representations is a topic of longstanding interest to neuroscience and deep learning. Although many methods have been proposed, only a few works have rigorously analyzed their

statistical efficiency or quantified estimator uncertainty in data-limited regimes. Here, we derive upper and lower bounds on the worst-case convergence of standard estimators of shape distance—a measure of representational dissimilarity proposed by Williams et al. (2021). These bounds reveal the challenging nature of the problem in high-dimensional feature spaces. To overcome these challenges, we introduce a novel method-of-moments estimator with a tunable bias-variance tradeoff parameterized by an upper bound on bias. We show that this estimator achieves superior performance to standard estimators in simulation and on neural data, particularly in high-dimensional settings. Our theoretical work and estimator thus respectively define and dramatically expand the scope of neural data for which geometric similarity can be accurately measured.
****
Jingyang Zhang,Shiwei Li,Yuanxun Lu,Tian Fang,David Neil McKinnon,Yanghai Tsin,Long Quan,Yao Yao
JointNet: Extending Text-to-Image Diffusion for Dense Distribution Modeling
We introduce JointNet, a novel neural network architecture for modeling the joint distribution of images and an additional dense modality (e.g., depth maps).
JointNet is extended from a pre-trained text-to-image diffusion model, where a copy of the original network is created for the new dense modality branch and is densely connected with the RGB branch.
The RGB branch is locked during network fine-tuning, which enables efficient learning of the new modality distribution while maintaining the strong generalization ability of the large-scale pre-trained diffusion model.
We demonstrate the effectiveness of JointNet by using the RGB-D diffusion as an example and through extensive experiments, showcasing its applicability in a variety of applications, including joint RGB-D generation, dense depth prediction, depth-conditioned image generation, and high-resolution 3D panorama generation.
****
Xihaier Luo,Wei Xu,Balu Nadiga,Yihui Ren,Shinjae Yoo
Continuous Field Reconstruction from Sparse Observations with Implicit Neural Networks
Reliably reconstructing physical fields from sparse sensor data is a challenge that frequenty arises in many scientific domains. In practice, the process generating the data is often not known to sufficient accuracy. Therefore, there is a growing interest in the deep neural network route to the problem. In this work, we present a novel approach that learns a continuous representation of the field using implicit neural representations (INR). Specifically, after factorizing spatiotemporal variability into spatial and temporal components using the technique of separation of variables, the method learns relevant basis functions from sparsely sampled irregular data points to thus develop a continuous representation of the data. In experimental evaluations, the proposed model outperforms recent INR methods, offering superior reconstruction quality on simulation data from a state of the art climate model and on a second dataset that comprises of ultra-high resolution satellite-based sea surface temperature field. [Website for the Project: Both data and code are accessible.](https://xihaier.github.io/ICLR-2024-MMGN/)
****
Marlene Careil,Matthew J. Muckley,Jakob Verbeek,Stéphane Lathuilière
Towards image compression with perfect realism at ultra-low bitrates
Image codecs are typically optimized to trade-off bitrate vs. distortion metrics. At low bitrates, this leads to compression artefacts which are easily perceptible, even when training with perceptual or adversarial losses. To improve image quality and remove dependency on the bitrate we propose to decode with iterative diffusion models. We condition the decoding process on a vector-quantized image representation, as well as a global image description to provide additional context. We dub our model `PerCo'' for ``perceptual compression'', and compare it to state-of-the-art codecs at rates from 0.1 down to 0.003 bits per pixel. The latter rate is more than an order of magnitude smaller than those considered in most prior work, compressing a 512x768 Kodak image with less than 153 bytes. Despite this ultra-low bitrate, our approach maintains the ability to reconstruct

realistic images. We find that our model leads to reconstructions with state-of-the-art visual quality as measured by FID and KID. As predicted by rate-distortion-perception theory, visual quality is less dependent on the bitrate than previous methods.
****
Guoqiang Zhang,Kenta Niwa,W. Bastiaan Kleijn
On Accelerating Diffusion-Based Sampling Processes via Improved Integration Approximation
A popular approach to sample a diffusion-based generative model is to solve an ordinary differential equation (ODE). In existing samplers, the coefficients of the ODE solvers are pre-determined by the ODE formulation, the reverse discrete timesteps, and the employed ODE methods. In this paper, we consider accelerating several popular ODE-based sampling processes (including EDM, DDIM, and DPM-Solver) by optimizing certain coefficients via improved integration approximation (IIA). We propose to minimize, for each time step, a mean squared error (MSE) function with respect to the selected coefficients. The MSE is constructed by applying the original ODE solver for a set of fine-grained timesteps, which in principle provides a more accurate integration approximation in predicting the next diffusion state. The proposed IIA technique does not require any change of a pre-trained model, and only introduces a very small computational overhead for solving a number of quadratic optimization problems. Extensive experiments show that considerably better FID scores can be achieved by using IIA-EDM, IIA-DDIM, and IIA-DPM-Solver than the original counterparts when the neural function evaluation (NFE) is small (i.e., less than 25).
****
Xin Yu,Yuan-Chen Guo,Yangguang Li,Ding Liang,Song-Hai Zhang,XIAOJUAN QI
Text-to-3D with Classifier Score Distillation
Text-to-3D generation has made remarkable progress recently, particularly with methods based on Score Distillation Sampling (SDS) that leverages pre-trained 2D diffusion models. While the usage of classifier-free guidance is well acknowledged to be crucial for successful optimization, it is considered an auxiliary trick rather than the most essential component. In this paper, we re-evaluate the role of classifier-free guidance in score distillation and discover a surprising finding: the guidance alone is enough for effective text-to-3D generation tasks. We name this method Classifier Score Distillation (CSD), which can be interpreted as using an implicit classification model for generation. This new perspective reveals new insights for understanding existing techniques. We validate the effectiveness of CSD across a variety of text-to-3D tasks including shape generation, texture synthesis, and shape editing, achieving results superior to those of state-of-the-art methods. Our project page is https://xinyu-andy.github.io/Classifier-Score-Distillation
****
Kostadin Garov,Dimitar Iliev Dimitrov,Nikola Jovanovi■,Martin Vechev
Hiding in Plain Sight: Disguising Data Stealing Attacks in Federated Learning
Malicious server (MS) attacks have enabled the scaling of data stealing in federated learning to large batch sizes and secure aggregation, settings previously considered private. However, many concerns regarding the client-side detectability of MS attacks were raised, questioning their practicality. In this work, for the first time, we thoroughly study client-side detectability. We first demonstrate that all prior MS attacks are detectable by principled checks, and formulate a necessary set of requirements that a practical MS attack must satisfy. Next, we propose SEER, a novel attack framework that satisfies these requirements. The key insight of SEER is the use of a secret decoder, jointly trained with the shared model. We show that SEER can steal user data from gradients of realistic networks, even for large batch sizes of up to 512 and under secure aggregation. Our work is a promising step towards assessing the true vulnerability of federated learning in real-world settings.
****
Yanpeng Zhao,Siyu Gao,Yunbo Wang,Xiaokang Yang
DynaVol: Unsupervised Learning for Dynamic Scenes through Object-Centric Voxeliz

ation
Unsupervised learning of object-centric representations in dynamic visual scenes is challenging. Unlike most previous approaches that learn to decompose 2D images, we present DynaVol, a 3D scene generative model that unifies geometric structures and object-centric learning in a differentiable volume rendering framework. The key idea is to perform object-centric voxelization to capture the 3D nature of the scene, which infers the probability distribution over objects at individual spatial locations. These voxel features evolve over time through a canonical-space deformation function, forming the basis for global representation learning via slot attention. The voxel features and global features are complementary and are both leveraged by a compositional NeRF decoder for volume rendering. DynaVol remarkably outperforms existing approaches for unsupervised dynamic scene decomposition. Once trained, the explicitly meaningful voxel features enable additional capabilities that 2D scene decomposition methods cannot achieve: it is possible to freely edit the geometric shapes or manipulate the motion trajectories of the objects.
****
Robin Staab,Mark Vero,Mislav Balunovic,Martin Vechev
Beyond Memorization: Violating Privacy via Inference with Large Language Models
Current privacy research on large language models (LLMs) primarily focuses on the issue of extracting memorized training data. At the same time, models' inference capabilities have increased drastically. This raises the key question of whether current LLMs could violate individuals' privacy by inferring personal attributes from text given at inference time. In this work, we present the first comprehensive study on the capabilities of pretrained LLMs to infer personal attributes from text. We construct a dataset consisting of real Reddit profiles, and show that current LLMs can infer a wide range of personal attributes (e.g., location, income, sex), achieving up to 85% top-1 and 95% top-3 accuracy at a fraction of the cost (100x) and time (240x) required by humans. As people increasingly interact with LLM-powered chatbots across all aspects of life, we also explore the emerging threat of privacy-invasive chatbots trying to extract personal information through seemingly benign questions. Finally, we show that common mitigations, i.e., text anonymization and model alignment, are currently ineffective at protecting user privacy against LLM inference. Our findings highlight that current LLMs can infer personal data at a previously unattainable scale. In the absence of working defenses, we advocate for a broader discussion around LLM privacy implications beyond memorization, striving for stronger and wider privacy protection.
****
Site Bai,Brian Bullins
Local Composite Saddle Point Optimization
Distributed optimization (DO) approaches for saddle point problems (SPP) have recently gained in popularity due to the critical role they play in machine learning (ML). Existing works mostly target smooth unconstrained objectives in Euclidean space, whereas ML problems often involve constraints or non-smooth regularization, which results in a need for composite optimization. Moreover, although non-smooth regularization often serves to induce structure (e.g., sparsity), standard aggregation schemes in distributed optimization break this structure. Addressing these issues, we propose Federated Dual Extrapolation (FeDualEx), an extra-step primal-dual algorithm with local updates, which is the first of its kind to encompass both saddle point optimization and composite objectives under the distributed paradigm. Using a generalized notion of Bregman divergence, we analyze its convergence and communication complexity in the homogeneous setting. Furthermore, the empirical evaluation demonstrates the effectiveness of FeDualEx for inducing structure in these challenging settings.
****
Junyi Li,Feihu Huang,Heng Huang
FedDA: Faster Adaptive Gradient Methods for Federated Constrained Optimization
Federated learning (FL) is an emerging learning paradigm where a set of distributed clients learns a task under the coordination of a server. The FedAvg algorit

hm is one of the most widely used methods in FL. In FedAvg, the learning rate is
 a constant rather than changing adaptively. Adaptive gradient methods have demo
nstrated superior performance over the constant learning rate schedules in non-d
istributed settings, and they have recently been adapted to FL. However, the maj
ority of these methods are designed for unconstrained settings. Meanwhile, many
crucial FL applications, like disease diagnosis and biomarker identification, of
ten rely on constrained formulations such as Lasso and group Lasso. It remains a
n open question as to whether adaptive gradient methods can be effectively appli
ed to FL problems with constrains. In this work, we introduce \textbf{FedDA}, a
novel adaptive gradient framework for FL. This framework utilizes a restarted du
al averaging technique and is compatible with a range of gradient estimation met
hods and adaptive learning rate schedules. Specifically, an instantiation of ou
r framework FedDA-MVR achieves sample complexity $\tilde{O}(K^{-1}\epsilon^{-1.5
})$ and communication complexity $\tilde{O}(K^{-0.25}\epsilon^{-1.25})$ for find
ing a stationary point ϵ in the constrained setting with K be the num
ber of clients. We conduct experiments over both constrained and unconstrained t
asks to confirm the effectiveness of our approach.
****
Tianxin Wei,Bowen Jin,Ruirui Li,Hansi Zeng,Zhengyang Wang,Jianhui Sun,Qingyu Yin
,Hanqing Lu,Suhang Wang,Jingrui He,Xianfeng Tang
Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Ge
nerative Recommendation and Beyond
Developing a universal model that can effectively harness heterogeneous resource
s and respond to a wide range of personalized needs has been a longstanding comm
unity aspiration. Our daily choices, especially in domains like fashion and reta
il, are substantially shaped by multi-modal data, such as pictures and textual d
escriptions. These modalities not only offer intuitive guidance but also cater t
o personalized user preferences. However, the predominant personalization approa
ches mainly focus on ID or text-based recommendation problems, failing to compre
hend the information spanning various tasks or modalities. In this paper, our go
al is to establish a Unified paradigm for Multi-modal Personalization systems (U
niMP), which effectively leverages multi-modal data while eliminating the comple
xities associated with task- and modality-specific customization. We argue that
the advancements in foundational generative modeling have provided the flexibili
ty and effectiveness necessary to achieve the objective. In light of this, we de
velop a generic and extensible personalization generative framework, that can ha
ndle a wide range of personalized needs including item recommendation, product s
earch, preference prediction, explanation generation, and further user-guided im
age generation. Our methodology enhances the capabilities of foundational langua
ge models for personalized tasks by seamlessly ingesting interleaved cross-modal
 user history information, ensuring a more precise and customized experience for
 users. To train and evaluate the proposed multi-modal personalized tasks, we al
so introduce a novel and comprehensive benchmark covering a variety of user requ
irements. Our experiments on the real-world benchmark showcase the model's poten
tial, outperforming competitive methods specialized for each task.
****
Yuanfeng Ji,Chongjian GE,Weikai Kong,Enze Xie,Zhengying Liu,Zhenguo Li,Ping Luo
Large Language Models as Automated Aligners for benchmarking Vision-Language M
odels
With the advancements in Large Language Models (LLMs), Vision-Language Models (V
LMs) have reached a new level of sophistication, showing notable competence in e
xecuting intricate cognition and reasoning tasks. However, existing evaluation b
enchmarks, primarily relying on rigid, hand-crafted datasets to measure task-spe
cific performance, face significant limitations in assessing the alignment of th
ese increasingly anthropomorphic models with human intelligence. In this work, w
e address the limitations via Auto-Bench, which delves into exploring LLMs as pr
oficient aligners, measuring the alignment between VLMs and human intelligence a
nd value through automatic data curation and assessment. Specifically, for data
curation, Auto-Bench utilizes LLMs (e.g., GPT-4) to automatically generate a vas
t set of question-answer-reasoning triplets via prompting on visual symbolic rep

resentations (e.g., captions, object locations, instance relationships, and etc. The curated data closely matches human intent, owing to the extensive world knowledge embedded in LLMs. Through this pipeline, a total of 28.5K human-verified and 3,504K unfiltered question-answer-reasoning triplets have been curated, covering 4 primary abilities and 16 sub-abilities. We subsequently engage LLMs like GPT-3.5 to serve as judges, implementing the quantitative and qualitative automated assessments to facilitate a comprehensive evaluation of VLMs. Our validation results reveal that LLMs are proficient in both evaluation data curation and model assessment, achieving an average agreement rate of 85%. We envision Auto-Bench as a flexible, scalable, and comprehensive benchmark for evaluating the evolving sophisticated VLMs.
****
Jorge Fernandez-de-Cossio-Diaz,Clément Roussel,Simona Cocco,Remi Monasson
Accelerated Sampling with Stacked Restricted Boltzmann Machines
Sampling complex distributions is an important but difficult objective in various fields, including physics, chemistry, and statistics. An improvement of standard Monte Carlo (MC) methods, intensively used in particular in the context of disordered systems, is Parallel Tempering, also called replica exchange MC, in which a sequence of MC Markov chains at decreasing temperatures are run in parallel and can swap their configurations. In this work we apply the ideas of parallel tempering in the context of restricted Boltzmann machines (RBM), a paradigm of unsupervised architectures, capable to learn complex, multimodal distributions. Inspired by Deep Tempering, an approach introduced for deep belief networks, we show how to learn on top of the first RBM a stack of nested RBMs, using the representations of a RBM as ’data’ for the next one along the stack. In our Stacked Tempering approach the hidden configurations of a machine can be exchanged with the visible configurations of the next one in the stack. Replica exchanges between the different RBMs is facilitated by the increasingly clustered representations learnt by deeper RBMs, allowing for fast transitions between the different modes of the data distribution. Analytical calculations of mixing times in a simplified theoretical setting shed light on why Stacked Tempering works, and how hyperparameters, such as the aspect ratios of the RBMs and weight regularization should be chosen. We illustrate the efficiency of the Stacked Tempering method with respect to standard and replica exchange MC on several datasets: MNIST, in-silico Lattice Proteins, and the 2D-Ising model.
****
Chang Chen,Fei Deng,Kenji Kawaguchi,Caglar Gulcehre,Sungjin Ahn
Simple Hierarchical Planning with Diffusion
Diffusion-based generative methods have proven effective in modeling trajectories with offline datasets. However, they often face computational challenges and can falter in generalization, especially in capturing temporal abstractions for long-horizon tasks. To overcome this, we introduce the Hierarchical Diffuser, a simple, fast, yet effective planning method combining the advantages of hierarchical and diffusion-based planning. Our model adopts a “jumpy” planning strategy at the high level, which allows it to have a larger receptive field but at a lower computational cost—a crucial factor for diffusion-based planning methods, as we have empirically verified. Additionally, the jumpy sub-goals guide our low-level planner, facilitating a fine-tuning stage and further improving our approach’s effectiveness. We conducted empirical evaluations on standard offline reinforcement learning benchmarks, demonstrating our method’s superior performance and efficiency in terms of training and planning speed compared to the non-hierarchical Diffuser as well as other hierarchical planning methods. Moreover, we explore our model’s generalization capability, particularly on how our method improves generalization capabilities on compositional out-of-distribution tasks.
****
Axel Laborieux,Friedemann Zenke
Improving equilibrium propagation without weight symmetry through Jacobian homeostasis
Equilibrium propagation (EP) is a compelling alternative to the back propagation of error algorithm (BP) for computing gradients of neural networks on biologica

l or analog neuromorphic substrates.
Still, the algorithm requires weight symmetry and infinitesimal equilibrium pert urbations, i.e., nudges, to yield unbiased gradient estimates.
Both requirements are challenging to implement in physical systems.
Yet, whether and how weight asymmetry contributes to bias is unknown because, in practice, its contribution may be masked by a finite nudge.
To address this question, we study generalized EP, which can be formulated witho ut weight symmetry, and analytically isolate the two sources of bias.
For complex-differentiable non-symmetric networks, we show that bias due to fini te nudge can be avoided by estimating exact derivatives via a Cauchy integral.
In contrast, weight asymmetry induces residual bias through poor alignment of E P's neuronal error vectors compared to BP resulting in low task performance.
To mitigate the latter issue, we present a new homeostatic objective that direct ly penalizes functional asymmetries of the Jacobian at the network's fixed point .
This homeostatic objective dramatically improves the network's ability to solve complex tasks such as ImageNet 32$\times$32.
Our results lay the theoretical groundwork for studying and mitigating the adver se effects of imperfections of physical networks on learning algorithms that rel y on the substrate's relaxation dynamics.
****
Jiatong Shi,Hirofumi Inaguma,Xutai Ma,Ilia Kulikov,Anna Sun
Multi-resolution HuBERT: Multi-resolution Speech Self-Supervised Learning with M asked Unit Prediction
Existing Self-Supervised Learning (SSL) models for speech typically process spee ch signals at a fixed resolution of 20 milliseconds. This approach overlooks the varying informational content present at different resolutions in speech signal s. In contrast, this paper aims to incorporate multi-resolution information into speech self-supervised representation learning. We introduce an SSL model that leverages a hierarchical Transformer architecture, complemented by HuBERT-style masked prediction objectives, to process speech at multiple resolutions. Experim ental results indicate that the proposed model not only achieves more efficient inference but also exhibits superior or comparable performance to the original H uBERT model over various tasks. Specifically, significant performance improvemen ts over the original HuBERT have been observed in fine-tuning experiments on the LibriSpeech speech recognition benchmark as well as in evaluations using the Sp eech Universal PERformance Benchmark (SUPERB) and Multilingual SUPERB (ML-SUPERB).
****
Uiwon Hwang,Jonghyun Lee,Juhyeon Shin,Sungroh Yoon
SF(DA)^2: Source-free Domain Adaptation Through the Lens of Data Augmentation
In the face of the deep learning model's vulnerability to domain shift, source-f ree domain adaptation (SFDA) methods have been proposed to adapt models to new, unseen target domains without requiring access to source domain data. Although t he potential benefits of applying data augmentation to SFDA are attractive, seve ral challenges arise such as the dependence on prior knowledge of class-preservi ng transformations and the increase in memory and computational requirements. In this paper, we propose Source-free Domain Adaptation Through the Lens of Data A ugmentation (SF(DA)^2), a novel approach that leverages the benefits of data a ugmentation without suffering from these challenges. We construct an augmentatio n graph in the feature space of the pretrained model using the neighbor relation ships between target features and propose spectral neighborhood clustering to id entify partitions in the prediction space. Furthermore, we propose implicit feat ure augmentation and feature disentanglement as regularization loss functions th at effectively utilize class semantic information within the feature space. Thes e regularizers simulate the inclusion of an unlimited number of augmented target features into the augmentation graph while minimizing computational and memory demands. Our method shows superior adaptation performance in SFDA scenarios, inc luding 2D image and 3D point cloud datasets and a highly imbalanced dataset.
****

Jicong Fan,Rui Chen,Zhao Zhang,Chris Ding
Neuron-Enhanced AutoEncoder Matrix Completion and Collaborative Filtering: Theory and Practice
Neural networks have shown promising performance in collaborative filtering and matrix completion but the theoretical analysis is limited and there is still room for improvement in terms of the accuracy of recovering missing values. This paper presents a neuron-enhanced autoencoder matrix completion (AEMC-NE) method and applies it to collaborative filtering. Our AEMC-NE adds an element-wise autoencoder to each output of the main autoencoder to enhance the reconstruction capability. Thus it can adaptively learn an activation function for the output layer to approximate possibly complicated response functions in real data. We provide theoretical analysis for AEMC-NE as well as AEMC to investigate the generalization ability of autoencoder and deep learning in matrix completion, considering both missing completely at random and missing not at random. We show that the element-wise neural network has the potential to reduce the generalization error bound, the data sparsity can be useful, and the prediction performance is closely related to the difference between the numbers of variables and samples. The numerical results on synthetic data and benchmark datasets demonstrated the effectiveness of AEMC-NE in comparison to many baselines.
****
Mike Lasby,Anna Golubeva,Utku Evci,Mihai Nica,Yani Ioannou
Dynamic Sparse Training with Structured Sparsity
Dynamic Sparse Training (DST) methods achieve state-of-the-art results in sparse neural network training, matching the generalization of dense models while enabling sparse training and inference. Although the resulting models are highly sparse and theoretically less computationally expensive, achieving speedups with unstructured sparsity on real-world hardware is challenging. In this work, we propose a sparse-to-sparse DST method, Structured RigL (SRigL), to learn a variant of fine-grained structured N:M sparsity by imposing a constant fan-in constraint. Using our empirical analysis of existing DST methods at high sparsity, we additionally employ a neuron ablation method which enables SRigL to achieve state-of-the-art sparse-to-sparse structured DST performance on a variety of Neural Network (NN) architectures. Using a 90% sparse linear layer, we demonstrate a real-world acceleration of 3.4×/2.5× on CPU for online inference and 1.7×/13.0× on GPU for inference with a batch size of 256 when compared to equivalent dense/unstructured (CSR) sparse layers, respectively.
****
Amin Rakhsha,Mete Kemertas,Mohammad Ghavamzadeh,Amir-massoud Farahmand
Maximum Entropy Model Correction in Reinforcement Learning
We propose and theoretically analyze an approach for planning with an approximate model in reinforcement learning that can reduce the adverse impact of model error. If the model is accurate enough, it accelerates the convergence to the true value function too. One of its key components is the MaxEnt Model Correction (MoCo) procedure that corrects the model’s next-state distributions based on a Maximum Entropy density estimation formulation. Based on MoCo, we introduce the Model Correcting Value Iteration (MoCoVI) algorithm, and its sampled-based variant MoCoDyna. We show that MoCoVI and MoCoDyna’s convergence can be much faster than the conventional model-free algorithms. Unlike traditional model-based algorithms, MoCoVI and MoCoDyna effectively utilize an approximate model and still converge to the correct value function.
****
Tianhong Li,Sangnie Bhardwaj,Yonglong Tian,Han Zhang,Jarred Barber,Dina Katabi,Guillaume Lajoie,Huiwen Chang,Dilip Krishnan
Leveraging Unpaired Data for Vision-Language Generative Models via Cycle Consistency
Current vision-language generative models rely on expansive corpora of $\textit{paired}$ image-text data to attain optimal performance and generalization capabilities. However, automatically collecting such data (e.g. via large-scale web scraping) leads to low quality and poor image-text correlation, while human annotation is more accurate but requires significant manual effort and expense. We int

roduce $\textbf{ITIT}$ ($\textbf{I}$n$\textbf{T}$egrating $\textbf{I}$mage $\tex
tbf{T}$ext): an innovative training paradigm grounded in the concept of cycle co
nsistency which allows vision-language training on $\textit{unpaired}$ image and
 text data. ITIT is comprised of a joint image-text encoder with disjoint image
and text decoders that enable bidirectional image-to-text and text-to-image gene
ration in a single framework. During training, ITIT leverages a small set of pai
red image-text data to ensure its output matches the input reasonably well in bo
th directions. Simultaneously, the model is also trained on much larger datasets
 containing only images or texts. This is achieved by enforcing cycle consistenc
y between the original unpaired samples and the cycle-generated counterparts. Fo
r instance, it generates a caption for a given input image and then uses the cap
tion to create an output image, and enforces similarity between the input and ou
tput images. Our experiments show that ITIT with unpaired datasets exhibits simi
lar scaling behavior as using high-quality paired data. We demonstrate image gen
eration and captioning performance on par with state-of-the-art text-to-image an
d image-to-text models with orders of magnitude fewer (only 3M) paired image-tex
t data. Code will be released at https://github.com/LTH14/itit.
****
Zhenghan Fang,Sam Buchanan,Jeremias Sulam
What's in a Prior? Learned Proximal Networks for Inverse Problems
Proximal operators are ubiquitous in inverse problems, commonly appearing as par
t of algorithmic strategies to regularize problems that are otherwise ill-posed.
 Modern deep learning models have been brought to bear for these tasks too, as i
n the framework of plug-and-play or deep unrolling, where they loosely resemble
proximal operators. Yet, something essential is lost in employing these purely d
ata-driven approaches: there is no guarantee that a general deep network represe
nts the proximal operator of any function, nor is there any characterization of
the function for which the network might provide some approximate proximal. This
 not only makes guaranteeing convergence of iterative schemes challenging but, m
ore fundamentally, complicates the analysis of what has been learned by these ne
tworks about their training data. Herein we provide a framework to develop *lear
ned proximal networks* (LPN), prove that they provide exact proximal operators f
or a data-driven nonconvex regularizer, and show how a new training strategy, du
bbed *proximal matching*, provably promotes the recovery of the log-prior of the
 true data distribution. Such LPN provide general, unsupervised, expressive prox
imal operators that can be used for general inverse problems with convergence gu
arantees. We illustrate our results in a series of cases of increasing complexit
y, demonstrating that these models not only result in state-of-the-art performan
ce, but provide a window into the resulting priors learned from data.
****
Jason Chun Lok Li,Steven Tin Sui Luo,Le Xu,Ngai Wong
ASMR: Activation-Sharing Multi-Resolution Coordinate Networks for Efficient Infe
rence
Coordinate network or implicit neural representation (INR) is a fast-emerging me
thod for encoding natural signals (such as images and videos) with the benefits
of a compact neural representation. While numerous methods have been proposed to
 increase the encoding capabilities of an INR, an often overlooked aspect is the
 inference efficiency, usually measured in multiply-accumulate (MAC) count. This
 is particularly critical in use cases where inference bandwidth is greatly limi
ted by hardware constraints. To this end, we propose the Activation-Sharing Mult
i-Resolution (ASMR) coordinate network that combines multi-resolution coordinate
 decomposition with hierarchical modulations. Specifically, an ASMR model enable
s the sharing of activations across grids of the data. This largely decouples it
s inference cost from its depth which is directly correlated to its reconstructi
on capability, and renders a near $O(1)$ inference complexity irrespective of th
e number of layers. Experiments show that ASMR can reduce the MAC of a vanilla S
IREN model by up to 350$\times$ while achieving an even higher reconstruction qu
ality than its SIREN baseline.
****
Joey Bose,Tara Akhound-Sadegh,Guillaume Huguet,Kilian FATRAS,Jarrid Rector-Brook

s,Cheng-Hao Liu,Andrei Cristian Nica,Maksym Korablyov,Michael M. Bronstein,Alexander Tong
SE(3)-Stochastic Flow Matching for Protein Backbone Generation
The computational design of novel protein structures has the potential to impact numerous scientific disciplines greatly. Toward this goal, we introduce \foldflow, a series of novel generative models of increasing modeling power based on the flow-matching paradigm over $3\mathrm{D}$ rigid motions---i.e. the group $\mathrm{SE(3)}$---enabling accurate modeling of protein backbones. We first introduce FoldFlow-Base, a simulation-free approach to learning deterministic continuous-time dynamics and matching invariant target distributions on $\mathrm{SE(3)}$. We next accelerate training by incorporating Riemannian optimal transport to create FoldFlow-OT, leading to the construction of both more simple and stable flows. Finally, we design \foldflowsfm, coupling both Riemannian OT and simulation-free training to learn stochastic continuous-time dynamics over $\mathrm{SE(3)}$. Our family of FoldFlow, generative models offers several key advantages over previous approaches to the generative modeling of proteins: they are more stable and faster to train than diffusion-based approaches, and our models enjoy the ability to map any invariant source distribution to any invariant target distribution over $\mathrm{SE(3)}$. Empirically, we validate FoldFlow, on protein backbone generation of up to 300 amino acids leading to high-quality designable, diverse, and novel samples.
****
Jian Chen,Ruiyi Zhang,Yufan Zhou,Changyou Chen
Towards Aligned Layout Generation via Diffusion Model with Aesthetic Constraints
Controllable layout generation refers to the process of creating a plausible visual arrangement of elements within a graphic design (*e.g.*, document and web designs) with constraints representing design intentions. Although recent diffusion-based models have achieved state-of-the-art FID scores, they tend to exhibit more pronounced misalignment compared to earlier transformer-based models. In this work, we propose the **LA**yout **C**onstraint diffusion mod**E**l (LACE), a unified model to handle a broad range of layout generation tasks, such as arranging elements with specified attributes and refining or completing a coarse layout design. The model is based on continuous diffusion models. Compared with existing methods that use discrete diffusion models, continuous state-space design can enable the incorporation of continuous aesthetic constraint functions in training more naturally. For conditional generation, we propose injecting layout conditions in the form of masks or gradient guidance during inference. Empirical results show that LACE produces high-quality layouts and outperforms existing state-of-the-art baselines. We will release our source code and model checkpoints.
****
Elan Rosenfeld,Andrej Risteski
Outliers with Opposing Signals Have an Outsized Effect on Neural Network Optimization
We identify a new phenomenon in neural network optimization which arises from the interaction of depth and a particular heavy-tailed structure in natural data. Our result offers intuitive explanations for several previously reported observations about network training dynamics, including a conceptually new cause for progressive sharpening and the edge of stability. We further draw connections to related phenomena in optimization including grokking and simplicity bias.

Experimentally, we demonstrate the significant influence of paired groups of outliers in the training data with strong \emph{opposing signals}: consistent, large magnitude features which dominate the network output and occur in both groups with similar frequency.
Due to these outliers, early optimization enters a narrow valley which carefully balances the opposing groups; subsequent sharpening causes their loss to rise rapidly, oscillating between high on one group and then the other, until the overall loss spikes. We complement these experiments with a theoretical analysis of a two-layer linear network on a simple model of opposing signals.

Our finding enables new qualitative predictions of training behavior which we co
nfirm experimentally. It also provides a new lens through which to study and imp
rove modern training practices for stochastic optimization. For instance, we ide
ntify two small modifications to Momentum SGD which result in performance that m
atches adaptive methods in settings where it has traditionally faltered---includ
ing on attention models.
****
Xunpeng Huang,Hanze Dong,Yifan HAO,Yian Ma,Tong Zhang
Reverse Diffusion Monte Carlo
We propose a Monte Carlo sampler from the reverse diffusion process. Unlike the
practice of diffusion models, where the intermediary updates---the score functio
ns---are learned with a neural network, we transform the score matching problem
into a mean estimation one.
By estimating the means of the regularized posterior distributions, we derive a
novel Monte Carlo sampling algorithm called reverse diffusion Monte Carlo (rdMC)
, which is distinct from the Markov chain Monte Carlo (MCMC) methods. We determi
ne the sample size from the error tolerance and the properties of the posterior
distribution to yield an algorithm that can approximately sample the target dist
ribution with any desired accuracy. Additionally, we demonstrate and prove under
 suitable conditions that sampling with rdMC can be significantly faster than th
at with MCMC. For multi-modal target distributions such as those in Gaussian mi
xture models, rdMC greatly improves over the Langevin-style MCMC sampling method
s both theoretically and in practice. The proposed rdMC method offers a new pers
pective and solution beyond classical MCMC algorithms for the challenging comple
x distributions.
****
Shuai Zhao,Xiaohan Wang,Linchao Zhu,Yi Yang
Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Lan
guage Models
One fascinating aspect of pre-trained vision-language models (VLMs) learning und
er language supervision is their impressive zero-shot generalization capability.
However, this ability is hindered by distribution shifts between the training an
d testing data.
Previous test time adaptation (TTA) methods for VLMs in zero-shot classification
 rely on minimizing the entropy of model outputs, tending to be stuck in incorre
ct model predictions.
In this work, we propose TTA with feedback to rectify the model output and preve
nt the model from becoming blindly confident.
Specifically, a CLIP model is adopted as the reward model during TTA and provide
s feedback for the VLM.
Given a single test sample,
the VLM is forced to maximize the CLIP reward between the input and sampled resu
lts from the VLM output distribution.
The proposed \textit{reinforcement learning with CLIP feedback~(RLCF)} framework
 is highly flexible and universal.
Beyond the classification task, with task-specific sampling strategies and a pro
per reward baseline choice, RLCF can be easily extended to not only discriminati
on tasks like retrieval but also generalization tasks like image captioning,
improving the zero-shot generalization capacity of VLMs.
According to the characteristics of these VL tasks, we build different fully TTA
 pipelines with RLCF to improve the zero-shot generalization ability of various
VLMs.
Extensive experiments along with promising
empirical results demonstrate the effectiveness of RLCF.
The code is available at https://github.com/mzhaoshuai/RLCF.
****
Linus Bleistein,Agathe Guilloux
On the Generalization and Approximation Capacities of Neural Controlled Differen
tial Equations
Neural Controlled Differential Equations (NCDE) are a state-of-the-art tool for

supervised learning with irregularly sampled time series (Kidger 2020). However, no theoretical analysis of their performance has been provided yet, and it remains unclear in particular how the roughness of the sampling affects their predictions. By merging the rich theory of controlled differential equations (CDE) and Lipschitz-based measures of the complexity of deep neural nets, we take a first step towards the theoretical understanding of NCDE. Our first result is a sampling-dependant generalization bound for this class of predictors. In a second time, we leverage the continuity of the flow of CDEs to provide a detailed analysis of both the sampling-induced bias and the approximation bias. Regarding this last result, we show how classical approximation results on neural nets may transfer to NCDE. Our theoretical results are validated through a series of experiments.

****

Shashank Gupta,Vaishnavi Shrivastava,Ameet Deshpande,Ashwin Kalyan,Peter Clark,Ashish Sabharwal,Tushar Khot

Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs

Recent work has showcased the ability of large-scale language models (LLMs) to embody diverse personas in their responses, exemplified by prompts like "_You are Julius Caesar. Compose a rap about Climate Change._" However, it remains unclear how these persona assignments indirectly influence LLMs' core capabilities. We present the first extensive study of this in the context of LLMs' ability to perform basic reasoning. Our study encompasses 16 personas spanning 5 diverse groups (race, gender, religion, disability, and political affiliation), across 24 reasoning datasets in diverse domains such as mathematics, history, law, ethics, and more. Our findings unveil that while LLMs, such as ChatGPT, overtly reject stereotypes when explicitly asked ("_Are Black people inept at mathematics?_"), they tend to manifest implicit stereotypical and often erroneous presumptions when prompted to take on a persona (e.g., abstentions in rationales such as "_As a Black person, I am unable to answer this question as it requires math knowledge_"). This results in substantial disparities in reasoning performance among personas. This inherent 'deep' bias permeates extensively, leading to a statistically significant performance drop in over 95\% of our datasets for certain personas, with as much as 70\% relative drop in accuracy on select datasets. Beyond explicit abstentions, these models also have implicitly biased reasoning not evident in their responses. We find that simple prompt-based mitigation approaches have minimal impact. Our findings serve as a cautionary tale that the practice of assigning personas to LLMs---a trend on the rise---can surface their deep-rooted biases and have unforeseeable and detrimental side-effects.

****

Ainaz Eftekhar,Kuo-Hao Zeng,Jiafei Duan,Ali Farhadi,Aniruddha Kembhavi,Ranjay Krishna

Selective Visual Representations Improve Convergence and Generalization for Embodied AI

Embodied AI models often employ off the shelf vision backbones like CLIP to encode their visual observations. Although such general purpose representations encode rich syntactic and semantic information about the scene, much of this information is often irrelevant to the specific task at hand. This introduces noise within the learning process and distracts the agent's focus from task-relevant visual cues.

Inspired by selective attention in humans—the process through which people filter their perception based on their experiences, knowledge, and the task at hand—we introduce a parameter-efficient approach to filter visual stimuli for embodied AI.

Our approach induces a task-conditioned bottleneck using a small learnable codebook module. This codebook is trained jointly to optimize task reward and acts as a task-conditioned selective filter over the visual observation.

Our experiments showcase state-of-the-art performance for object goal navigation and object displacement across 5 benchmarks, ProcTHOR, ArchitecTHOR, RoboTHOR, AI2-iTHOR, and ManipulaTHOR. The filtered representations produced by the codebook are also able generalize better and converge faster when adapted to other s

imulation environments such as Habitat. Our qualitative analyses show that agent
s explore their environments more effectively and their representations retain t
ask-relevant information like target object recognition while ignoring superfluo
us information about other objects.
****
Xiaxia Wang,David Jaime Tena Cucala,Bernardo Cuenca Grau,Ian Horrocks
Faithful Rule Extraction for Differentiable Rule Learning Models
There is increasing interest in methods for extracting interpretable rules from
ML models trained to solve a wide range of tasks over knowledge graphs (KGs), su
ch as KG completion, node classification, question answering and recommendation.
 Many such approaches, however, lack formal guarantees establishing the precise
relationship between the model and the extracted rules, and this lack of assuran
ce becomes especially problematic when the extracted rules are applied in safety
-critical contexts or to ensure compliance with legal requirements. Recent resea
rch has examined whether the rules derived from the influential Neural-LP model
exhibit soundness (or completeness), which means that the results obtained by ap
plying the model to any dataset always contain (or are contained in) the results
 obtained by applying the rules to the same dataset. In this paper, we extend th
is analysis to the context of DRUM, an approach that has demonstrated superior p
ractical performance. After observing that the rules currently extracted from a
DRUM model can be unsound and/or incomplete, we propose a novel algorithm where
the output rules, expressed in an extension of Datalog, ensure both soundness an
d completeness. This algorithm, however, can be inefficient in practice and henc
e we propose additional constraints to DRUM models facilitating rule extraction,
 albeit at the expense of reduced expressive power.
****
Peter Sorrenson,Felix Draxler,Armand Rousselot,Sander Hummerich,Lea Zimmermann,U
llrich Koethe
Lifting Architectural Constraints of Injective Flows
Normalizing Flows explicitly maximize a full-dimensional likelihood on the train
ing data. However, real data is typically only supported on a lower-dimensional
manifold leading the model to expend significant compute on modeling noise. Inje
ctive Flows fix this by jointly learning a manifold and the distribution on it.
So far, they have been limited by restrictive architectures and/or high computat
ional cost. We lift both constraints by a new efficient estimator for the maximu
m likelihood loss, compatible with free-form bottleneck architectures. We furthe
r show that naively learning both the data manifold and the distribution on it c
an lead to divergent solutions, and use this insight to motivate a stable maximu
m likelihood training objective. We perform extensive experiments on toy, tabula
r and image data, demonstrating the competitive performance of the resulting mod
el.
****
Ali Hatamizadeh,Greg Heinrich,Hongxu Yin,Andrew Tao,Jose M. Alvarez,Jan Kautz,Pa
vlo Molchanov
FasterViT: Fast Vision Transformers with Hierarchical Attention
We design a new family of hybrid CNN-ViT neural networks, named FasterViT, with
a focus on high image throughput for computer vision (CV) applications. FasterVi
T combines the benefits of fast local representation learning in CNNs and global
 modeling properties in ViT. Our newly introduced Hierarchical Attention (HAT) a
pproach decomposes global self-attention with quadratic complexity into a multi-
level attention with reduced computational costs. We benefit from efficient wind
ow-based self-attention. Each window has access to dedicated carrier tokens that
 participate in local and global representation learning. At a high level, globa
l self-attentions enable the efficient cross-window communication at lower costs
. FasterViT achieves a SOTA Pareto-front in terms of accuracy and image throughp
ut. We have extensively validated its effectiveness on various CV tasks includin
g classification, object detection and segmentation. We also show that HAT can b
e used as a plug-and-play module for existing networks and enhance them. We furt
her demonstrate significantly faster and more accurate performance than competit
ive counterparts for images with high resolution. Code is available at https://g

ithub.com/NVlabs/FasterViT.
****
Matteo Alleman,Jack Lindsey,Stefano Fusi
Task structure and nonlinearity jointly determine learned representational geome
try
The utility of a learned neural representation depends on how well its geometry supports performance in downstream tasks. This geometry depends on the structure of the inputs, the structure of the target outputs, and on the architecture of the network. By studying the learning dynamics of networks with one hidden laye r, we discovered that the network's activation function has an unexpectedly stro ng impact on the representational geometry: Tanh networks tend to learn represen tations that reflect the structure of the target outputs, while ReLU networks re tain more information about the structure of the raw inputs. This difference is consistently observed across a broad class of parameterized tasks in which we mo dulated the degree of alignment between the geometry of the task inputs and that of the task labels. We analyzed the learning dynamics in weight space and show how the differences between the networks with Tanh and ReLU nonlinearities arise from the asymmetric saturation of ReLU, which leads feature neurons to speciali ze for different regions of input space. Feature neurons in Tanh networks, by co ntrast, tend to inherit the task label structure. Consequently, when the target outputs are low dimensional, Tanh networks generate neural representations that are more disentangled than those obtained with a ReLU nonlinearity. Our findings shed light on the interplay between input-output geometry, nonlinearity, and le arned representations in neural networks.
****
Stylianos Poulakakis-Daktylidis,Hadi Jamali-Rad
BECLR: Batch Enhanced Contrastive Few-Shot Learning
Learning quickly from very few labeled samples is a fundamental attribute that s eparates machines and humans in the era of deep representation learning. Unsuper vised few-shot learning (U-FSL) aspires to bridge this gap by discarding the rel iance on annotations at training time. Intrigued by the success of contrastive l earning approaches in the realm of U-FSL, we structurally approach their shortco mings in both pretraining and downstream inference stages. We propose a novel Dy namic Clustered mEmory (DyCE) module to promote a highly separable latent repres entation space for enhancing positive sampling at the pretraining phase and infu sing implicit class-level insights into unsupervised contrastive learning. We th en tackle the, somehow overlooked yet critical, issue of sample bias at the few-shot inference stage. We propose an iterative Optimal Transport-based distributi on Alignment (OpTA) strategy and demonstrate that it efficiently addresses the p roblem, especially in low-shot scenarios where FSL approaches suffer the most fr om sample bias. We later on discuss that DyCE and OpTA are two intertwined piece s of a novel end-to-end approach (we coin as BECLR), constructively magnifying e ach other's impact. We then present a suite of extensive quantitative and qualit ative experimentation to corroborate that BECLR sets a new state-of-the-art acro ss ALL existing U-FSL benchmarks (to the best of our knowledge), and significant ly outperforms the best of the current baselines (codebase available at https:// github.com/stypoumic/BECLR).
****
Emanuele Palumbo,Laura Manduchi,Sonia Laguna,Daphné Chopard,Julia E Vogt
Deep Generative Clustering with Multimodal Diffusion Variational Autoencoders
Multimodal VAEs have recently gained significant attention as generative models for weakly-supervised learning with multiple heterogeneous modalities. In parall el, VAE-based methods have been explored as probabilistic approaches for cluster ing tasks. At the intersection of these two research directions, we propose a no vel multimodal VAE model in which the latent space is extended to learn data clu sters, leveraging shared information across modalities. Our experiments show tha t our proposed model improves generative performance over existing multimodal VA Es, particularly for unconditional generation. Furthermore, we propose a post-ho c procedure to automatically select the number of true clusters thus mitigating critical limitations of previous clustering frameworks. Notably, our method favo

rably compares to alternative clustering approaches, in weakly-supervised settin
gs. Finally, we integrate recent advancements in diffusion models into the propo
sed method to improve generative quality for real-world images.
****
Aliyah R. Hsu,Yeshwanth Cherapanamjeri,Briton Park,Tristan Naumann,Anobel Odisho
,Bin Yu
Diagnosing Transformers: Illuminating Feature Spaces for Clinical Decision-Makin
g
Pre-trained transformers are often fine-tuned to aid clinical decision-making us
ing limited clinical notes. Model interpretability is crucial, especially in hig
h-stakes domains like medicine, to establish trust and ensure safety, which requ
ires human engagement. We introduce SUFO, a systematic framework that enhances i
nterpretability of fine-tuned transformer feature spaces. SUFO utilizes a range
of analytic and visualization techniques, including Supervised probing, Unsuperv
ised similarity analysis, Feature dynamics, and Outlier analysis to address key
questions about model trust and interpretability (e.g. model suitability for a t
ask, feature space evolution during fine-tuning, and interpretation of fine-tune
d features and failure modes). We conduct a case study investigating the impact
of pre-training data where we focus on real-world pathology classification tasks
, and validate our findings on MedNLI. We evaluate five 110M-sized pre-trained t
ransformer models, categorized into general-domain (BERT, TNLR), mixed-domain (B
ioBERT, Clinical BioBERT), and domain-specific (PubMedBERT) groups. Our SUFO ana
lyses reveal that: (1) while PubMedBERT, the domain-specific model, contains val
uable information for fine-tuning, it can overfit to minority classes when class
 imbalances exist. In contrast, mixed-domain models exhibit greater resistance t
o overfitting, suggesting potential improvements in domain-specific model robust
ness; (2) in-domain pre-training accelerates feature disambiguation during fine-
tuning; and (3) feature spaces undergo significant sparsification during this pr
ocess, enabling clinicians to identify common outlier modes among fine-tuned mod
els as demonstrated in this paper. These findings showcase the utility of SUFO i
n enhancing trust and safety when using transformers in medicine, and we believe
 SUFO can aid practitioners in evaluating fine-tuned language models (LMs) for o
ther applications in medicine and in more critical domains.
****
Haiyan Jiang,Vincent Zoonekynd,Giulia De Masi,Bin Gu,Huan Xiong
TAB: Temporal Accumulated Batch Normalization in Spiking Neural Networks
Spiking Neural Networks (SNNs) are attracting growing interest for their energy-
efficient computing when implemented on neuromorphic hardware. However, directly
 training SNNs, even adopting batch normalization (BN), is highly challenging du
e to their non-differentiable activation function and the temporally delayed acc
umulation of outputs over time.
    For SNN training, this temporal accumulation gives rise to Temporal Covariat
e Shifts (TCS) along the temporal dimension, a phenomenon that would become incr
easingly pronounced with layer-wise computations across multiple layers and mult
iple time-steps.
    In this paper, we introduce TAB (Temporal Accumulated Batch Normalization),
a novel SNN batch normalization method that addresses the temporal covariate shi
ft issue by aligning with neuron dynamics (specifically the accumulated membrane
 potential) and utilizing temporal accumulated statistics for data normalization
.
    Within its framework, TAB effectively encapsulates the historical temporal d
ependencies that underlie the membrane potential accumulation process, thereby e
stablishing a natural connection between neuron dynamics and TAB batch normaliza
tion.
    Experimental results on CIFAR-10, CIFAR-100, and DVS-CIFAR10 show that our T
AB method outperforms other state-of-the-art methods.
****
Hee Suk Yoon,Eunseop Yoon,Joshua Tian Jin Tee,Mark A. Hasegawa-Johnson,Yingzhen
Li,Chang D. Yoo
C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Fe

ature Dispersion
In deep learning, test-time adaptation has gained attention as a method for mode
l fine-tuning without the need for labeled data. A prime exemplification is the
recently proposed test-time prompt tuning for large-scale vision-language models
 such as CLIP. Unfortunately, these prompts have been mainly developed to improv
e accuracy, overlooking the importance of calibration—a crucial aspect for quant
ifying prediction uncertainty. However, traditional calibration methods rely on
substantial amounts of labeled data, making them impractical for test-time scena
rios. To this end, this paper explores calibration during test-time prompt tunin
g by leveraging the inherent properties of CLIP. Through a series of observation
s, we find that the prompt choice significantly affects the calibration in CLIP,
 where the prompts leading to higher text feature dispersion result in better-ca
librated predictions. Introducing the Average Text Feature Dispersion (ATFD), we
 establish its relationship with calibration error and present a novel method, C
alibrated Test-time Prompt Tuning (C-TPT), for optimizing prompts during test-ti
me with enhanced calibration. Through extensive experiments on different CLIP ar
chitectures and datasets, we show that C-TPT can effectively improve the calibra
tion of test-time prompt tuning without needing labeled data.
****
Gregoire Deletang,Anian Ruoss,Paul-Ambroise Duquenne,Elliot Catt,Tim Genewein,Ch
ristopher Mattern,Jordi Grau-Moya,Li Kevin Wenliang,Matthew Aitchison,Laurent Or
seau,Marcus Hutter,Joel Veness
Language Modeling Is Compression
It has long been established that predictive models can be transformed into loss
less compressors and vice versa. Incidentally, in recent years, the machine lear
ning community has focused on training increasingly large and powerful self-supe
rvised (language) models. Since these large language models exhibit impressive p
redictive capabilities, they are well-positioned to be strong compressors. In th
is work, we advocate for viewing the prediction problem through the lens of comp
ression and evaluate the compression capabilities of large (foundation) models.
We show that large language models are powerful general-purpose predictors and t
hat the compression viewpoint provides novel insights into scaling laws, tokeniz
ation, and in-context learning. For example, Chinchilla 70B, while trained prima
rily on text, compresses ImageNet patches to 43.4% and LibriSpeech samples to 16
.4% of their raw size, beating domain-specific compressors like PNG (58.5%) or F
LAC (30.3%), respectively. Finally, we show that the prediction-compression equi
valence allows us to use any compressor (like gzip) to build a conditional gener
ative model.
****
Joan Puigcerver,Carlos Riquelme Ruiz,Basil Mustafa,Neil Houlsby
From Sparse to Soft Mixtures of Experts
Sparse mixture of expert architectures (MoEs) scale model capacity without signi
ficant increases in training or inference costs.
Despite their success, MoEs suffer from a number of issues: training instability
, token dropping, inability to scale the number of experts, or ineffective finet
uning.
In this work, we propose Soft MoE, a fully-differentiable sparse Transformer tha
t addresses these challenges, while maintaining the benefits of MoEs.
Soft MoE performs an implicit soft assignment by passing different weighted comb
inations of all input tokens to each expert.
As in other MoEs, experts in Soft MoE only process a subset of the (combined) to
kens, enabling larger model capacity (and performance) at lower inference cost.
In the context of visual recognition, Soft MoE greatly outperforms dense Transfo
rmers (ViTs) and popular MoEs (Tokens Choice and Experts Choice).
Soft MoE scales well: Soft MoE Huge/14 with 128 experts in 16 MoE layers has ove
r 40x more parameters than ViT Huge/14, with only 2% increased inference time, a
nd substantially better quality.
****
Tongxin Yin,Jean-Francois Ton,Ruocheng Guo,Yuanshun Yao,Mingyan Liu,Yang Liu
Fair Classifiers that Abstain without Harm

In critical applications, it is vital for classifiers to defer decision-making to humans. We propose a post-hoc method that makes existing classifiers selectively abstain from predicting certain samples. Our abstaining classifier is incentivized to maintain the original accuracy for each sub-population (i.e. no harm) while achieving a set of group fairness definitions to a user specified degree. To this end, we design an Integer Programming (IP) procedure that assigns abstention decisions for each training sample to satisfy a set of constraints. To generalize the abstaining decisions to test samples, we then train a surrogate model to learn the abstaining decisions based on the IP solutions in an end-to-end manner. We analyze the feasibility of the IP procedure to determine the possible abstention rate for different levels of unfairness tolerance and accuracy constraint for achieving no harm. To the best of our knowledge, this work is the first to identify the theoretical relationships between the constraint parameters and the required abstention rate. Our theoretical results are important since a high abstention rate is often infeasible in practice due to a lack of human resources. Our framework outperforms existing methods in terms of fairness disparity without sacrificing accuracy at similar abstention rates.
****
Huy Nguyen,Pedram Akbarian,Fanqi Yan,Nhat Ho
Statistical Perspective of Top-K Sparse Softmax Gating Mixture of Experts
Top-K sparse softmax gating mixture of experts has been widely used for scaling up massive deep-learning architectures without increasing the computational cost. Despite its popularity in real-world applications, the theoretical understanding of that gating function has remained an open problem. The main challenge comes from the structure of the top-K sparse softmax gating function, which partitions the input space into multiple regions with distinct behaviors. By focusing on a Gaussian mixture of experts, we establish theoretical results on the effects of the top-K sparse softmax gating function on both density and parameter estimations. Our results hinge upon defining novel loss functions among parameters to capture different behaviors of the input regions. When the true number of experts $k_{\ast}$ is known, we demonstrate that the convergence rates of density and parameter estimations are both parametric on the sample size. However, when $k_{\ast}$ becomes unknown and the true model is over-specified by a Gaussian mixture of k experts where $k > k_{\ast}$, our findings suggest that the number of experts selected from the top-K sparse softmax gating function must exceed the total cardinality of a certain number of Voronoi cells associated with the true parameters to guarantee the convergence of the density estimation. Moreover, while the density estimation rate remains parametric under this setting, the parameter estimation rates become substantially slow due to an intrinsic interaction between the softmax gating and expert functions.
****
Sharon Lee,Yunzhi Zhang,Shangzhe Wu,Jiajun Wu
Language-Informed Visual Concept Learning
Our understanding of the visual world is centered around various concept axes, characterizing different aspects of visual entities. While different concept axes can be easily specified by language, e.g., color, the exact visual nuances along each axis often exceed the limitations of linguistic articulations, e.g., a particular style of painting. In this work, our goal is to learn a language-informed visual concept representation, by simply distilling large pre-trained vision-language models. Specifically, we train a set of concept encoders to encode the information pertinent to a set of language-informed concept axes, with an objective of reproducing the input image through a pre-trained Text-to-Image (T2I) model. To encourage better disentanglement of different concept encoders, we anchor the concept embeddings to a set of text embeddings obtained from a pre-trained Visual Question Answering (VQA) model. At inference time, the model extracts concept embeddings along various axes from new test images, which can be remixed to generate images with novel compositions of visual concepts. With a lightweight test-time finetuning procedure, it can also generalize to novel concepts unseen at training.
Project page at https://cs.stanford.edu/~yzzhang/projects/concept-axes.

****
Qian Wang,Zhen Zhang,Zemin Liu,Shengliang Lu,Bingqiao Luo,Bingsheng He
EX-Graph: A Pioneering Dataset Bridging Ethereum and X
While numerous public blockchain datasets are available, their utility is constrained by an exclusive focus on blockchain data. This constraint limits the incorporation of relevant social network data into blockchain analysis, thereby diminishing the breadth and depth of insight that can be derived. To address the above limitation, we introduce EX-Graph, a novel dataset that authentically links Ethereum and X, marking the first and largest dataset of its kind. EX-Graph combines Ethereum transaction records (2 million nodes and 30 million edges) and X following data (1 million nodes and 3 million edges), bonding 30,667 Ethereum addresses with verified X accounts sourced from OpenSea. Detailed statistical analysis on EX- Graph highlights the structural differences between X-matched and non-X-matched Ethereum addresses. Extensive experiments, including Ethereum link prediction, wash-trading Ethereum addresses detection, and X-Ethereum matching link pre- diction, emphasize the significant role of X data in enhancing Ethereum analysis. EX-Graph is available at https://exgraph.deno.dev/.
****
Gabryel Mason-Williams,Fredrik Dahlqvist
What Makes a Good Prune? Maximal Unstructured Pruning for Maximal Cosine Similarity
Pruning is an effective method to reduce the size of deep neural network models, maintain accuracy, and, in some cases, improve the network's overall performance. However, the mechanisms underpinning pruning remain unclear. Why can different methods prune by different percentages yet achieve similar performance? Why can we not prune at the start of training? Why are some models more amenable to being pruned than others? Given a model, what is the maximum amount it can be pruned before significantly affecting the performance? This paper explores and answers these questions from the global unstructured magnitude pruning perspective with one epoch of fine-tuning. We develop the idea that cosine similarity is an effective proxy measure for functional similarity between the parent and the pruned network. We prove that the L1 pruning method is optimal when pruning by cosine similarity. We show that the higher the kurtosis of a model's parameter distribution, the more it can be pruned while maintaining performance. Finally, we present a simple method to determine the optimal amount by which a network can be L1-pruned based on its parameter distribution. The code demonstrating the method is available at https://github.com/gmw99/what_makes_a_good_prune
****
Cong Zhang,Zhiguang Cao,Wen Song,Yaoxin Wu,Jie Zhang
Deep Reinforcement Learning Guided Improvement Heuristic for Job Shop Scheduling
Recent studies in using deep reinforcement learning (DRL) to solve Job-shop scheduling problems (JSSP) focus on construction heuristics. However, their performance is still far from optimality, mainly because the underlying graph representation scheme is unsuitable for modelling partial solutions at each construction step. This paper proposes a novel DRL-guided improvement heuristic for solving JSSP, where graph representation is employed to encode complete solutions. We design a Graph-Neural-Network-based representation scheme, consisting of two modules to effectively capture the information of dynamic topology and different types of nodes in graphs encountered during the improvement process. To speed up solution evaluation during improvement, we present a novel message-passing mechanism that can evaluate multiple solutions simultaneously. We prove that the computational complexity of our method scales linearly with problem size. Experiments on classic benchmarks show that the improvement policy learned by our method outperforms state-of-the-art DRL-based methods by a large margin.
****
André Cruz,Moritz Hardt
Unprocessing Seven Years of Algorithmic Fairness
Seven years ago, researchers proposed a postprocessing method to equalize the error rates of a model across different demographic groups. The work launched hundreds of papers purporting to improve over the postprocessing baseline. We empiri

cally evaluate these claims through thousands of model evaluations on several tabular datasets. We find that the fairness-accuracy Pareto frontier achieved by postprocessing contains all other methods we were feasibly able to evaluate. In doing so, we address two common methodological errors that have confounded previous observations. One relates to the comparison of methods with different unconstrained base models. The other concerns methods achieving different levels of constraint relaxation. At the heart of our study is a simple idea we call unprocessing that roughly corresponds to the inverse of postprocessing. Unprocessing allows for a direct comparison of methods using different underlying models and levels of relaxation.
****
Xingyao Wang,Zihan Wang,Jiateng Liu,Yangyi Chen,Lifan Yuan,Hao Peng,Heng Ji
MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback
To solve complex tasks, large language models (LLMs) often require multiple rounds of interactions with the user, sometimes assisted by external tools.
However, current evaluation protocols often emphasize benchmark performance with single-turn exchanges, neglecting the nuanced interactions among the user, LLMs, and external tools, while also underestimating the importance of natural language feedback from users. These oversights contribute to discrepancies between research benchmark evaluations and real-world use cases.
We introduce MINT, a benchmark that evaluates LLMs' ability to solve tasks with multi-turn interactions by (1) using tools and (2) leveraging natural language feedback.
To ensure reproducibility, we provide an evaluation framework where LLMs can access tools by executing Python code and receive users' natural language feedback simulated by GPT-4.
We repurpose a diverse set of established evaluation datasets focusing on reasoning, coding, and decision-making and carefully curate them into a compact subset for efficient evaluation.
Our analysis of 20 open- and closed-source LLMs offers intriguing findings.
(a) LLMs generally benefit from tools and language feedback, with performance gains (absolute, same below) of 1--8% for each turn of tool use and 2--17% with natural language feedback.
(b) Better single-turn performance does not guarantee better multi-turn performance.
(c) Surprisingly, on the LLMs evaluated, supervised instruction-finetuning (SIFT) and reinforcement learning from human feedback (RLHF) generally hurt multi-turn capabilities.
We expect MINT can help measure progress and incentivize research in improving LLMs' capabilities in multi-turn interactions, especially for open-source communities where multi-turn human evaluation can be less accessible compared to commercial LLMs with a larger user base.
****
Xiyao Wang,Ruijie Zheng,Yanchao Sun,Ruonan Jia,Wichayaporn Wongkamjan,Huazhe Xu,Furong Huang
COPlanner: Plan to Roll Out Conservatively but to Explore Optimistically for Model-Based RL
Dyna-style model-based reinforcement learning contains two phases: model rollouts to generate sample for policy learning and real environment exploration using current policy for dynamics model learning. However, due to the complex real-world environment, it is inevitable to learn an imperfect dynamics model with model prediction error, which can further mislead policy learning and result in sub-optimal solutions. In this paper, we propose $\texttt{COPlanner}$, a planning-driven framework for model-based methods to address the inaccurately learned dynamics model problem with conservative model rollouts and optimistic environment exploration. $\texttt{COPlanner}$ leverages an uncertainty-aware policy-guided model predictive control (UP-MPC) component to plan for multi-step uncertainty estimation. This estimated uncertainty then serves as a penalty during model rollouts and as a bonus during real environment exploration respectively, to choose actions. Consequently, $\texttt{COPlanner}$ can avoid model uncertain regions throu

gh conservative model rollouts, thereby alleviating the influence of model erro
r. Simultaneously, it explores high-reward model uncertain regions to reduce mod
el error actively through optimistic real environment exploration. $\texttt{COPl
anner}$ is a plug-and-play framework that can be applied to any dyna-style model
-based methods. Experimental results on a series of proprioceptive and visual co
ntinuous control tasks demonstrate that both sample efficiency and asymptotic pe
rformance of strong model-based methods are significantly improved combined with
 $\texttt{COPlanner}$.
****
Xuefeng Du,Zhen Fang,Ilias Diakonikolas,Yixuan Li
How Does Unlabeled Data Provably Help Out-of-Distribution Detection?
Using unlabeled data to regularize the machine learning models has demonstrated
promise for improving safety and reliability in detecting out-of-distribution (O
OD) data. Harnessing the power of unlabeled in-the-wild data is non-trivial due
to the heterogeneity of both in-distribution (ID) and OOD data. This lack of a c
lean set of OOD samples poses significant challenges in learning an optimal OOD
classifier. Currently, there is a lack of research on formally understanding how
 unlabeled data helps OOD detection. This paper bridges the gap by introducing a
 new learning framework SAL (Separate And Learn) that offers both strong theoret
ical guarantees and empirical effectiveness. The framework separates candidate o
utliers from the unlabeled data and then trains an OOD classifier using the cand
idate outliers and the labeled ID data. Theoretically, we provide rigorous error
 bounds from the lens of separability and learnability, formally justifying the
two components in our algorithm. Our theory shows that SAL can separate the cand
idate outliers with small error rates, which leads to a generalization guarantee
 for the learned OOD classifier. Empirically, SAL achieves state-of-the-art perf
ormance on common benchmarks, reinforcing our theoretical insights. Code is publ
icly available at https://github.com/deeplearning-wisc/sal.
****
Rui Jiao,Wenbing Huang,Yu Liu,Deli Zhao,Yang Liu
Space Group Constrained Crystal Generation
Crystals are the foundation of numerous scientific and industrial applications.
While various learning-based approaches have been proposed for crystal generatio
n, existing methods neglect the spacegroup constraint which is crucial in descri
bing the geometry of crystals and closely relevant to many desirable properties.
 However, considering spacegroup constraint is challenging owing to its diverse
and nontrivial forms. In this paper, we reduce the spacegroup constraint into an
 equivalent formulation that is more tractable to be handcrafted into the genera
tion process. In particular, we translate the spacegroup constraint into two cas
es: the basis constraint of the invariant exponential space of the lattice matri
x and the Wyckoff position constraint of the fractional coordinates. Upon the de
rived constraints, we then propose DiffCSP++, a novel diffusion model that has e
nhanced a previous work DiffCSP by further taking spacegroup constraint into acc
ount. Experiments on several popular datasets verify the benefit of the involvem
ent of the spacegroup constraint, and show that our DiffCSP++ achieves the best
or comparable performance on crystal structure prediction and ab initio crystal
generation.
****
Zhilu Zhang,Haoyu Wang,Shuai Liu,Xiaotao Wang,LEI LEI,Wangmeng Zuo
Self-Supervised High Dynamic Range Imaging with Multi-Exposure Images in Dynamic
 Scenes
Merging multi-exposure images is a common approach for obtaining high dynamic ra
nge (HDR) images, with the primary challenge being the avoidance of ghosting art
ifacts in dynamic scenes. Recent methods have proposed using deep neural network
s for deghosting. However, the methods typically rely on sufficient data with HD
R ground-truths, which are difficult and costly to collect. In this work, to eli
minate the need for labeled data, we propose SelfHDR, a self-supervised HDR reco
nstruction method that only requires dynamic multi-exposure images during traini
ng. Specifically, SelfHDR learns a reconstruction network under the supervision
of two complementary components, which can be constructed from multi-exposure im

ages and focus on HDR color as well as structure, respectively. The color component is estimated from aligned multi-exposure images, while the structure one is generated through a structure-focused network that is supervised by the color component and an input reference (\eg, medium-exposure) image. During testing, the learned reconstruction network is directly deployed to predict an HDR image. Experiments on real-world images demonstrate our SelfHDR achieves superior results against the state-of-the-art self-supervised methods, and comparable performance to supervised ones. Codes are available at https://github.com/cszhilu1998/SelfHDR
****
Rachit Bansal,Bidisha Samanta,Siddharth Dalmia,Nitish Gupta,Sriram Ganapathy,Abhishek Bapna,Prateek Jain,Partha Talukdar
LLM Augmented LLMs: Expanding Capabilities through Composition
Foundational models with billions of parameters which have been trained on large corpus of data have demonstrated non-trivial skills in a variety of domains. However, due to their monolithic structure, it is challenging and expensive to augment them or impart new skills. On the other hand, due to their adaptation abilities,several new instances of these models are being trained towards new domains and tasks. In this work, we study the problem of efficient and practical composition of existing foundation models with more specific models to enable newer capabilities. To this end, we propose CALM—Composition to Augment Language Models—which introduces cross-attention between models to compose their representations and enable new capabilities. Salient features of CALM are: (i) Scales up LLMs on new tasks by 're-using' existing LLMs along with a few additional parameters and data, (ii) Existing model weights are kept intact, and hence preserves existing capabilities, and (iii) Applies to diverse domains and settings. We illustrate that augmenting PaLM2-S with a smaller model trained on low-resource languages results in an absolute improvement of up to 13% on tasks like translation into English and arithmetic reasoning for low-resource languages. Similarly,when PaLM2-S is augmented with a code-specific model, we see a relative improvement of 40% over the base model for code generation and explanation tasks—on-par with fully fine-tuned counterparts.
****
Xinmeng Huang,Ping Li,Xiaoyun Li
Stochastic Controlled Averaging for Federated Learning with Communication Compression
Communication compression has been an important topic in Federated Learning (FL) for alleviating the communication overhead. However, communication compression brings forth new challenges in FL due to the interplay of compression-incurred information distortion and inherent characteristics of FL such as partial participation and data heterogeneity. Despite the recent development, the existing approaches either cannot accommodate arbitrary data heterogeneity or partial participation, or require stringent conditions on compression. In this paper, we revisit the seminal stochastic controlled averaging method by proposing an equivalent but more efficient/simplified formulation with halved uplink communication costs, building upon which we propose two compressed FL algorithms, SCALLION and SCAFCOM, to support unbiased and biased compression, respectively. Both the proposed methods outperform the existing compressed FL methods in terms of communication and computation complexities. Moreover,SCALLION and SCAFCOM attain fast convergence rates under arbitrary data heterogeneity without any additional assumptions on compression errors. Experiments show that \scallion and \scafcom outperform recent compressed FL methods under the same communication budget.
****
Fanqi Wan,Xinting Huang,Deng Cai,Xiaojun Quan,Wei Bi,Shuming Shi
Knowledge Fusion of Large Language Models
While training large language models (LLMs) from scratch can generate models with distinct functionalities and strengths, it comes at significant costs and may result in redundant capabilities. Alternatively, a cost-effective and compelling approach is to merge existing pre-trained LLMs into a more potent model. However, due to the varying architectures of these LLMs, directly blending their weigh

ts is impractical. In this paper, we introduce the notion of knowledge fusion for LLMs, aimed at combining the capabilities of existing LLMs and transferring them into a single LLM. By leveraging the generative distributions of source LLMs, we externalize their collective knowledge and unique strengths, thereby potentially elevating the capabilities of the target model beyond those of any individual source LLM. We validate our approach using three popular LLMs with different architectures—Llama-2, MPT, and OpenLLaMA—across various benchmarks and tasks. Our findings confirm that the fusion of LLMs can improve the performance of the target model across a range of capabilities such as reasoning, commonsense, and code generation. Our code, model weights, and data are public at \url{https://github.com/fanqiwan/FuseLLM}.
****
Apratim Bhattacharyya,Sunny Panchal,Reza Pourreza,Mingu Lee,Pulkit Madan,Roland Memisevic
Look, Remember and Reason: Grounded Reasoning in Videos with Language Models
Multi-modal language models (LM) have recently shown promising performance
in high-level reasoning tasks on videos. However, existing methods still fall short in tasks like causal or compositional spatiotemporal reasoning over actions, in which model predictions need to be grounded in fine-grained low-level details, such as object motions and object interactions.
In this work, we propose training an LM end-to-end on low-level surrogate tasks, including object detection, re-identification, and tracking, to endow the model with the required low-level visual capabilities. We show that a two-stream video encoder with spatiotemporal attention is effective at capturing the required static and motion-based cues in the video. By leveraging the LM's ability to perform the low-level surrogate tasks, we can cast reasoning in videos as the three-step process of *Look, Remember, Reason*, wherein visual information is extracted using low-level visual skills step-by-step and then integrated to arrive at a final answer. We demonstrate the effectiveness of our framework on diverse visual reasoning tasks from the ACRE, CATER, Something-Else and STAR datasets. Our approach is trainable end-to-end and surpasses state-of-the-art task-specific methods across these tasks by a large margin.
****
Zayne Rea Sprague,Xi Ye,Kaj Bostrom,Swarat Chaudhuri,Greg Durrett
MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning
While large language models (LLMs) equipped with techniques like chain-of-thought prompting have demonstrated impressive capabilities, they still fall short in their ability to reason robustly in complex settings. However, evaluating LLM reasoning is challenging because system capabilities continue to grow while benchmark datasets for tasks like logical deduction have remained static. We introduce MuSR, a dataset for evaluating language models on multistep soft reasoning tasks specified in a natural language narrative. This dataset has two crucial features. First, it is created through a novel neurosymbolic synthetic-to-natural generation algorithm, enabling the construction of complex reasoning instances that challenge GPT-4 (e.g., murder mysteries roughly 1000 words in length) and which can be scaled further as more capable LLMs are released. Second, our data instances are free text narratives corresponding to real-world domains of reasoning; this makes it simultaneously much more challenging than other synthetically-crafted benchmarks while remaining realistic and tractable for human annotators to solve with high accuracy. We evaluate a range of LLMs and prompting techniques on this dataset and characterize the gaps that remain for techniques like chain-of-thought to perform robust reasoning.
****
Duo Cheng,Xingyu Zhou,Bo Ji
Follow-the-Perturbed-Leader for Adversarial Bandits: Heavy Tails, Robustness, and Privacy
We study adversarial bandit problems with potentially heavy-tailed losses. Unlike standard settings with non-negative and bounded losses, managing negative and unbounded losses introduces a unique challenge in controlling the ``stability'' of the algorithm and hence the regret. To tackle this challenge, we propose a Fo

llow-the-Perturbed-Leader (FTPL) based learning algorithm. Notably, our method achieves (nearly) optimal worst-case regret, eliminating the need for an undesired assumption inherent in the Follow-the-Regularized-Leader (FTRL) based approach. Thanks to this distinctive advantage, our algorithmic framework finds novel applications in two important scenarios with unbounded heavy-tailed losses. For adversarial bandits with heavy-tailed losses and Huber contamination, which we call the robust setting, our algorithm is the first to match the lower bound (up to a $\polylog(K)$ factor, where K is the number of actions). In the private setting, where true losses are in a bounded range (e.g., $[0,1]$) but with additional Local Differential Privacy (LDP) guarantees, our algorithm achieves an improvement of a $\polylog(T)$ factor in the regret bound compared to the best-known results, where T is the total number of rounds. Furthermore, when compared to state-of-the-art FTRL-based algorithms, our FTPL-based algorithm has a more streamlined design. It eliminates the need for additional explicit exploration and solely maintains the absolute value of loss estimates below a predetermined threshold.
****
Peiyan Zhang,Haoyang Liu,Chaozhuo Li,Xing Xie,Sunghun Kim,Haohan Wang
Foundation Model-oriented Robustness: Robust Image Model Evaluation with Pretrained Models
Machine learning has demonstrated remarkable performance over finite datasets, yet whether the scores over the fixed benchmarks can sufficiently indicate the model’s performance in the real world is still in discussion. In reality, an ideal robust model will probably behave similarly to the oracle (e.g., the human users), thus a good evaluation protocol is probably to evaluate the models’ behaviors in comparison to the oracle. In this paper, we introduce a new robustness measurement that directly measures the image classification model’s performance compared with a surrogate oracle (i.e., a zoo of foundation models). Besides, we design a simple method that can accomplish the evaluation beyond the scope of the benchmarks. Our method extends the image datasets with new samples that are sufficiently perturbed to be distinct from the ones in the original sets, but are still bounded within the same image-label structure the original test image represents, constrained by a zoo of foundation models pretrained with a large amount of samples. As a result, our new method will offer us a new way to evaluate the models’ robustness performance, free of limitations of fixed benchmarks or constrained perturbations, although scoped by the power of the oracle. In addition to the evaluation results, we also leverage our generated data to understand the behaviors of the model and our new evaluation strategies.
****
Urszula Julia Komorowska,Simon V Mathis,Kieran Didi,Francisco Vargas,Pietro Lio,Mateja Jamnik
Dynamics-Informed Protein Design with Structure Conditioning
Current protein generative models are able to design novel backbones with desired shapes or functional motifs. However, despite the importance of a protein’s dynamical properties for its function, conditioning on dynamical properties remains elusive. We present a new approach to protein generative modeling by leveraging Normal Mode Analysis that enables us to capture dynamical properties too. We introduce a method for conditioning the diffusion probabilistic models on protein dynamics, specifically on the lowest non-trivial normal mode of oscillation. Our method, similar to the classifier guidance conditioning, formulates the sampling process as being driven by conditional and unconditional terms. However, unlike previous works, we approximate the conditional term with a simple analytical function rather than an external neural network, thus making the eigenvector calculations approachable. We present the corresponding SDE theory as a formal justification of our approach. We extend our framework to conditioning on structure and dynamics at the same time, enabling scaffolding of the dynamical motifs. We demonstrate the empirical effectiveness of our method by turning the open-source unconditional protein diffusion model Genie into the conditional model with no retraining. Generated proteins exhibit the desired dynamical and structural properties while still being biologically plausible. Our work represents a first ste

p towards incorporating dynamical behaviour in protein design and may open the door to designing more flexible and functional proteins in the future.
****
florence regol,Joud Chataoui,Mark Coates
Jointly-Learned Exit and Inference for a Dynamic Neural Network
Large pretrained models, coupled with fine-tuning, are slowly becoming established as the dominant architecture in machine learning. Even though these models offer impressive performance, their practical application is often limited by the prohibitive amount of resources required for $\textit{every}$ inference. Early-exiting dynamic neural networks (EDNN) circumvent this issue by allowing a model to make some of its predictions from intermediate layers (i.e., early-exit). Training an EDNN architecture is challenging as it consists of two intertwined components: the gating mechanism (GM) that controls early-exiting decisions and the intermediate inference modules (IMs) that perform inference from intermediate representations. As a result, most existing approaches rely on thresholding confidence metrics for the gating mechanism and strive to improve the underlying backbone network and the inference modules. Although successful, this approach has two fundamental shortcomings: 1) the GMs and the IMs are decoupled during training, leading to a train-test mismatch; and 2) the thresholding gating mechanism introduces a positive bias into the predictive probabilities, making it difficult to readily extract uncertainty information. We propose a novel architecture that connects these two modules. This leads to significant performance improvements on classification datasets and enables better uncertainty characterization capabilities.
****
Mikhail Galkin,Xinyu Yuan,Hesham Mostafa,Jian Tang,Zhaocheng Zhu
Towards Foundation Models for Knowledge Graph Reasoning
Foundation models in language and vision have the ability to run inference on any textual and visual inputs thanks to the transferable representations such as a vocabulary of tokens in language.
Knowledge graphs (KGs) have different entity and relation vocabularies that generally do not overlap.
The key challenge of designing foundation models on KGs is to learn such transferable representations that enable inference on any graph with arbitrary entity and relation vocabularies.
In this work, we make a step towards such foundation models and present ULTRA, an approach for learning universal and transferable graph representations.
ULTRA builds relational representations as a function conditioned on their interactions.
Such a conditioning strategy allows a pre-trained ULTRA model to inductively generalize to any unseen KG with any relation vocabulary and to be fine-tuned on any graph.
Conducting link prediction experiments on 57 different KGs, we find that the zero-shot inductive inference performance of a single pre-trained ULTRA model on unseen graphs of various sizes is often on par or better than strong baselines trained on specific graphs.
Fine-tuning further boosts the performance.
****
Bowen Shi,XIAOPENG ZHANG,Yaoming Wang,Jin Li,Wenrui Dai,Junni Zou,Hongkai Xiong,Qi Tian
Hybrid Distillation: Connecting Masked Autoencoders with Contrastive Learners
As two prominent strategies for representation learning, Contrastive Learning (CL) and Masked Image Modeling (MIM) have witnessed significant progress. Previous studies have demonstrated the advantages of each approach in specific scenarios. CL, resembling supervised pre-training, excels at capturing longer-range global patterns and enhancing feature discrimination, while MIM is adept at introducing local and diverse attention across transformer layers. Considering the respective strengths, previous studies utilize feature distillation to inherit both discrimination and diversity. In this paper, we thoroughly examine previous feature distillation methods and observe that the increase in diversity mainly stems f

rom asymmetric designs, which may in turn compromise the discrimination ability. To strike a balance between the two properties, we propose a simple yet effective strategy termed Hybrid Distill, which leverages both the CL and MIM teachers to jointly guide the student model. Hybrid Distill emulates the token relations of the MIM teacher at intermediate layers for diversity, while simultaneously distilling the final features of the CL teacher to enhance discrimination. A progressive redundant token masking strategy is employed to reduce the expenses associated with distillation and aid in preventing the model from converging to local optima. Experimental results demonstrate that Hybrid Distill achieves superior performance on various benchmark datasets.
****
Quinn LeBlanc Fisher,Haoming Meng,Vardan Papyan
Pushing Boundaries: Mixup's Influence on Neural Collapse
Mixup is a data augmentation strategy that employs convex combinations of training instances and their respective labels to improve the robustness and calibration of deep neural networks. Despite its widespread adoption, the nuanced mechanisms that underpin its success are not entirely understood. The observed phenomenon of Neural Collapse, where the last-layer activations and classifier of deep networks converge to a simplex equiangular tight frame (ETF), provides a compelling motivation to explore whether mixup induces alternative geometric configurations and whether those could explain its success. In this study, we delve into the last-layer activations of training data for deep networks subjected to mixup, aiming to uncover insights into its operational efficacy. Our investigation, spanning various architectures and dataset pairs, reveals that mixup's last-layer activations predominantly converge to a distinctive configuration different than one might expect. In this configuration, activations from mixed-up examples of identical classes align with the classifier, while those from different classes delineate channels along the decision boundary. These findings are unexpected, as mixed-up features are not simple convex combinations of feature class means (as one might get, for example, by training mixup with the mean squared error loss). By analyzing this distinctive geometric configuration, we elucidate the mechanisms by which mixup enhances model calibration. To further validate our empirical observations, we conduct a theoretical analysis under the assumption of an unconstrained features model, utilizing the mixup loss. Through this, we characterize and derive the optimal last-layer features under the assumption that the classifier forms a simplex ETF.
****
Shengbo Wang,Jose Blanchet,Peter Glynn
Optimal Sample Complexity for Average Reward Markov Decision Processes
We resolve the open question regarding the sample complexity of policy learning for maximizing the long-run average reward associated with a uniformly ergodic Markov decision process (MDP), assuming a generative model. In this context, the existing literature provides a sample complexity upper bound of $\widetilde O(|S||A|t_{\text{mix}}^2 \epsilon^{-2})$ and a lower bound of $\Omega(|S||A|t_{\text{mix}} \epsilon^{-2})$. In these expressions, $|S|$ and $|A|$ denote the cardinalities of the state and action spaces respectively, t_{mix} serves as a uniform upper limit for the total variation mixing times, and ϵ signifies the error tolerance. Therefore, a notable gap of t_{mix} still remains to be bridged. Our primary contribution is the development of an estimator for the optimal policy of average reward MDPs with a sample complexity of $\widetilde O(|S||A|t_{\text{mix}}\epsilon^{-2})$. This marks the first algorithm and analysis to reach the literature's lower bound. Our new algorithm draws inspiration from ideas in Li et al. (2020), Jin \& Sidford (2021), and Wang et al. (2023). Additionally, we conduct numerical experiments to validate our theoretical findings.
****
Jaemoo Choi,Jaewoong Choi,Myungjoo Kang
Analyzing and Improving Optimal-Transport-based Adversarial Networks
Optimal Transport (OT) problem aims to find a transport plan that bridges two distributions while minimizing a given cost function. OT theory has been widely ut

ilized in generative modeling. In the beginning, OT distance has been used as a measure for assessing the distance between data and generated distributions. Rec ently, OT transport map between data and prior distributions has been utilized a s a generative model. These OT-based generative models share a similar adversari al training objective. In this paper, we begin by unifying these OT-based advers arial methods within a single framework. Then, we elucidate the role of each com ponent in training dynamics through a comprehensive analysis of this unified fra mework. Moreover, we suggest a simple but novel method that improves the previou sly best-performing OT-based model. Intuitively, our approach conducts a gradual refinement of the generated distribution, progressively aligning it with the da ta distribution. Our approach achieves a FID score of 2.51 on CIFAR-10 and 5.99 on CelebA-HQ-256, outperforming unified OT-based adversarial approaches.
****
Marius Memmel,Andrew Wagenmaker,Chuning Zhu,Dieter Fox,Abhishek Gupta
ASID: Active Exploration for System Identification in Robotic Manipulation
Model-free control strategies such as reinforcement learning have shown the abil ity to learn control strategies without requiring an accurate model or simulator of the world. While this is appealing due to the lack of modeling requirements, such methods can be sample inefficient, making them impractical in many real-wo rld domains. On the other hand, model-based control techniques leveraging accura te simulators can circumvent these challenges and use a large amount of cheap si mulation data to learn controllers that can effectively transfer to the real wor ld. The challenge with such model-based techniques is the requirement for an ext remely accurate simulation, requiring both the specification of appropriate simu lation assets and physical parameters. This requires considerable human effort t o design for every environment being considered. In this work, we propose a lear ning system that can leverage a small amount of real-world data to autonomously refine a simulation model and then plan an accurate control strategy that can be deployed in the real world. Our approach critically relies on utilizing an init ial (possibly inaccurate) simulator to design effective exploration policies tha t, when deployed in the real world, collect high-quality data. We demonstrate th e efficacy of this paradigm in identifying articulation, mass, and other physica l parameters in several challenging robotic manipulation tasks, and illustrate t hat only a small amount of real-world data can allow for effective sim-to-real t ransfer.
****
Shenyu Lu,Yipei Wang,Xiaoqian Wang
Debiasing Attention Mechanism in Transformer without Demographics
Although transformers demonstrate impressive capabilities in a variety of tasks, the fairness issue remains a significant concern when deploying these models. E xisting works to address fairness issues in transformers require sensitive label s (such as age, gender, etc.), which can raise privacy concerns or violate legal regulations. An alternative way is through fairness without demographics. Howev er, existing works that improve Rawlsian Max-Min fairness may impose overly rest rictive constraints. Other methods that use auxiliary networks could be paramete r inefficient. In this paper, we present a new approach to debiasing transformer s by leveraging their inherent structure. By reconsidering the roles of importa nt components (queries, keys, and values) in the attention mechanism, we introdu ce a simple yet effective debiasing strategy from two perspectives: 1) Grounded in theoretical analysis, we normalize and apply absolute value operations to que ries and keys to minimize the bias in attention weight allocation; 2) We reduce the bias within values through local alignment via contrastive learning. Through out the entire process, our approach does not require any sensitive labels. Furt hermore, to enhance memory efficiency in the training phase, we propose a strate gy that debias only the last encoder to improve fairness in pre-trained models. We conduct experiments in computer vision and natural language processing tasks and show that our method is comparable and even outperforms the state-of-the-art method with substantially lower energy consumption.
****
Yufei Kuang,Jie Wang,Haoyang Liu,Fangzhou Zhu,Xijun Li,Jia Zeng,Jianye HAO,Bin L

i,Feng Wu
Rethinking Branching on Exact Combinatorial Optimization Solver: The First Deep Symbolic Discovery Framework
Machine learning (ML) has been shown to successfully accelerate solving NP-hard combinatorial optimization (CO) problems under the branch and bound framework. However, the high training and inference cost and limited interpretability of ML approaches severely limit their wide application to modern exact CO solvers. In contrast, human-designed policies---though widely integrated in modern CO solvers due to their compactness and reliability---can not capture data-driven patterns for higher performance. To combine the advantages of the two paradigms, we propose the first symbolic discovery framework---namely, deep symbolic discovery for exact combinatorial optimization solver (Symb4CO)---to learn high-performance symbolic policies on the branching task. Specifically, we show the potential existence of small symbolic policies empirically, employ a large neural network to search in the high-dimensional discrete space, and compile the learned symbolic policies directly for fast deployment. Experiments show that the Symb4CO learned purely CPU-based policies consistently achieve *comparable* performance to previous GPU-based state-of-the-art approaches.
Furthermore, the appealing features of Symb4CO include its high training (*ten training instances*) and inference (*one CPU core*) efficiency and good interpretability (*one-line expressions*), making it simple and reliable for deployment. The results show encouraging potential for the *wide* deployment of ML to modern CO solvers.
****
Ruoyu Chen,Hua Zhang,Siyuan Liang,Jingzhi Li,Xiaochun Cao
Less is More: Fewer Interpretable Region via Submodular Subset Selection
Image attribution algorithms aim to identify important regions that are highly relevant to model decisions. Although existing attribution solutions can effectively assign importance to target elements, they still face the following challenges: 1) existing attribution methods generate inaccurate small regions thus misleading the direction of correct attribution, and 2) the model cannot produce good attribution results for samples with wrong predictions. To address the above challenges, this paper re-models the above image attribution problem as a submodular subset selection problem, aiming to enhance model interpretability using fewer regions. To address the lack of attention to local regions, we construct a novel submodular function to discover more accurate small interpretation regions. To enhance the attribution effect for all samples, we also impose four different constraints on the selection of sub-regions, i.e., confidence, effectiveness, consistency, and collaboration scores, to assess the importance of various subsets. Moreover, our theoretical analysis substantiates that the proposed function is in fact submodular. Extensive experiments show that the proposed method outperforms SOTA methods on two face datasets (Celeb-A and VGG-Face2) and one fine-grained dataset (CUB-200-2011). For correctly predicted samples, the proposed method improves the Deletion and Insertion scores with an average of 4.9\% and 2.5\% gain relative to HSIC-Attribution. For incorrectly predicted samples, our method achieves gains of 81.0\% and 18.4\% compared to the HSIC-Attribution algorithm in the average highest confidence and Insertion score respectively. The code is released at https://github.com/RuoyuChen10/SMDL-Attribution.
****
Keiran Paster,Marco Dos Santos,Zhangir Azerbayev,Jimmy Ba
OpenWebMath: An Open Dataset of High-Quality Mathematical Web Text
There is growing evidence that pretraining on high quality, carefully thought-out tokens such as code or mathematics plays an important role in improving the reasoning abilities of large language models. For example, Minerva, a PaLM model finetuned on billions of tokens of mathematical documents from arXiv and the web, reported dramatically improved performance on problems that require quantitative reasoning. However, because all known open source web datasets employ preprocessing that does not faithfully preserve mathematical notation, the benefits of large scale training on quantitive web documents are unavailable to the research community. We introduce OpenWebMath, an open dataset inspired by these works con

taining 14.7B tokens of mathematical webpages from Common Crawl. We describe in detail our method for extracting text and LaTeX content and removing boilerplate from HTML documents, as well as our methods for quality filtering and deduplication. Additionally, we run small-scale experiments by training 1.4B language models on OpenWebMath, showing that models trained on 14.7B tokens of our dataset surpass the performance of models trained on over 20x the amount of general language data. We hope that our dataset, open-sourced and released on the Hugging Face Hub, will help spur advances in the reasoning abilities of large language models.
****
Karan Mirakhor,Sourav Ghosh,Dipanjan Das,Brojeshwar Bhowmick
Task Planning for Visual Room Rearrangement under Partial Observability
This paper presents a novel hierarchical task planner under partial observability that empowers an embodied agent to use visual input to efficiently plan a sequence of actions for simultaneous object search and rearrangement in an untidy room, to achieve a desired tidy state. The paper introduces (i) a novel Search Network that utilizes commonsense knowledge from large language models to find unseen objects, (ii) a Deep RL network trained with proxy reward, along with (iii) a novel graph-based state representation to produce a scalable and effective planner that interleaves object search and rearrangement to minimize the number of steps taken and overall traversal of the agent, as well as to resolve blocked goal and swap cases, and (iv) a sample-efficient cluster-biased sampling for simultaneous training of the proxy reward network along with the Deep RL network. Furthermore, the paper presents new metrics and a benchmark dataset - RoPOR, to measure the effectiveness of rearrangement planning. Experimental results show that our method significantly outperforms the state-of-the-art rearrangement methods Weihs et al. (2021a); Gadre et al. (2022); Sarch et al. (2022); Ghosh et al. (2022).
****
Zifeng Wang,Zichen Wang,Balasubramaniam Srinivasan,Vassilis N. Ioannidis,Huzefa Rangwala,RISHITA ANUBHAI
BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs
Foundation models (FMs) learn from large volumes of unlabeled data to demonstrate superior performance across a wide range of tasks. However, FMs developed for biomedical domains have largely remained unimodal, i.e., independently trained and used for tasks on protein sequences alone, small molecule structures alone, or clinical data alone.
To overcome this limitation, we present BioBridge, a parameter-efficient learning framework, to bridge independently trained unimodal FMs to establish multimodal behavior. BioBridge achieves it by utilizing Knowledge Graphs (KG) to learn transformations between one unimodal FM and another without fine-tuning any underlying unimodal FMs.
Our results demonstrate that BioBridge can beat the best baseline KG embedding methods (on average by ~ 76.3%) in cross-modal retrieval tasks. We also identify BioBridge demonstrates out-of-domain generalization ability by extrapolating to unseen modalities or relations. Additionally, we also show that BioBridge presents itself as a general-purpose retriever that can aid biomedical multimodal question answering as well as enhance the guided generation of novel drugs. Code is at https://github.com/RyanWangZf/BioBridge.
****
Ruiquan Huang,Yingbin Liang,Jing Yang
Provably Efficient UCB-type Algorithms For Learning Predictive State Representations
The general sequential decision-making problem, which includes Markov decision p

rocesses (MDPs) and partially observable MDPs (POMDPs) as special cases, aims at maximizing a cumulative reward by making a sequence of decisions based on a history of observations and actions over time. Recent studies have shown that the sequential decision-making problem is statistically learnable if it admits a low-rank structure modeled by predictive state representations (PSRs). Despite these advancements, existing approaches typically involve oracles or steps that are computationally intractable. On the other hand, the upper confidence bound (UCB) based approaches, which have served successfully as computationally efficient methods in bandits and MDPs, have not been investigated for more general PSRs, due to the difficulty of optimistic bonus design in these more challenging settings. This paper proposes the first known UCB-type approach for PSRs, featuring a novel bonus term that upper bounds the total variation distance between the estimated and true models. We further characterize the sample complexity bounds for our designed UCB-type algorithms for both online and offline PSRs. In contrast to existing approaches for PSRs, our UCB-type algorithms enjoy computational tractability, last-iterate guaranteed near-optimal policy, and guaranteed model accuracy.
****
Xin Liu,Muhammad Khalifa,Lu Wang
LitCab: Lightweight Language Model Calibration over Short- and Long-form Responses
A model is considered well-calibrated when its probability estimate aligns with the actual likelihood of the output being correct. Calibrating language models (LMs) is crucial, as it plays a vital role in detecting and mitigating hallucinations of LMs as well as building more trustworthy models. However, standard calibration techniques may not be suited for LM calibration. For instance, post-processing methods such as temperature scaling do not reorder the candidate generations. On the other hand, training-based methods require fine-tuning the entire model, which is impractical for LMs of large scale. We present LitCab, a lightweight calibration mechanism consisting of a single linear layer that takes the input text representation and predicts a bias term, which is then added to the LM output logits. LitCab improves model calibration by only adding < 2% of the original model parameters. For evaluation, we construct CaT, a benchmark consisting of eight text generation tasks, covering responses ranging from short phrases to paragraphs. We test LitCab with Llama2-7B, where it improves calibration across all tasks, reducing the average ECE score by as large as 30%. We further conduct a comprehensive evaluation with multiple popular open-sourced LMs from GPT and LLaMA families, yielding the following key findings: (i) Larger models within the same family exhibit better calibration on tasks with short generation tasks, but not necessarily for longer ones. (ii) GPT-family models show superior calibration compared to LLaMA, Llama2, and Vicuna models, despite having much fewer parameters. (iii) Fine-tuning pretrained model (e.g., LLaMA) with samples of limited purpose (e.g., conversations) may lead to worse calibration, highlighting the importance of fine-tuning setups for calibrating LMs.
****
Yue Cao,Tianlin Li,Xiaofeng Cao,Ivor Tsang,Yang Liu,Qing Guo
IRAD: Implicit Representation-driven Image Resampling against Adversarial Attacks
We introduce a novel approach to counter adversarial attacks, namely, image resampling. Image resampling transforms a discrete image into a new one, simulating the process of scene recapturing or rerendering as specified by a geometrical transformation. The underlying rationale behind our idea is that image resampling can alleviate the influence of adversarial perturbations while preserving essential semantic information, thereby conferring an inherent advantage in defending against adversarial attacks. To validate this concept, we present a comprehensive study on leveraging image resampling to defend against adversarial attacks. We have developed basic resampling methods that employ interpolation strategies and coordinate shifting magnitudes. Our analysis reveals that these basic methods can partially mitigate adversarial attacks. However, they come with apparent limitations: the accuracy of clean images noticeably decreases, while the improveme

nt in accuracy on adversarial examples is not substantial.We propose implicit representation-driven image resampling (IRAD) to overcome these limitations. First, we construct an implicit continuous representation that enables us to represent any input image within a continuous coordinate space. Second, we introduce SampleNet, which automatically generates pixel-wise shifts for resampling in response to different inputs. Furthermore, we can extend our approach to the state-of-the-art diffusion-based method, accelerating it with fewer time steps while preserving its defense capability. Extensive experiments demonstrate that our method significantly enhances the adversarial robustness of diverse deep models against various attacks while maintaining high accuracy on clean images.
****

Qiuyi Chen,Mark Fuge
Compressing Latent Space via Least Volume
This paper introduces Least Volume---a simple yet effective regularization inspired by geometric intuition---that can reduce the necessary number of latent dimensions needed by an autoencoder without requiring any prior knowledge of the intrinsic dimensionality of the dataset. We show that the Lipschitz continuity of the decoder is the key to making it work, provide a proof that PCA is just a linear special case of it, and reveal that it has a similar PCA-like importance ordering effect when applied to nonlinear models. We demonstrate the intuition behind the regularization on some pedagogical toy problems, and its effectiveness on several benchmark problems, including MNIST, CIFAR-10 and CelebA.
****

Wes Gurnee,Max Tegmark
Language Models Represent Space and Time
The capabilities of large language models (LLMs) have sparked debate over whether such systems just learn an enormous collection of superficial statistics or a set of more coherent and grounded representations that reflect the real world. We find evidence for the latter by analyzing the learned representations of three spatial datasets (world, US, NYC places) and three temporal datasets (historical figures, artworks, news headlines) in the Llama-2 family of models. We discover that LLMs learn linear representations of space and time across multiple scales. These representations are robust to prompting variations and unified across different entity types (e.g. cities and landmarks). In addition, we identify individual "space neurons" and "time neurons" that reliably encode spatial and temporal coordinates. While further investigation is needed, our results suggest modern LLMs learn rich spatiotemporal representations of the real world and possess basic ingredients of a world model.
****

Bowen Song,Soo Min Kwon,Zecheng Zhang,Xinyu Hu,Qing Qu,Liyue Shen
Solving Inverse Problems with Latent Diffusion Models via Hard Data Consistency
Latent diffusion models have been demonstrated to generate high-quality images, while offering efficiency in model training compared to diffusion models operating in the pixel space. However, incorporating latent diffusion models to solve inverse problems remains a challenging problem due to the nonlinearity of the encoder and decoder. To address these issues, we propose ReSample, an algorithm that can solve general inverse problems with pre-trained latent diffusion models. Our algorithm incorporates data consistency by solving an optimization problem during the reverse sampling process, a concept that we term as hard data consistency. Upon solving this optimization problem, we propose a novel resampling scheme to map the measurement-consistent sample back onto the noisy data manifold and theoretically demonstrate its benefits. Lastly, we apply our algorithm to solve a wide range of linear and nonlinear inverse problems in both natural and medical images, demonstrating that our approach outperforms existing state-of-the-art approaches, including those based on pixel-space diffusion models.
****

Yinan Zheng,Jianxiong Li,Dongjie Yu,Yujie Yang,Shengbo Eben Li,Xianyuan Zhan,Jingjing Liu
Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model
Safe offline reinforcement learning is a promising way to bypass risky online in

teractions towards safe policy learning. Most existing methods only enforce soft
 constraints, i.e., constraining safety violations in expectation below threshol
ds predetermined. This can lead to potentially unsafe outcomes, thus unacceptabl
e in safety-critical scenarios. An alternative is to enforce the hard constraint
 of zero violation. However, this can be challenging in offline setting, as it n
eeds to strike the right balance among three highly intricate and correlated asp
ects: safety constraint satisfaction, reward maximization, and behavior regulari
zation imposed by offline datasets. Interestingly, we discover that via reachabi
lity analysis of safe-control theory, the hard safety constraint can be equivale
ntly translated to identifying the largest feasible region given the offline dat
aset. This seamlessly converts the original trilogy problem to a feasibility-dep
endent objective, i.e., maximizing reward value within the feasible region while
 minimizing safety risks in the infeasible region. Inspired by these, we propose
 FISOR (FeasIbility-guided Safe Offline RL), which allows safety constraint adhe
rence, reward maximization, and offline policy learning to be realized via three
 decoupled processes, while offering strong safety performance and stability. In
 FISOR, the optimal policy for the translated optimization problem can be derive
d in a special form of weighted behavior cloning, which can be effectively extra
cted with a guided diffusion model thanks to its expressiveness. We compare FIS
OR against baselines on DSRL benchmark for safe offline RL. Evaluation results s
how that FISOR is the only method that can guarantee safety satisfaction in all
tasks, while achieving top returns in most tasks. Code: https://github.com/Zheng
Yinan-AIR/FISOR.
****
António Farinhas,Chrysoula Zerva,Dennis Thomas Ulmer,Andre Martins
Non-Exchangeable Conformal Risk Control
Split conformal prediction has recently sparked great interest due to its abilit
y to provide formally guaranteed uncertainty sets or intervals for predictions m
ade by black-box neural models, ensuring a predefined probability of containing
the actual ground truth. While the original formulation assumes data exchangeabi
lity, some extensions handle non-exchangeable data, which is often the case in m
any real-world scenarios. In parallel, some progress has been made in conformal
methods that provide statistical guarantees for a broader range of objectives, s
uch as bounding the best F_1-score or minimizing the false negative rate in ex
pectation. In this paper, we leverage and extend these two lines of work by prop
osing non-exchangeable conformal risk control, which allows controlling the expe
cted value of any monotone loss function when the data is not exchangeable. Our
framework is flexible, makes very few assumptions, and allows weighting the data
 based on its relevance for a given test example; a careful choice of weights ma
y result in tighter bounds, making our framework useful in the presence of chang
e points, time series, or other forms of distribution drift. Experiments with bo
th synthetic and real world data show the usefulness of our method.
****
Yisong Huang,Jin Li,Xinlong Chen,Yang-Geng Fu
Training Graph Transformers via Curriculum-Enhanced Attention Distillation
Recent studies have shown that Graph Transformers (GTs) can be effective for spe
cific graph-level tasks. However, when it comes to node classification, training
 GTs remains challenging, especially in semi-supervised settings with a severe s
carcity of labeled data. Our paper aims to address this research gap by focusing
 on semi-supervised node classification. To accomplish this, we develop a curric
ulum-enhanced attention distillation method that involves utilizing a Local GT t
eacher and a Global GT student. Additionally, we introduce the concepts of in-cl
ass and out-of-class and then propose two improvements, out-of-class entropy and
 top-k pruning, to facilitate the student's out-of-class exploration under the t
eacher's in-class guidance. Taking inspiration from human learning, our method i
nvolves a curriculum mechanism for distillation that initially provides strict g
uidance to the student and gradually allows for more out-of-class exploration by
 a dynamic balance. Extensive experiments show that our method outperforms many
state-of-the-art approaches on seven public graph benchmarks, proving its effect
iveness.

****
Elisa Kreiss,Eric Zelikman,Christopher Potts,Nick Haber
ContextRef: Evaluating Referenceless Metrics for Image Description Generation
Referenceless metrics (e.g., CLIPScore) use pretrained vision--language models to assess image descriptions directly without costly ground-truth reference texts. Such methods can facilitate rapid progress, but only if they truly align with human preference judgments. In this paper, we introduce ContextRef, a benchmark for assessing referenceless metrics for such alignment. ContextRef has two components: human ratings along a variety of established quality dimensions, and ten diverse robustness checks designed to uncover fundamental weaknesses. A crucial aspect of ContextRef is that images and descriptions are presented in context, reflecting prior work showing that context is important for description quality. Using ContextRef, we assess a variety of pretrained models, scoring functions, and techniques for incorporating context. None of the methods is successful with ContextRef, but we show that careful fine-tuning yields substantial improvements. ContextRef remains a challenging benchmark though, in large part due to the challenge of context dependence.
****
Eliya Nachmani,Alon Levkovitch,Roy Hirsch,Julian Salazar,Chulayuth Asawaroengchai,Soroosh Mariooryad,Ehud Rivlin,RJ Skerry-Ryan,Michelle Tadmor Ramanovich
Spoken Question Answering and Speech Continuation Using Spectrogram-Powered LLM
We present Spectron, a novel approach to adapting pre-trained large language models (LLMs) to perform spoken question answering (QA) and speech continuation. By endowing the LLM with a pre-trained speech encoder, our model becomes able to take speech inputs and generate speech outputs. The entire system is trained end-to-end and operates directly on spectrograms, simplifying our architecture. Key to our approach is a training objective that jointly supervises speech recognition, text continuation, and speech synthesis using only paired speech-text pairs, enabling a `cross-modal' chain-of-thought within a single decoding pass. Our method surpasses existing spoken language models in speaker preservation and semantic coherence. Furthermore, the proposed model improves upon direct initialization in retaining the knowledge of the original LLM as demonstrated through spoken QA datasets. We release our audio samples and spoken QA dataset via our website.
****
Anke Tang,Li Shen,Yong Luo,Yibing Zhan,Han Hu,Bo Du,Yixin Chen,Dacheng Tao
Parameter-Efficient Multi-Task Model Fusion with Partial Linearization
Large pre-trained models have enabled significant advances in machine learning and served as foundation components.
Model fusion methods, such as task arithmetic, have been proven to be powerful and scalable to incorporate fine-tuned weights from different tasks into a multi-task model.
However, efficiently fine-tuning large pre-trained models on multiple downstream tasks remains challenging, leading to inefficient multi-task model fusion.
In this work, we propose a novel method to improve multi-task fusion for parameter-efficient fine-tuning techniques like LoRA fine-tuning.
Specifically, our approach partially linearizes only the adapter modules and applies task arithmetic over the linearized adapters.
This allows us to leverage the the advantages of model fusion over linearized fine-tuning, while still performing fine-tuning and inference efficiently.
We demonstrate that our partial linearization technique enables a more effective fusion of multiple tasks into a single model, outperforming standard adapter tuning and task arithmetic alone.
Experimental results demonstrate the capabilities of our proposed partial linearization technique to effectively construct unified multi-task models via the fusion of fine-tuned task vectors.
We evaluate performance over an increasing number of tasks and find that our approach outperforms standard parameter-efficient fine-tuning techniques. The results highlight the benefits of partial linearization for scalable and efficient multi-task model fusion.

****
Nicholas Konz,Maciej A Mazurowski
The Effect of Intrinsic Dataset Properties on Generalization: Unraveling Learnin g Differences Between Natural and Medical Images
This paper investigates discrepancies in how neural networks learn from differen t imaging domains, which are commonly overlooked when adopting computer vision t echniques from the domain of natural images to other specialized domains such as medical images. Recent works have found that the generalization error of a trai ned network typically increases with the intrinsic dimension (d_{data}) of its training set. Yet, the steepness of this relationship varies significantly betw een medical (radiological) and natural imaging domains, with no existing theoret ical explanation. We address this gap in knowledge by establishing and empirical ly validating a generalization scaling law with respect to d_{data}, and propo se that the substantial scaling discrepancy between the two considered domains m ay be at least partially attributed to the higher intrinsic ``label sharpness'' ($K_\mathcal{F}$) of medical imaging datasets, a metric which we propose. Next, we demonstrate an additional benefit of measuring the label sharpness of a train ing set: it is negatively correlated with the trained model's adversarial robust ness, which notably leads to models for medical images having a substantially hi gher vulnerability to adversarial attack. Finally, we extend our d_{data} form alism to the related metric of learned representation intrinsic dimension ($d_{r epr}$), derive a generalization scaling law with respect to d_{repr}, and show that d_{data} serves as an upper bound for d_{repr}. Our theoretical result s are supported by thorough experiments with six models and eleven natural and m edical imaging datasets over a range of training set sizes. Our findings offer i nsights into the influence of intrinsic dataset properties on generalization, re presentation learning, and robustness in deep neural networks. *Code link: https ://github.com/mazurowski-lab/intrinsic-properties*
****
Alexander G Shypula,Aman Madaan,Yimeng Zeng,Uri Alon,Jacob R. Gardner,Yiming Yan g,Milad Hashemi,Graham Neubig,Parthasarathy Ranganathan,Osbert Bastani,Amir Yazd anbakhsh
Learning Performance-Improving Code Edits
With the waning of Moore's law, optimizing program performance has become a majo r focus of software research. However, high-level optimizations such as API and algorithm changes remain elusive due to the difficulty of understanding the sema ntics of code.
Simultaneously, pretrained large language models (LLMs) have demonstrated strong capabilities at solving a wide range of programming tasks.
To that end, we introduce a framework for adapting LLMs to high-level program op timization.
First, we curate a dataset of performance-improving edits made by human programm ers of over 77,000 competitive C++ programming submission pairs, accompanied by extensive unit tests.
A major challenge is the significant variability of measuring performance on com modity hardware, which can lead to spurious "improvements".
To isolate and reliably evaluate the impact of program optimizations, we design an environment based on the gem5 full system simulator, the de facto simulator u sed in academia and industry.
Next, we propose a broad range of adaptation strategies for code optimization; f or prompting, these include retrieval-based few-shot prompting and chain-of-thou ght, and for finetuning, these include performance-conditioned generation and sy nthetic data augmentation based on self-play.
A combination of these techniques achieves an average speedup of 5.65 times on C odeLlama-13B and 6.86 times on GPT-3.5, surpassing the best human performance (4 .06 times).
We find our proposed performance-conditioned generation is particularly effectiv e at improving performance as well as increasing the fraction of optimized progr ams.
****

Lirong Wu,Yijun Tian,Yufei Huang,Siyuan Li,Haitao Lin,Nitesh V Chawla,Stan Z. Li
MAPE-PPI: Towards Effective and Efficient Protein-Protein Interaction Prediction via Microenvironment-Aware Protein Embedding
Protein-Protein Interactions (PPIs) are fundamental in various biological processes and play a key role in life activities. The growing demand and cost of experimental PPI assays require computational methods for efficient PPI prediction. While existing methods rely heavily on protein sequence for PPI prediction, it is the protein structure that is the key to determine the interactions. To take both protein modalities into account, we define the microenvironment of an amino acid residue by its sequence and structural contexts, which describe the surrounding chemical properties and geometric features. In addition, microenvironments defined in previous work are largely based on experimentally assayed physicochemical properties, for which the "vocabulary" is usually extremely small. This makes it difficult to cover the diversity and complexity of microenvironments. In this paper, we propose Microenvironment-Aware Protein Embedding for PPI prediction (MPAE-PPI), which encodes microenvironments into chemically meaningful discrete codes via a sufficiently large microenvironment "vocabulary" (i.e., codebook). Moreover, we propose a novel pre-training strategy, namely Masked Codebook Modeling (MCM), to capture the dependencies between different microenvironments by randomly masking the codebook and reconstructing the input. With the learned microenvironment codebook, we can reuse it as an off-the-shelf tool to efficiently and effectively encode proteins of different sizes and functions for large-scale PPI prediction. Extensive experiments show that MAPE-PPI can scale to PPI prediction with millions of PPIs with superior trade-offs between effectiveness and computational efficiency than the state-of-the-art competitors.
****
Wenyu Jiang,Hao Cheng,MingCai Chen,Chongjun Wang,Hongxin Wei
DOS: Diverse Outlier Sampling for Out-of-Distribution Detection
Modern neural networks are known to give overconfident predictions for out-of-distribution inputs when deployed in the open world. It is common practice to leverage a surrogate outlier dataset to regularize the model during training, and recent studies emphasize the role of uncertainty in designing the sampling strategy for outlier datasets. However, the OOD samples selected solely based on predictive uncertainty can be biased towards certain types, which may fail to capture the full outlier distribution. In this work, we empirically show that diversity is critical in sampling outliers for OOD detection performance. Motivated by the observation, we propose a straightforward and novel sampling strategy named DOS (Diverse Outlier Sampling) to select diverse and informative outliers. Specifically, we cluster the normalized features at each iteration, and the most informative outlier from each cluster is selected for model training with absent category loss. With DOS, the sampled outliers efficiently shape a globally compact decision boundary between ID and OOD data. Extensive experiments demonstrate the superiority of DOS, reducing the average FPR95 by up to 25.79% on CIFAR-100 with TI-300K.
****
Xu Ma,Xiyang Dai,Jianwei Yang,Bin Xiao,Yinpeng Chen,Yun Fu,Lu Yuan
Efficient Modulation for Vision Networks
In this work, we present efficient modulation, a novel design for efficient vision networks. We revisit the modulation mechanism, which operates input through convolutional context modeling and feature projection layers, and fuses features via element-wise multiplication and an MLP block. We demonstrate that the abstracted modulation mechanism is particularly well suited for efficient networks and further tailor the modulation design by proposing the efficient modulation (EfficientMod) block, which is considered the essential building block for our networks. Bene- fiting from the prominent representational ability of modulation mechanism and the efficiency of efficient modulation design, our network can accomplish better accuracy-efficiency trade-offs and set new state-of-the-art performance for efficient networks. When integrating EfficientMod block with the vanilla self-attention block, we obtain the hybrid architecture and further improve the performance without sacrificing the efficiency. We carry out comprehensive exper

iments to verify EfficientMod's performance. With fewer parameters, our EfficientMod-s performs 0.6 top-1 accuracy better than the prior state-of-the-art approach EfficientFormerV2-s2 without any training tricks and is 25% faster on GPU. Additionally, our method presents a notable improvement in downstream tasks, outperforming EfficientFormerV2-s by 3.6 mIoU on the ADE20K benchmark. Code and checkpoints are available at https://github.com/ma-xu/EfficientMod.
****
Siming Yan,Chen Song,Youkang Kong,Qixing Huang
Multi-View Representation is What You Need for Point-Cloud Pre-Training
A promising direction for pre-training 3D point clouds is to leverage the massive amount of data in 2D, whereas the domain gap between 2D and 3D creates a fundamental challenge. This paper proposes a novel approach to point-cloud pre-training that learns 3D representations by leveraging pre-trained 2D networks. Different from the popular practice of predicting 2D features first and then obtaining 3D features through dimensionality lifting, our approach directly uses a 3D network for feature extraction. We train the 3D feature extraction network with the help of the novel 2D knowledge transfer loss, which enforces the 2D projections of the 3D feature to be consistent with the output of pre-trained 2D networks. To prevent the feature from discarding 3D signals, we introduce the multi-view consistency loss that additionally encourages the projected 2D feature representations to capture pixel-wise correspondences across different views. Such correspondences induce 3D geometry and effectively retain 3D features in the projected 2D features. Experimental results demonstrate that our pre-trained model can be successfully transferred to various downstream tasks, including 3D shape classification, part segmentation, 3D object detection, and semantic segmentation, achieving state-of-the-art performance.
****
Tianyu Guo,Wei Hu,Song Mei,Huan Wang,Caiming Xiong,Silvio Savarese,Yu Bai
How Do Transformers Learn In-Context Beyond Simple Functions? A Case Study on Learning with Representations
While large language models based on the transformer architecture have demonstrated remarkable in-context learning (ICL) capabilities, understandings of such capabilities are still in an early stage, where existing theory and mechanistic understanding focus mostly on simple scenarios such as learning simple function classes. This paper takes initial steps on understanding ICL in more complex scenarios, by studying learning with \emph{representations}. Concretely, we construct synthetic in-context learning problems with a compositional structure, where the label depends on the input through a possibly complex but \emph{fixed} representation function, composed with a linear function that \emph{differs} in each instance. By construction, the optimal ICL algorithm first transforms the inputs by the representation function, and then performs linear ICL on top of the transformed dataset. We show theoretically the existence of transformers that approximately implement such algorithms with mild depth and size. Empirically, we find trained transformers consistently achieve near-optimal ICL performance in this setting, and exhibit the desired dissection where lower layers transforms the dataset and upper layers perform linear ICL. Through extensive probing and a new pasting experiment, we further reveal several mechanisms within the trained transformers, such as concrete copying behaviors on both the inputs and the representations, linear ICL capability of the upper layers alone, and a post-ICL representation selection mechanism in a harder mixture setting. These observed mechanisms align well with our theory and may shed light on how transformers perform ICL in more realistic scenarios.
****
Arda Sahiner,Tolga Ergen,Batu Ozturkler,John M. Pauly,Morteza Mardani,Mert Pilanci
Scaling Convex Neural Networks with Burer-Monteiro Factorization
It has been demonstrated that the training problem for a variety of (non) linear two-layer neural networks (such as two-layer perceptrons, convolutional networks, and self-attention) can be posed as equivalent convex optimization problems, with an induced regularizer which encourages low rank. However, this regularizer

becomes prohibitively expensive to compute at moderate scales, impeding training convex neural networks. To this end, we propose applying the Burer-Monteiro factorization to convex neural networks, which for the first time enables a Burer-Monteiro perspective on neural networks with non-linearities. This factorization leads to an equivalent yet computationally tractable non-convex alternative with no spurious local minima. We develop a novel relative optimality bound of stationary points of the Burer-Monteiro factorization, providing verifiable conditions under which any stationary point is a global optimum. Further, for the first time, we show that linear self-attention with sufficiently many heads has no spurious local minima. Our experiments validate the novel relative optimality bound and the utility of the Burer-Monteiro factorization for scaling convex neural networks.
****

Dennis Frauen,Fergus Imrie,Alicia Curth,Valentyn Melnychuk,Stefan Feuerriegel,Mihaela van der Schaar

A Neural Framework for Generalized Causal Sensitivity Analysis

Unobserved confounding is common in many applications, making causal inference from observational data challenging. As a remedy, causal sensitivity analysis is an important tool to draw causal conclusions under unobserved confounding with mathematical guarantees. In this paper, we propose NeuralCSA, a neural framework for generalized causal sensitivity analysis. Unlike previous work, our framework is compatible with (i) a large class of sensitivity models, including the marginal sensitivity model, f-sensitivity models, and Rosenbaum's sensitivity model; (ii) different treatment types (i.e., binary and continuous); and (iii) different causal queries, including (conditional) average treatment effects and simultaneous effects on multiple outcomes. This generality is achieved by learning a latent distribution shift that corresponds to a treatment intervention using two conditional normalizing flows. We provide theoretical guarantees that NeuralCSA is able to infer valid bounds on the causal query of interest and also demonstrate this empirically using both simulated and real-world data.
****

Haonan Qiu,Menghan Xia,Yong Zhang,Yingqing He,Xintao Wang,Ying Shan,Ziwei Liu

FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling

With the availability of large-scale video datasets and the advances of diffusion models, text-driven video generation has achieved substantial progress. However, existing video generation models are typically trained on a limited number of frames, resulting in the inability to generate high-fidelity long videos during inference. Furthermore, these models only support single-text conditions, whereas real-life scenarios often require multi-text conditions as the video content changes over time. To tackle these challenges, this study explores the potential of extending the text-driven capability to generate longer videos conditioned on multiple texts. 1) We first analyze the impact of initial noise in video diffusion models. Then building upon the observation of noise, we propose FreeNoise, a tuning-free and time-efficient paradigm to enhance the generative capabilities of pretrained video diffusion models while preserving content consistency. Specifically, instead of initializing noises for all frames, we reschedule a sequence of noises for long-range correlation and perform temporal attention over them by window-based fusion. 2) Additionally, we design a novel motion injection method to support the generation of videos conditioned on multiple text prompts. Extensive experiments validate the superiority of our paradigm in extending the generative capabilities of video diffusion models. It is noteworthy that compared with the previous best-performing method which brought about 255% extra time cost, our method incurs only negligible time cost of approximately 17%. Generated video samples are available at our website: http://haonanqiu.com/projects/FreeNoise.html.
****

Derek Lim,Haggai Maron,Marc T. Law,Jonathan Lorraine,James Lucas

Graph Metanetworks for Processing Diverse Neural Architectures

Neural networks efficiently encode learned information within their parameters. Consequently, many tasks can be unified by treating neural networks themselves a

s input data. When doing so, recent studies demonstrated the importance of accou
nting for the symmetries and geometry of parameter spaces. However, those works
developed architectures tailored to specific networks such as MLPs and CNNs with
out normalization layers, and generalizing such architectures to other types of
networks can be challenging. In this work, we overcome these challenges by build
ing new metanetworks --- neural networks that take weights from other neural net
works as input. Put simply, we carefully build graphs representing the input neu
ral networks and process the graphs using graph neural networks. Our approach, G
raph Metanetworks (GMNs), generalizes to neural architectures where competing me
thods struggle, such as multi-head attention layers, normalization layers, convo
lutional layers, ResNet blocks, and group-equivariant linear layers. We prove th
at GMNs are expressive and equivariant to parameter permutation symmetries that
leave the input neural network functions unchanged. We validate the effectivenes
s of our method on several metanetwork tasks over diverse neural network archite
ctures.
****
Moyang Li,Peng Wang,Lingzhe Zhao,Bangyan Liao,Peidong Liu
USB-NeRF: Unrolling Shutter Bundle Adjusted Neural Radiance Fields
Neural Radiance Fields (NeRF) has received much attention recently due to its im
pressive capability to represent 3D scene and synthesize novel view images. Exis
ting works usually assume that the input images are captured by a global shutter
camera. Thus, rolling shutter (RS) images cannot be trivially applied to an off
-the-shelf NeRF algorithm for novel view synthesis. Rolling shutter effect would
also affect the accuracy of the camera pose estimation (e.g. via COLMAP), which
further prevents the success of NeRF algorithm with RS images.
In this paper, we propose Unrolling Shutter Bundle Adjusted Neural Radiance Fiel
ds (USB-NeRF). USB-NeRF is able to correct rolling shutter distortions and recov
er accurate camera motion trajectory simultaneously under the framework of NeRF,
by modeling the physical image formation process of a RS camera.
Experimental results demonstrate that USB-NeRF achieves better performance compa
red to prior works, in terms of RS effect removal, novel view image synthesis as
well as camera motion estimation. Furthermore, our algorithm can also be used t
o recover high-fidelity high frame-rate global shutter video from a sequence of
RS images.
****
O■uz Kaan Yüksel,Etienne Boursier,Nicolas Flammarion
First-order ANIL provably learns representations despite overparametrisation
Due to its empirical success in few-shot classification and reinforcement learni
ng, meta-learning has recently received significant interest. Meta-learning meth
ods leverage data from previous tasks to learn a new task in a sample-efficient
manner. In particular, model-agnostic methods look for initialization points fro
m which gradient descent quickly adapts to any new task. Although it has been em
pirically suggested that such methods perform well by learning shared representa
tions during pretraining, there is limited theoretical evidence of such behavior
. More importantly, it has not been shown that these methods still learn a share
d structure, despite architectural misspecifications. In this direction, this wo
rk shows, in the limit of an infinite number of tasks, that first-order ANIL wit
h a linear two-layer network architecture successfully learns linear shared repr
esentations. This result even holds with _overparametrization_; having a width l
arger than the dimension of the shared representations results in an asymptotica
lly low-rank solution. The learned solution then yields a good adaptation perfor
mance on any new task after a single gradient step. Overall, this illustrates ho
w well model-agnostic methods such as first-order ANIL can learn shared represen
tations.
****
Nabeel Seedat,Fergus Imrie,Mihaela van der Schaar
Dissecting Sample Hardness: A Fine-Grained Analysis of Hardness Characterization
Methods for Data-Centric AI
Characterizing samples that are difficult to learn from is crucial to developing
highly performant ML models. This has led to numerous Hardness Characterization

Methods (HCMs) that aim to identify ''hard'' samples. However, there is a lack of consensus regarding the definition and evaluation of ''hardness''. Unfortunately, current HCMs have only been evaluated on specific types of hardness and often only qualitatively or with respect to downstream performance, overlooking the fundamental quantitative identification task. We address this gap by presenting a fine-grained taxonomy of hardness types. Additionally, we propose the Hardness Characterization Analysis Toolkit (H-CAT), which supports comprehensive and quantitative benchmarking of HCMs across the hardness taxonomy and can easily be extended to new HCMs, hardness types, and datasets. We use H-CAT to evaluate 13 different HCMs across 8 hardness types. This comprehensive evaluation encompassing over 14K setups uncovers strengths and weaknesses of different HCMs, leading to practical tips to guide HCM selection and future development. Our findings highlight the need for more comprehensive HCM evaluation, while we hope our hardness taxonomy and toolkit will advance the principled evaluation and uptake of data-centric AI methods.
****
Yingtao Zhang,Jialin Zhao,Wenjing Wu,Alessandro Muscoloni,Carlo Vittorio Cannistraci
Epitopological learning and Cannistraci-Hebb network shape intelligence brain-inspired theory for ultra-sparse advantage in deep learning
Sparse training (ST) aims to ameliorate deep learning by replacing fully connected artificial neural networks (ANNs) with sparse or ultra-sparse ones, such as brain networks are, therefore it might benefit to borrow brain-inspired learning paradigms from complex network intelligence theory. Here, we launch the ultra-sparse advantage challenge, whose goal is to offer evidence on the extent to which ultra-sparse (around 1\% connection retained) topologies can achieve any leaning advantage against fully connected. Epitopological learning is a field of network science and complex network intelligence that studies how to implement learning on complex networks by changing the shape of their connectivity structure (epitopological plasticity). One way to implement Epitopological (epi- means new) Learning is via link prediction: predicting the likelihood of non-observed links to appear in the network. Cannistraci-Hebb learning theory inspired the CH3-L3 network automata rule for link prediction which is effective for general-purpose link prediction. Here, starting from CH3-L3 we propose Epitopological Sparse Meta-deep Learning (ESML) to apply Epitopological Learning to sparse training. In empirical experiments, we find that ESML learns ANNs with ultra-sparse hyperbolic (epi-)topology in which emerges a community layer organization that is meta-deep (meaning that each layer also has an internal depth due to power-law node hierarchy). Furthermore, we discover that ESML can in many cases automatically sparse the neurons during training (arriving even to 30\% neurons left in hidden layers), this process of node dynamic removal is called percolation. Starting from this network science evidence, we design Cannistraci-Hebb training (CHT), a 4-step training methodology that puts ESML at its heart. We conduct experiments on 7 datasets and 5 network structures comparing CHT to dynamic sparse training SOTA algorithms and the fully connected counterparts. The results indicate that, with a mere 1\% of links retained during training, CHT surpasses fully connected networks on VGG16, GoogLeNet, ResNet50, and ResNet152. This key finding is an evidence for ultra-sparse advantage and signs a milestone in deep learning. CHT acts akin to a gradient-free oracle that adopts CH3-L3-based epitopological learning to guide the placement of new links in the ultra-sparse network topology to facilitate sparse-weight gradient learning, and this in turn reduces the convergence time of ultra-sparse training. Finally, CHT offers the first examples of parsimony dynamic sparse training because, in many datasets, it can retain network performance by percolating and significantly reducing the node network size.
Our code is available at: https://github.com/biomedical-cybernetics/Cannistraci-Hebb-training
****
Yuxiao Cheng,Ziqian Wang,Tingxiong Xiao,Qin Zhong,Jinli Suo,Kunlun He
CausalTime: Realistically Generated Time-series for Benchmarking of Causal Discovery

Time-series causal discovery (TSCD) is a fundamental problem of machine learning . However, existing synthetic datasets cannot properly evaluate or predict the algorithms' performance on real data. This study introduces the CausalTime pipeline to generate time-series that highly resemble the real data and with ground truth causal graphs for quantitative performance evaluation. The pipeline starts from real observations in a specific scenario and produces a matching benchmark dataset. Firstly, we harness deep neural networks along with normalizing flow to accurately capture realistic dynamics. Secondly, we extract hypothesized causal graphs by performing importance analysis on the neural network or leveraging prior knowledge. Thirdly, we derive the ground truth causal graphs by splitting the causal model into causal term, residual term, and noise term. Lastly, using the fitted network and the derived causal graph, we generate corresponding versatile time-series proper for algorithm assessment. In the experiments, we validate the fidelity of the generated data through qualitative and quantitative experiments, followed by a benchmarking of existing TSCD algorithms using these generated datasets. CausalTime offers a feasible solution to evaluating TSCD algorithms in real applications and can be generalized to a wide range of fields. For easy use of the proposed approach, we also provide a user-friendly website, hosted on www.causaltime.cc.
****
Yuhang Liu,Zhen Zhang,Dong Gong,Mingming Gong,Biwei Huang,Anton van den Hengel,Kun Zhang,Javen Qinfeng Shi
Identifiable Latent Polynomial Causal Models through the Lens of Change
Causal representation learning aims to unveil latent high-level causal representations from observed low-level data. One of its primary tasks is to provide reliable assurance of identifying these latent causal models, known as \textit{identifiability}. A recent breakthrough explores identifiability by leveraging the change of causal influences among latent causal variables across multiple environments \citep{liu2022identifying}. However, this progress rests on the assumption that the causal relationships among latent causal variables adhere strictly to linear Gaussian models. In this paper, we extend the scope of latent causal models to involve nonlinear causal relationships, represented by polynomial models, and general noise distributions conforming to the exponential family. Additionally, we investigate the necessity of imposing changes on all causal parameters and present partial identifiability results when part of them remains unchanged. Further, we propose a novel empirical estimation method, grounded in our theoretical finding, that enables learning consistent latent causal representations. Our experimental results, obtained from both synthetic and real-world data, validate our theoretical contributions concerning identifiability and consistency.
****
Ziwei Guan,Yi Zhou,Yingbin Liang
On the Hardness of Online Nonconvex Optimization with Single Oracle Feedback
Online nonconvex optimization has been an active area of research recently. Previous studies either considered the global regret with full information about the objective functions, or studied the local regret with window-smoothed objective functions, which required access to unlimited number of gradient oracles per time step. In this paper, we focus on the more challenging and practical setting, where access to only a single oracle is allowed per time step, and take the local regret of the original (i.e., unsmoothed) objective functions as the performance metric. Specifically, for both settings respectively with a single exact and stochastic gradient oracle feedback, we derive lower bounds on the local regret and show that the classical online (stochastic) gradient descent algorithms are optimal. Moreover, for the more challenging setting with a single function value oracle feedback, we develop an online algorithm based on a one-point running difference gradient estimator, and show that such an algorithm achieves a local regret that a generic stochastic gradient oracle can best achieve.
****
Joey Hejna,Rafael Rafailov,Harshit Sikchi,Chelsea Finn,Scott Niekum,W. Bradley Knox,Dorsa Sadigh
Contrastive Preference Learning: Learning from Human Feedback without Reinforcem

ent Learning
Reinforcement Learning from Human Feedback (RLHF) has emerged as a popular paradigm for aligning models with human intent. Typically RLHF algorithms operate in two phases: first, use human preferences to learn a reward function and second, align the model by optimizing the learned reward via reinforcement learning (RL). This paradigm assumes that human preferences are distributed according to reward, but recent work suggests that they instead follow the regret under the user's optimal policy. Thus, learning a reward function from feedback is not only based on a flawed assumption of human preference, but also leads to unwieldy optimization challenges that stem from policy gradients or bootstrapping in the RL phase. Because of these optimization challenges, contemporary RLHF methods restrict themselves to contextual bandit settings (e.g., as in large language models) or limit observation dimensionality (e.g., state-based robotics). We overcome these limitations by introducing a new family of algorithms for optimizing behavior from human feedback using the regret model of human preferences. Using the principle of maximum entropy, we derive Contrastive Preference Learning (CPL), an algorithm for learning optimal policies from preferences without learning reward functions, circumventing the need for RL. CPL is fully off-policy, uses only a simple contrastive objective, and can be applied to arbitrary MDPs. In contrast to prior work, this enables CPL to elegantly scale to high-dimensional and sequential RLHF problems.
****
Zhan Zhuang,Yu Zhang,Ying Wei
Gradual Domain Adaptation via Gradient Flow
Domain shift degrades classification models on new data distributions. Conventional unsupervised domain adaptation (UDA) aims to learn features that bridge labeled source and unlabeled target domains. In contrast to feature learning, gradual domain adaptation (GDA) leverages extra continuous intermediate domains with pseudo-labels to boost the source classifier. However, real intermediate domains are sometimes unavailable or ineffective. In this paper, we propose $\textbf{G}$radual Domain Adaptation via $\textbf{G}$radient $\textbf{F}$low (GGF) to generate intermediate domains with preserving labels, thereby enabling us a fine-tuning method for GDA. We employ the Wasserstein gradient flow in Kullback–Leibler divergence to transport samples from the source to the target domain. To simulate the dynamics, we utilize the Langevin algorithm. Since the Langevin algorithm disregards label information and introduces diffusion noise, we introduce classifier-based and sample-based potentials to avoid label switching and dramatic deviations in the sampling process. For the proposed GGF model, we analyze its generalization bound. Experiments on several benchmark datasets demonstrate the superiority of the proposed GGF method compared to state-of-the-art baselines.
****
Eshant English,Matthias Kirchler,Christoph Lippert
Kernelised Normalising Flows
Normalising Flows are non-parametric statistical models known for their dual capabilities of density estimation and generation. They are distinguished by their inherently invertible architecture. However, the requirement of invertibility imposes constraints on their expressiveness, necessitating a large number of parameters and innovative architectural designs to achieve satisfactory outcomes. Whilst flow-based models predominantly rely on neural-network-based transformations for expressive designs, alternative transformation methods have received limited attention. In this work, we present Ferumal flow, a novel kernelised normalising flow paradigm that integrates kernels into the framework. Our results demonstrate that a kernelised flow can yield competitive or superior results compared to neural network-based flows whilst maintaining parameter efficiency. Kernelised flows excel especially in the low-data regime, enabling flexible non-parametric density estimation in applications with sparse data availability.
****
Jian Kang,Yinglong Xia,Ross Maciejewski,Jiebo Luo,Hanghang Tong
Deceptive Fairness Attacks on Graphs via Meta Learning
We study deceptive fairness attacks on graphs to answer the following question:

How can we achieve poisoning attacks on a graph learning model to exacerbate the bias deceptively? We answer this question via a bi-level optimization problem and propose a meta learning-based framework named FATE. FATE is broadly applicable with respect to various fairness definitions and graph learning models, as well as arbitrary choices of manipulation operations. We further instantiate FATE to attack statistical parity or individual fairness on graph neural networks. We conduct extensive experimental evaluations on real-world datasets in the task of semi-supervised node classification. The experimental results demonstrate that FATE could amplify the bias of graph neural networks with or without fairness consideration while maintaining the utility on the downstream task. We hope this paper provides insights into the adversarial robustness of fair graph learning and can shed light on designing robust and fair graph learning in future studies.
****

Jan Schneider,Pierre Schumacher,Simon Guist,Le Chen,Daniel Haeufle,Bernhard Schölkopf,Dieter Büchler

Identifying Policy Gradient Subspaces

Policy gradient methods hold great potential for solving complex continuous control tasks. Still, their training efficiency can be improved by exploiting structure within the optimization problem. Recent work indicates that supervised learning can be accelerated by leveraging the fact that gradients lie in a low-dimensional and slowly-changing subspace. In this paper, we conduct a thorough evaluation of this phenomenon for two popular deep policy gradient methods on various simulated benchmark tasks. Our results demonstrate the existence of such gradient subspaces despite the continuously changing data distribution inherent to reinforcement learning. These findings reveal promising directions for future work on more efficient reinforcement learning, e.g., through improving parameter-space exploration or enabling second-order optimization.
****

Yi-Lin Sung,Jaehong Yoon,Mohit Bansal

ECoFLaP: Efficient Coarse-to-Fine Layer-Wise Pruning for Vision-Language Models

Large Vision-Language Models (LVLMs) can understand the world comprehensively by integrating rich information from different modalities, achieving remarkable performance improvements on various multimodal downstream tasks. However, deploying LVLMs is often problematic due to their massive computational/energy costs and carbon consumption, making it infeasible to adopt conventional iterative global pruning, which is costly due to computing the Hessian matrix of the entire large model for sparsification. Alternatively, several studies have recently proposed layer-wise pruning approaches to avoid the expensive computation of global pruning and efficiently compress model weights according to their importance within a layer. However, these methods often suffer from suboptimal model compression due to their lack of a global perspective. To address this limitation in recent efficient pruning methods for large models, we propose Efficient Coarse-to-Fine Layer-Wise Pruning (ECoFLaP), a two-stage coarse-to-fine weight pruning approach for LVLMs. We first determine the sparsity ratios of different layers or blocks by leveraging the global importance score, which is efficiently computed based on the zeroth-order approximation of the global model gradients. Then, the multimodal model performs layer-wise unstructured weight pruning. We validate our proposed method across various multi-modal and single-modal models and datasets, demonstrating significant performance improvements over prevalent pruning techniques in the high-sparsity regime.
****

Khalid Oublal,Said Ladjal,David Benhaiem,Emmanuel LE BORGNE,François Roueff

Disentangling Time Series Representations via Contrastive Independence-of-Support on l-Variational Inference

Learning disentangled representations for Time Series is a promising path to facilitate reliable generalization to in- and out-of distribution, offering benefits like feature derivation and improved interpretability and fairness, thereby enhancing downstream tasks. We focus on disentangled representation learning for home appliance electricity usage, enabling users to understand and optimize their consumption for a reduced carbon footprint. Our approach frames the problem as

disentangling each attribute's role in total consumption. Unlike existing method s assuming attribute independence which leads to non-identiability, we acknowled ge real-world time series attribute correlations, learned up to a smooth bijecti on using contrastive learning and a single encoder. To address this, we propose a Disentanglement under Independence-of-Support via Contrastive Learning, facili tating representation generalization across diverse correlated scenarios. Our me thod utilizes innovative l-variational inference layers with self-attention, eff ectively addressing temporal dependencies across bottom-up and top-down networks . We find that DIOSC can enhance the task of representation of time series elect ricity consumption. We introduce TDS (Time Disentangling Score) to gauge disenta nglement quality. TDS reliably reflects disentanglement performance, making it a valuable metric for evaluating time series representations disentanglement. Cod e available at https://github.com/time-disentanglement-lib.
****
Amitis Shidani,R Devon Hjelm,Jason Ramapuram,Russell Webb,Eeshan Gunesh Dhekane, Dan Busbridge
Poly-View Contrastive Learning
Contrastive learning typically matches pairs of related views among a number of unrelated negative views. Views can be generated (e.g. by augmentations) or be o bserved. We investigate matching when there are more than two related views whic h we call poly-view tasks,
and derive new representation learning objectives using information maximization and sufficient statistics. We show that with unlimited computation, one should maximize the number of related views, and with a fixed compute budget, it is ben eficial to decrease the number of unique samples whilst increasing the number of views of those samples. In particular, poly-view contrastive models trained for 128 epochs with batch size 256 outperform SimCLR trained for 1024 epochs at bat ch size 4096 on ImageNet1k, challenging the belief that contrastive models requi re large batch sizes and many training epochs.
****
Ming-Yu Chung,Sheng-Yen Chou,Chia-Mu Yu,Pin-Yu Chen,Sy-Yen Kuo,Tsung-Yi Ho
Rethinking Backdoor Attacks on Dataset Distillation: A Kernel Method Perspective
Dataset distillation offers a potential means to enhance data efficiency in deep learning. Recent studies have shown its ability to counteract backdoor risks pr esent in original training samples. In this study, we delve into the theoretical aspects of backdoor attacks and dataset distillation based on kernel methods. W e introduce two new theory-driven trigger pattern generation methods specialized for dataset distillation. Following a comprehensive set of analyses and experim ents, we show that our optimization-based trigger design framework informs effec tive backdoor attacks on dataset distillation. Notably, datasets poisoned by our designed trigger prove resilient against conventional backdoor attack detection and mitigation methods. Our empirical results validate that the triggers devel oped using our approaches are proficient at executing resilient backdoor attacks .
****
Hadi Beik Mohammadi,Søren Hauberg,Georgios Arvanitidis,Nadia Figueroa,Gerhard Ne umann,Leonel Rozo
Neural Contractive Dynamical Systems
Stability guarantees are crucial when ensuring that a fully autonomous robot doe s not take undesirable or potentially harmful actions. Unfortunately, global sta bility guarantees are hard to provide in dynamical systems learned from data, es pecially when the learned dynamics are governed by neural networks. We propose a novel methodology to learn \emph{neural contractive dynamical systems}, where o ur neural architecture ensures contraction, and hence, global stability. To effi ciently scale the method to high-dimensional dynamical systems, we develop a var iant of the variational autoencoder that learns dynamics in a low-dimensional la tent representation space while retaining contractive stability after decoding. We further extend our approach to learning contractive systems on the Lie group of rotations to account for full-pose end-effector dynamic motions. The result i s the first highly flexible learning architecture that provides contractive stab

ility guarantees with capability to perform obstacle avoidance. Empirically, we demonstrate that our approach encodes the desired dynamics more accurately than the current state-of-the-art, which provides less strong stability guarantees.
****
Ziqi Wang,Chengpeng Hu,Jialin Liu,Xin Yao
Negatively Correlated Ensemble Reinforcement Learning for Online Diverse Game Level Generation
Deep reinforcement learning has recently been successfully applied to online procedural content generation in which a policy determines promising game-level segments. However, existing methods can hardly discover diverse level patterns, while the lack of diversity makes the gameplay boring. This paper proposes an ensemble reinforcement learning approach that uses multiple negatively correlated sub-policies to generate different alternative level segments, and stochastically selects one of them following a selector model. A novel policy regularisation technique is integrated into the approach to diversify the generated alternatives. In addition, we develop theorems to provide general methodologies for optimising policy regularisation in a Markov decision process. The proposed approach is compared with several state-of-the-art policy ensemble methods and classic methods on a well-known level generation benchmark, with two different reward functions expressing game-design goals from different perspectives. Results show that our approach boosts level diversity notably with competitive performance in terms of the reward. Furthermore, by varying the regularisation coefficient, the trained generators form a well-spread Pareto front, allowing explicit trade-offs between diversity and rewards of generated levels.
****
Depen Morwani,Benjamin L. Edelman,Costin-Andrei Oncescu,Rosie Zhao,Sham M. Kakade
Feature emergence via margin maximization: case studies in algebraic tasks
Understanding the internal representations learned by neural networks is a cornerstone challenge in the science of machine learning. While there have been significant recent strides in some cases towards understanding *how* neural networks implement specific target functions, this paper explores a complementary question -- *why* do networks arrive at particular computational strategies?
Our inquiry focuses on the algebraic learning tasks of modular addition, sparse parities, and finite group operations. Our primary theoretical findings analytically characterize the features learned by stylized neural networks for these algebraic tasks. Notably, our main technique demonstrates how the principle of margin maximization alone can be used to fully specify the features learned by the network.
Specifically, we prove that the trained networks utilize Fourier features to perform modular addition and employ features corresponding to irreducible group-theoretic representations to perform compositions in general groups, aligning closely with the empirical observations of Nanda et al. (2023) and Chughtai et al. (2023). More generally, we hope our techniques can help to foster a deeper understanding of why neural networks adopt specific computational strategies.
****
Pratik Patil,Daniel LeJeune
Asymptotically Free Sketched Ridge Ensembles: Risks, Cross-Validation, and Tuning
We employ random matrix theory to establish consistency of generalized cross validation (GCV) for estimating prediction risks of sketched ridge regression ensembles, enabling efficient and consistent tuning of regularization and sketching parameters. Our results hold for a broad class of asymptotically free sketches under very mild data assumptions. For squared prediction risk, we provide a decomposition into an unsketched equivalent implicit ridge bias and a sketching-based variance, and prove that the risk can be globally optimized by only tuning sketch size in infinite ensembles. For general subquadratic prediction risk functionals, we extend GCV to construct consistent risk estimators, and thereby obtain distributional convergence of the GCV-corrected predictions in Wasserstein-2 metric. This in particular allows construction of prediction intervals with asymptoti

cally correct coverage conditional on the training data. We also propose an "ens emble trick" whereby the risk for unsketched ridge regression can be efficiently estimated via GCV using small sketched ridge ensembles. We empirically validate our theoretical results using both synthetic and real large-scale datasets with practical sketches including CountSketch and subsampled randomized discrete cos ine transforms.
****
Elias Frantar,Carlos Riquelme Ruiz,Neil Houlsby,Dan Alistarh,Utku Evci
Scaling Laws for Sparsely-Connected Foundation Models
We explore the impact of parameter sparsity on the scaling behavior of Transform ers trained on massive datasets (i.e., "foundation models"), in both vision and language domains. In this setting, we identify the first scaling law describing the relationship between weight sparsity, number of non-zero parameters, and amo unt of training data, which we validate empirically across model and data scales ; on ViT/JFT-4B and T5/C4. These results allow us to characterize the "optimal s parsity", the sparsity level which yields the best performance for a given effec tive model size and training budget. For a fixed number of non-zero parameters, we identify that the optimal sparsity increases with the amount of data used for training. We also extend our study to different sparsity structures (such as th e hardware-friendly n:m pattern) and strategies (such as starting from a pretrai ned dense model). Our findings shed light on the power and limitations of weight sparsity across various parameter and computational settings, offering both the oretical understanding and practical implications for leveraging sparsity toward s computational efficiency improvements. We provide pruning and scaling law fitt ing code at: github.com/google-research/jaxpruner/tree/main/jaxpruner/projects/b igsparse.
****
Zichen Liu,Chao Du,Wee Sun Lee,Min Lin
Locality Sensitive Sparse Encoding for Learning World Models Online
Acquiring an accurate world model $\textit{online}$ for model-based reinforcemen t learning (MBRL) is challenging due to data nonstationarity, which typically ca uses catastrophic forgetting for neural networks (NNs). From the online learning perspective, a Follow-The-Leader (FTL) world model is desirable, which optimall y fits all previous experiences at each round. Unfortunately, NN-based models ne ed re-training on all accumulated data at every interaction step to achieve FTL, which is computationally expensive for lifelong agents. In this paper, we revis it models that can achieve FTL with incremental updates. Specifically, our world model is a linear regression model supported by nonlinear random features. The linear part ensures efficient FTL update while the nonlinear random feature empo wers the fitting of complex environments. To best trade off model capacity and c omputation efficiency, we introduce a locality sensitive sparse encoding, which allows us to conduct efficient sparse updates even with very high dimensional no nlinear features. We validate the representation power of our encoding and verif y that it allows efficient online learning under data covariate shift. We also s how, in the Dyna MBRL setting, that our world models learned online using a $\te xtit{single pass}$ of trajectory data either surpass or match the performance of deep world models trained with replay and other continual learning methods.
****
Mingyuan Sun,Donghao Zhang,Zongyuan Ge,WANG Jiaxu,Jia Li,Zheng Fang,Renjing Xu
EventRPG: Event Data Augmentation with Relevance Propagation Guidance
Event camera, a novel bio-inspired vision sensor, has drawn a lot of attention f or its low latency, low power consumption, and high dynamic range. Currently, ov erfitting remains a critical problem in event-based classification tasks for Spi king Neural Network (SNN) due to its relatively weak spatial representation capa bility. Data augmentation is a simple but efficient method to alleviate overfitt ing and improve the generalization ability of neural networks, and saliency-base d augmentation methods are proven to be effective in the image processing field. However, there is no approach available for extracting saliency maps from SNNs. Therefore, for the first time, we present Spiking Layer-Time-wise Relevance Pro pagation rule (SLTRP) and Spiking Layer-wise Relevance Propagation rule (SLRP) i

n order for SNN to generate stable and accurate CAMs and saliency maps. Based on this, we propose EventRPG, which leverages relevance propagation on the spiking neural network for more efficient augmentation. Our proposed method has been evaluated on several SNN structures, achieving state-of-the-art performance in object recognition tasks including N-Caltech101, CIFAR10-DVS, with accuracies of 85.62% and 85.55%, as well as action recognition task SL-Animals with an accuracy of 91.59%. Our code is available at https://github.com/myuansun/EventRPG.
****
Lin Chen,Michal Lukasik,Wittawat Jitkrittum,Chong You,Sanjiv Kumar
On Bias-Variance Alignment in Deep Models
Classical wisdom in machine learning holds that the generalization error can be decomposed into bias and variance, and these two terms exhibit a \emph{trade-off}. However, in this paper, we show that for an ensemble of deep learning based classification models, bias and variance are \emph{aligned} at a sample level, where squared bias is approximately \emph{equal} to variance for correctly classified sample points. We present empirical evidence confirming this phenomenon in a variety of deep learning models and datasets. Moreover, we study this phenomenon from two theoretical perspectives: calibration and neural collapse. We first show theoretically that under the assumption that the models are well calibrated, we can observe the bias-variance alignment. Second, starting from the picture provided by the neural collapse theory, we show an approximate correlation between bias and variance.
****
Chunming He,Kai Li,Yachao Zhang,Yulun Zhang,Chenyu You,Zhenhua Guo,Xiu Li,Martin Danelljan,Fisher Yu
Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects
Camouflaged object detection (COD) is the challenging task of identifying camouflaged objects visually blended into surroundings. Albeit achieving remarkable success, existing COD detectors still struggle to obtain precise results in some challenging cases. To handle this problem, we draw inspiration from the prey-vs-predator game that leads preys to develop better camouflage and predators to acquire more acute vision systems and develop algorithms from both the prey side and the predator side. On the prey side, we propose an adversarial training framework, Camouflageator, which introduces an auxiliary generator to generate more camouflaged objects that are harder for a COD method to detect. Camouflageator trains the generator and detector in an adversarial way such that the enhanced auxiliary generator helps produce a stronger detector. On the predator side, we introduce a novel COD method, called Internal Coherence and Edge Guidance (ICEG), which introduces a camouflaged feature coherence module to excavate the internal coherence of camouflaged objects, striving to obtain more complete segmentation results. Additionally, ICEG proposes a novel edge-guided separated calibration module to remove false predictions to avoid obtaining ambiguous boundaries. Extensive experiments show that ICEG outperforms existing COD detectors and Camouflageator is flexible to improve various COD detectors, including ICEG, which brings state-of-the-art COD performance.
****
Yunchong Song,Siyuan Huang,Xinbing Wang,Chenghu Zhou,Zhouhan Lin
Graph Parsing Networks
Graph pooling compresses graph information into a compact representation. State-of-the-art graph pooling methods follow a hierarchical approach, which reduces the graph size step-by-step. These methods must balance memory efficiency with preserving node information, depending on whether they use node dropping or node clustering. Additionally, fixed pooling ratios or numbers of pooling layers are predefined for all graphs, which prevents personalized pooling structures from being captured for each individual graph. In this work, inspired by bottom-up grammar induction, we propose an efficient graph parsing algorithm to infer the pooling structure, which then drives graph pooling. The resulting Graph Parsing Network (GPN) adaptively learns personalized pooling structure for each individual graph. GPN benefits from the discrete assignments generated by the graph parsing

algorithm, allowing good memory efficiency while preserving node information int
act. Experimental results on standard benchmarks demonstrate that GPN outperform
s state-of-the-art graph pooling methods in graph classification tasks while bei
ng able to achieve competitive performance in node classification tasks. We also
 conduct a graph reconstruction task to show GPN's ability to preserve node info
rmation and measure both memory and time efficiency through relevant tests.
****
Ivan Butakov,Alexander Tolmachev,Sofia Malanchuk,Anna Neopryatnaya,Alexey Frolov
,Kirill Andreev
Information Bottleneck Analysis of Deep Neural Networks via Lossy Compression
The Information Bottleneck (IB) principle offers an information-theoretic framew
ork for analyzing the training process of deep neural networks (DNNs). Its essen
ce lies in tracking the dynamics of two mutual information (MI) values: between
the hidden layer output and the DNN input/target. According to the hypothesis pu
t forth by Shwartz-Ziv & Tishby (2017), the training process consists of two dis
tinct phases: fitting and compression. The latter phase is believed to account f
or the good generalization performance exhibited by DNNs. Due to the challenging
 nature of estimating MI between high-dimensional random vectors, this hypothesi
s was only partially verified for NNs of tiny sizes or specific types, such as q
uantized NNs. In this paper, we introduce a framework for conducting IB analysis
 of general NNs. Our approach leverages the stochastic NN method proposed by Gol
dfeld et al. (2019) and incorporates a compression step to overcome the obstacle
s associated with high dimensionality. In other words, we estimate the MI betwee
n the compressed representations of high-dimensional random vectors. The propose
d method is supported by both theoretical and practical justifications. Notably,
 we demonstrate the accuracy of our estimator through synthetic experiments feat
uring predefined MI values and comparison with MINE (Belghazi et al., 2018). Fin
ally, we perform IB analysis on a close-to-real-scale convolutional DNN, which r
eveals new features of the MI dynamics.
****
Ronghao Dang,Jiangyan Feng,Haodong Zhang,Chongjian GE,Lin Song,Lijun GONG,Chengj
u Liu,Qijun Chen,Feng Zhu,Rui Zhao,Yibing Song
InstructDET: Diversifying Referring Object Detection with Generalized Instructio
ns
We propose InstructDET, a data-centric method for referring object detection (RO
D) that localizes target objects based on user instructions. While deriving from
 referring expressions (REC), the instructions we leverage are greatly diversifi
ed to encompass common user intentions related to object detection. For one imag
e, we produce tremendous instructions that refer to every single object and diff
erent combinations of multiple objects. Each instruction and its corresponding o
bject bounding boxes (bbxs) constitute one training data pair. In order to encom
pass common detection expressions, we involve emerging vision-language model (VL
M) and large language model (LLM) to generate instructions guided by text prompt
s and object bbxs, as the generalizations of foundation models are effective to
produce human-like expressions (e.g., describing object property, category, and
relationship). We name our constructed dataset as InDET. It contains images, bbx
s and generalized instructions that are from foundation models. Our InDET is dev
eloped from existing REC datasets and object detection datasets, with the expand
ing potential that any image with object bbxs can be incorporated through using
our InstructDET method. By using our InDET dataset, we show that a conventional
ROD model surpasses existing methods on standard REC datasets and our InDET test
 set. Our data-centric method InstructDET, with automatic data expansion by leve
raging foundation models, directs a promising field that ROD can be greatly dive
rsified to execute common object detection instructions.
****
Mingxuan Li,Junzhe Zhang,Elias Bareinboim
Causally Aligned Curriculum Learning
A pervasive challenge in Reinforcement Learning (RL) is the ``curse of dimension
ality'' which is the exponential growth in the state-action space when optimizin
g a high-dimensional target task. The framework of curriculum learning trains th

e agent in a curriculum composed of a sequence of related and more manageable so urce tasks. The expectation is that when some optimal decision rules are shared across source tasks and the target task, the agent could more quickly pick up th e necessary skills to behave optimally in the environment, thus accelerating the learning process.
However, this critical assumption of invariant optimal decision rules does not n ecessarily hold in many practical applications, specifically when the underlying environment contains unobserved confounders. This paper studies the problem of curriculum RL through causal lenses. We derive a sufficient graphical condition characterizing causally aligned source tasks, i.e., the invariance of optimal de cision rules holds. We further develop an efficient algorithm to generate a caus ally aligned curriculum, provided with qualitative causal knowledge of the targe t environment. Finally, we validate our proposed methodology through experiments in confounded environments.
****
Soroush H. Zargarbashi,Aleksandar Bojchevski
Conformal Inductive Graph Neural Networks
Conformal prediction (CP) transforms any model's output into prediction sets gua ranteed to include (cover) the true label. CP requires exchangeability, a relaxa tion of the i.i.d. assumption, to obtain a valid distribution-free coverage guar antee. This makes it directly applicable to transductive node-classification. Ho wever, conventional CP cannot be applied in inductive settings due to the implic it shift in the (calibration) scores caused by message passing with the new node s. We fix this issue for both cases of node and edge-exchangeable graphs, recov ering the standard coverage guarantee without sacrificing statistical efficiency . We further prove that the guarantee holds independently of the prediction time , e.g. upon arrival of a new node/edge or at any subsequent moment.
****
Xudong Shen,Chao Du,Tianyu Pang,Min Lin,Yongkang Wong,Mohan Kankanhalli
Finetuning Text-to-Image Diffusion Models for Fairness
The rapid adoption of text-to-image diffusion models in society underscores an u rgent need to address their biases. Without interventions, these biases could pr opagate a skewed worldview and restrict opportunities for minority groups. In th is work, we frame fairness as a distributional alignment problem. Our solution c onsists of two main technical contributions: (1) a distributional alignment loss that steers specific characteristics of the generated images towards a user-def ined target distribution, and (2) adjusted direct finetuning of diffusion model' s sampling process (adjusted DFT), which leverages an adjusted gradient to direc tly optimize losses defined on the generated images. Empirically, our method mar kedly reduces gender, racial, and their intersectional biases for occupational p rompts. Gender bias is significantly reduced even when finetuning just five soft tokens. Crucially, our method supports diverse perspectives of fairness beyond absolute equality, which is demonstrated by controlling age to a $75\\%$ young a nd $25\\%$ old distribution while simultaneously debiasing gender and race. Fina lly, our method is scalable: it can debias multiple concepts at once by simply i ncluding these prompts in the finetuning data. We share code and various fair di ffusion model adaptors at https://sail-sg.github.io/finetune-fair-diffusion/.
****
Chenhui Deng,Zichao Yue,Zhiru Zhang
Polynormer: Polynomial-Expressive Graph Transformer in Linear Time
Graph transformers (GTs) have emerged as a promising architecture that is theore tically more expressive than message-passing graph neural networks (GNNs). Howev er, typical GT models have at least quadratic complexity and thus cannot scale t o large graphs. While there are several linear GTs recently proposed, they still lag behind GNN counterparts on several popular graph datasets, which poses a cr itical concern on their practical expressivity. To balance the trade-off between expressivity and scalability of GTs, we propose Polynormer, a polynomial-expres sive GT model with linear complexity. Polynormer is built upon a novel base mode l that learns a high-degree polynomial on input features. To enable the base mod el permutation equivariant, we integrate it with graph topology and node feature

s separately, resulting in local and global equivariant attention models. Conseq uently, Polynormer adopts a linear local-to-global attention scheme to learn hig h-degree equivariant polynomials whose coefficients are controlled by attention scores. Polynormer has been evaluated on 13 homophilic and heterophilic datase ts, including large graphs with millions of nodes. Our extensive experiment resu lts show that Polynormer outperforms state-of-the-art GNN and GT baselines on mo st datasets, even without the use of nonlinear activation functions. Source code of Polynormer is freely available at: [github.com/cornell-zhang/Polynormer](htt ps://github.com/cornell-zhang/Polynormer).
****
Weiqiang He,Hendrik Fichtenberger,Pan Peng
A Differentially Private Clustering Algorithm for Well-Clustered Graphs
We study differentially private (DP) algorithms for recovering clusters in well-clustered graphs, which are graphs whose vertex set can be partitioned into a sm all number of sets, each inducing a subgraph of high inner conductance and small outer conductance. Such graphs have widespread application as a benchmark in th e theoretical analysis of spectral clustering.
We provide an efficient (ϵ,δ)-DP algorithm tailored specifically for such graphs. Our algorithm draws inspiration from the recent work of Chen e t al., who developed DP algorithms for recovery of stochastic block models in ca ses where the graph comprises exactly two nearly-balanced clusters. Our algorith m works for well-clustered graphs with k nearly-balanced clusters, and the mis classification ratio almost matches the one of the best-known non-private algori thms. We conduct experimental evaluations on datasets with known ground truth cl usters to substantiate the prowess of our algorithm. We also show that any (pure) ϵ-DP algorithm would result in substantial error.
****
Peiyan Hu,Yue Wang,Zhi-Ming Ma
Better Neural PDE Solvers Through Data-Free Mesh Movers
Recently, neural networks have been extensively employed to solve partial differ ential equations (PDEs) in physical system modeling. While major studies focus o n learning system evolution on predefined static mesh discretizations, some meth ods utilize reinforcement learning or supervised learning techniques to create a daptive and dynamic meshes, due to the dynamic nature of these systems. However, these approaches face two primary challenges: (1) the need for expensive optima l mesh data, and (2) the change of the solution space's degree of freedom and to pology during mesh refinement. To address these challenges, this paper proposes a neural PDE solver with a neural mesh adapter. To begin with, we introduce a no vel data-free neural mesh adaptor, called Data-free Mesh Mover (DMM), with two m ain innovations. Firstly, it is an operator that maps the solution to adaptive m eshes and is trained using the Monge-Ampère equation without optimal mesh data. Secondly, it dynamically changes the mesh by moving existing nodes rather than a dding or deleting nodes and edges. Theoretical analysis shows that meshes genera ted by DMM have the lowest interpolation error bound. Based on DMM, to efficient ly and accurately model dynamic systems, we develop a moving mesh based neural P DE solver (MM-PDE) that embeds the moving mesh with a two-branch architecture an d a learnable interpolation framework to preserve information within the data. E mpirical experiments demonstrate that our method generates suitable meshes and c onsiderably enhances accuracy when modeling widely considered PDE systems. The c ode can be found at: https://github.com/Peiyannn/MM-PDE.git.
****
Antoine Gonon,Nicolas Brisebarre,Elisa Riccietti,Rémi Gribonval
A path-norm toolkit for modern networks: consequences, promises and challenges
This work introduces the first toolkit around path-norms that fully encompasses general DAG ReLU networks with biases, skip connections and any operation based on the extraction of order statistics: max pooling, GroupSort etc.
This toolkit notably allows us to establish generalization bounds for modern neu ral networks that are not only the most widely applicable path-norm based ones, but also recover or beat the sharpest known bounds of this type.
These extended path-norms further enjoy the usual benefits of path-norms: ease o

f computation, invariance under the symmetries of the network, and improved sharpness on layered fully-connected networks compared to the product of operator norms, another complexity measure most commonly used.

The versatility of the toolkit and its ease of implementation allow us to challenge the concrete promises of path-norm-based generalization bounds, by numerically evaluating the sharpest known bounds for ResNets on ImageNet.
****
Xuangeng Chu,Yu Li,Ailing Zeng,Tianyu Yang,Lijian Lin,Yunfei Liu,Tatsuya Harada
GPAvatar: Generalizable and Precise Head Avatar from Image(s)
Head avatar reconstruction, crucial for applications in virtual reality, online meetings, gaming, and film industries, has garnered substantial attention within the computer vision community. The fundamental objective of this field is to faithfully recreate the head avatar and precisely control expressions and postures. Existing methods, categorized into 2D-based warping, mesh-based, and neural rendering approaches, present challenges in maintaining multi-view consistency, incorporating non-facial information, and generalizing to new identities. In this paper, we propose a framework named GPAvatar that reconstructs 3D head avatars from one or several images in a single forward pass. The key idea of this work is to introduce a dynamic point-based expression field driven by a point cloud to precisely and effectively capture expressions. Furthermore, we use a Multi Tri-planes Attention (MTA) fusion module in tri-planes canonical field to leverage information from multiple input images. The proposed method achieves faithful identity reconstruction, precise expression control, and multi-view consistency, demonstrating promising results for free-viewpoint rendering and novel view synthesis.
****
Xichen Pan,Li Dong,Shaohan Huang,Zhiliang Peng,Wenhu Chen,Furu Wei
Kosmos-G: Generating Images in Context with Multimodal Large Language Models
Recent advancements in subject-driven image generation have made significant strides. However, current methods still fall short in diverse application scenarios, as they require test-time tuning and cannot accept interleaved multi-image and text input. These limitations keep them far from the ultimate goal of "image as a foreign language in image generation." This paper presents Kosmos-G, a model that leverages the advanced multimodal perception capabilities of Multimodal Large Language Models (MLLMs) to tackle the aforementioned challenge. Our approach aligns the output space of MLLM with CLIP using the textual modality as an anchor and performs compositional instruction tuning on curated data. Kosmos-G demonstrates an impressive capability of zero-shot subject-driven generation with interleaved multi-image and text input. Notably, the score distillation instruction tuning requires no modifications to the image decoder. This allows for a seamless substitution of CLIP and effortless integration with a myriad of U-Net techniques ranging from fine-grained controls to personalized image decoder variants. We posit Kosmos-G as an initial attempt towards the goal of "image as a foreign language in image generation."
****
Kai Yi,Nidham Gazagnadou,Peter Richtárik,Lingjuan Lyu
FedP3: Federated Personalized and Privacy-friendly Network Pruning under Model Heterogeneity
The interest in federated learning has surged in recent research due to its unique ability to train a global model using privacy-secured information held locally on each client. This paper pays particular attention to the issue of client-side model heterogeneity, a pervasive challenge in the practical implementation of FL that escalates its complexity. Assuming a scenario where each client possesses varied memory storage, processing capabilities and network bandwidth - a phenomenon referred to as system heterogeneity - there is a pressing need to customize a unique model for each client. In response to this, we present an effective and adaptable federated framework FedP3, representing Federated Personalized and Privacy-friendly network Pruning, tailored for model heterogeneity scenarios. Our proposed methodology can incorporate and adapt well-established techniques t

o its specific instances. We offer a theoretical interpretation of FedP3 and its locally differential-private variant, DP-FedP3, and theoretically validate their efficiencies.
****

Rujie Wu,Xiaojian Ma,Zhenliang Zhang,Wei Wang,Qing Li,Song-Chun Zhu,Yizhou Wang
Bongard-OpenWorld: Few-Shot Reasoning for Free-form Visual Concepts in the Real World
We introduce Bongard-OpenWorld, a new benchmark for evaluating real-world few-shot reasoning for machine vision. It originates from the classical Bongard Problems (BPs): Given two sets of images (positive and negative), the model needs to identify the set that query images belong to by inducing the visual concepts, which is exclusively depicted by images from the positive set. Our benchmark inherits the few-shot concept induction of the original BPs while adding the two novel layers of challenge: 1) open-world free-form concepts, as the visual concepts in Bongard-OpenWorld are unique compositions of terms from an open vocabulary, ranging from object categories to abstract visual attributes and commonsense factual knowledge; 2) real-world images, as opposed to the synthetic diagrams used by many counterparts. In our exploration, Bongard-OpenWorld already imposes a significant challenge to current few-shot reasoning algorithms. We further investigate to which extent the recently introduced Large Language Models (LLMs) and Vision-Language Models (VLMs) can solve our task, by directly probing VLMs, and combining VLMs and LLMs in an interactive reasoning scheme. We even conceived a neuro-symbolic reasoning approach that reconciles LLMs & VLMs with logical reasoning to emulate the human problem-solving process for Bongard Problems. However, none of these approaches manage to close the human-machine gap, as the best learner achieves 64% accuracy while human participants easily reach 91%. We hope Bongard-OpenWorld can help us better understand the limitations of current visual intelligence and facilitate future research on visual agents with stronger few-shot visual reasoning capabilities.
****

Xiangyu Qi,Yi Zeng,Tinghao Xie,Pin-Yu Chen,Ruoxi Jia,Prateek Mittal,Peter Henderson
Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!
Optimizing large language models (LLMs) for downstream use cases often involves the customization of pre-trained LLMs through further fine-tuning. Meta's open-source release of Llama models and OpenAI's APIs for fine-tuning GPT-3.5 Turbo on customized datasets accelerate this trend. But, what are the safety costs associated with such customized fine-tuning? While existing safety alignment techniques restrict harmful behaviors of LLMs at inference time, they do not cover safety risks when fine-tuning privileges are extended to end-users. Our red teaming studies find that the safety alignment of LLMs can be compromised by fine-tuning with only a few adversarially designed training examples. For instance, we jailbreak GPT-3.5 Turbo's safety guardrails by fine-tuning it on only 10 such examples at a cost of less than $0.20 via OpenAI's APIs, making the model responsive to nearly any harmful instructions. Disconcertingly, our research also reveals that, even without malicious intent, simply fine-tuning with benign and commonly used datasets can also inadvertently degrade the safety alignment of LLMs, though to a lesser extent. These findings suggest that fine-tuning aligned LLMs introduces new safety risks that current safety infrastructures fall short of addressing --- even if a model's initial safety alignment is impeccable, how can it be maintained after customized fine-tuning? We outline and critically analyze potential mitigations and advocate for further research efforts toward reinforcing safety protocols for the customized fine-tuning of aligned LLMs. (This paper contains red-teaming data and model-generated content that can be offensive in nature.)
****

Akari Asai,Zeqiu Wu,Yizhong Wang,Avirup Sil,Hannaneh Hajishirzi
Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection
Despite their remarkable capabilities, large language models (LLMs) often produc

e responses containing factual inaccuracies due to their sole reliance on the pa
rametric knowledge they encapsulate. Retrieval-Augmented Generation (RAG), an ad
 hoc approach that augments LMs with retrieval of relevant knowledge, decreases
such issues. However, indiscriminately retrieving and incorporating a fixed numb
er of retrieved passages, regardless of whether retrieval is necessary, or passa
ges are relevant, diminishes LM versatility or can lead to unhelpful response ge
neration. We introduce a new framework called **Self-Reflective Retrieval-Augmen
ted Generation (Self-RAG)** that enhances an LM's quality and factuality through
 retrieval and self-reflection.
Our framework trains a single arbitrary LM that adaptively retrieves passages on
-demand, and generates and reflects on retrieved passages and its generations us
ing special tokens, called {\it reflection} tokens. Generating reflection tokens
 makes the LM controllable during the inference phase, enabling it to tailor its
 behavior to diverse task requirements.
Experiments show that Self-RAG (7B and 13B parameters) significantly outperforms
 state-of-the-art LLMs and retrieval-augmented models on a diverse set of tasks.

Specifically, Self-RAG outperforms ChatGPT and retrieval-augmented Llama2-chat o
n Open-domain QA, reasoning, and fact verification tasks, and it shows significa
nt gains in improving factuality and citation accuracy for long-form generations
 relative to these models. Our code and trained models are available at https://
selfrag.github.io/
****
Murtaza Dalal,Tarun Chiruvolu,Devendra Singh Chaplot,Ruslan Salakhutdinov
Plan-Seq-Learn: Language Model Guided RL for Solving Long Horizon Robotics Tasks
Large Language Models (LLMs) are highly capable of performing planning for long-
horizon robotics tasks, yet existing methods require access to a pre-defined ski
ll library (*e.g.* picking, placing, pulling, pushing, navigating). However, LLM
 planning does not address how to design or learn those behaviors, which remains
 challenging particularly in long-horizon settings. Furthermore, for many tasks
of interest, the robot needs to be able to adjust its behavior in a fine-grained
 manner, requiring the agent to be capable of modifying *low-level* control acti
ons. Can we instead use the internet-scale knowledge from LLMs for high-level po
licies, guiding reinforcement learning (RL) policies to efficiently solve roboti
c control tasks online without requiring a pre-determined set of skills? In this
 paper, we propose **Plan-Seq-Learn** (PSL): a modular approach that uses motion
 planning to bridge the gap between abstract language and learned low-level cont
rol for solving long-horizon robotics tasks from scratch. We demonstrate that PS
L is capable of solving 20+ challenging single and multi-stage robotics tasks on
 four benchmarks at success rates of over 80% from raw visual input, out-perform
ing language-based, classical, and end-to-end approaches. Video results and code
 at https://planseqlearn.github.io/
****
Lu Yu,Avetik Karagulyan,Arnak S. Dalalyan
Langevin Monte Carlo for strongly log-concave distributions: Randomized midpoint
 revisited
We revisit the problem of sampling from a target distribution that has a smooth
strongly log-concave density everywhere in $\mathbb{R}^p$. In this context, if n
o additional density information is available, the randomized midpoint discretiz
ation for the kinetic Langevin diffusion is known to be the most scalable method
 in high dimensions with large condition numbers. Our main result is a nonasympt
otic and easy to compute upper bound on the W_2-error of this method. To provi
de a more thorough explanation of our method for establishing the computable upp
er bound, we conduct an analysis of the midpoint discretization for the vanilla
Langevin process. This analysis helps to clarify the underlying principles and p
rovides valuable insights that we use to establish an improved upper bound for t
he kinetic Langevin process with the midpoint discretization. Furthermore, by ap
plying these techniques we establish new guarantees for the kinetic Langevin pro
cess with Euler discretization, which have a better dependence on the condition
number than existing upper bounds

****
Siqi Zhang,Sayantan Choudhury,Sebastian U Stich,Nicolas Loizou
Communication-Efficient Gradient Descent-Accent Methods for Distributed Variatio nal Inequalities: Unified Analysis and Local Updates
Distributed and federated learning algorithms and techniques associated primaril y with minimization problems. However, with the increase of minimax optimization and variational inequality problems in machine learning, the necessity of desig ning efficient distributed/federated learning approaches for these problems is b ecoming more apparent. In this paper, we provide a unified convergence analysis of communication-efficient local training methods for distributed variational in equality problems (VIPs). Our approach is based on a general key assumption on t he stochastic estimates that allows us to propose and analyze several novel loca l training algorithms under a single framework for solving a class of structured non-monotone VIPs. We present the first local gradient descent-accent algorithm s with provable improved communication complexity for solving distributed variat ional inequalities on heterogeneous data. The general algorithmic framework reco vers state-of-the-art algorithms and their sharp convergence guarantees when the setting is specialized to minimization or minimax optimization problems. Finall y, we demonstrate the strong performance of the proposed algorithms compared to state-of-the-art methods when solving federated minimax optimization problems.
****
Maxime Wabartha,Joelle Pineau
Piecewise Linear Parametrization of Policies: Towards Interpretable Deep Reinfor cement Learning
Learning inherently interpretable policies is a central challenge in the path to developing autonomous agents that humans can trust. Linear policies can justify their decisions while interacting in a dynamic environment, but their reduced e xpressivity prevents them from solving hard tasks. Instead, we argue for the use of piecewise-linear policies. We carefully study to what extent they can retain the interpretable properties of linear policies while reaching competitive perf ormance with neural baselines. In particular, we propose the HyperCombinator (HC), a piecewise-linear neural architecture expressing a policy with a controllabl y small number of sub-policies. Each sub-policy is linear with respect to interp retable features, shedding light on the decision process of the agent without re quiring an additional explanation model. We evaluate HC policies in control and navigation experiments, visualize the improved interpretability of the agent and highlight its trade-off with performance. Moreover, we validate that the restri cted model class that the HyperCombinator belongs to is compatible with the algo rithmic constraints of various reinforcement learning algorithms.
****
Yiheng Xu,Hongjin SU,Chen Xing,Boyu Mi,Qian Liu,Weijia Shi,Binyuan Hui,Fan Zhou, Yitao Liu,Tianbao Xie,Zhoujun Cheng,Siheng Zhao,Lingpeng Kong,Bailin Wang,Caimin g Xiong,Tao Yu
Lemur: Harmonizing Natural Language and Code for Language Agents
We introduce Lemur and Lemur-Chat, openly accessible language models optimized for both natural language and coding capabilities to serve as the backbone of versatile language agents. The evolution from language chat models to functional language agents demands that models not only master human interaction ,
reasoning, and planning but also ensure grounding in the relevant environments. This calls for a harmonious blend of language and coding capabilities in the models. Lemur and Lemur-Chat are proposed to address this necessity, demonstrating balanced proficiencies in both domains, unlike existing open-source models that tend to specialize in either. Through meticulous pretrai ning
using a code-intensive corpus and instruction fine-tuning on text and code data, our models achieve state-of-the-art averaged performance across diverse text and coding benchmarks. Comprehensive experiments demonstrate Lemur’s superiority over existing open-source models and its proficiency across various agent tasks involving human communication, tool usage, and interaction under

fully- and partially- observable environments. The harmonization between natural and programming languages enables Lemur-Chat to significantly narrow the
gap with proprietary models on agent abilities, providing key insights into deve loping
advanced open-source agents adept at reasoning, planning, and operating
seamlessly across environments. Our model and code have been open-sourced at
https://github.com/OpenLemur/Lemur.
****
Zhiquan Tan,Yifan Zhang,Jingqin Yang,Yang Yuan
Contrastive Learning is Spectral Clustering on Similarity Graph
Contrastive learning is a powerful self-supervised learning method, but we have a limited theoretical understanding of how it works and why it works. In this pa per, we prove that contrastive learning with the standard InfoNCE loss is equiva lent to spectral clustering on the similarity graph. Using this equivalence as t he building block, we extend our analysis to the CLIP model and rigorously chara cterize how similar multi-modal objects are embedded together. Motivated by our theoretical insights, we introduce the Kernel-InfoNCE loss, incorporating mixtur es of kernel functions that outperform the standard Gaussian kernel on several v ision datasets.
****
Weihao Tan,Wentao Zhang,Shanqi Liu,Longtao Zheng,Xinrun Wang,Bo An
True Knowledge Comes from Practice: Aligning Large Language Models with Embodied Environments via Reinforcement Learning
Despite the impressive performance across numerous tasks, large language models (LLMs) often fail in solving simple decision-making tasks due to the misalignmen t of the knowledge in LLMs with environments. On the contrary, reinforcement lea rning (RL) agents learn policies from scratch, which makes them always align wit h environments but difficult to incorporate prior knowledge for efficient explor ations. To narrow the gap, we propose TWOSOME, a novel general online framework that deploys LLMs as decision-making agents to efficiently interact and align wi th embodied environments via RL without requiring any prepared datasets or prior knowledge of the environments. Firstly, we query the joint probabilities of eac h valid action with LLMs to form behavior policies. Then, to enhance the stabili ty and robustness of the policies, we propose two normalization methods and summ arize four prompt design principles. Finally, we design a novel parameter-effici ent training architecture where the actor and critic share one frozen LLM equipp ed with low-rank adapters (LoRA) updated by PPO. We conduct extensive experiment s to evaluate TWOSOME. i) TWOSOME exhibits significantly better sample efficienc y and performance compared to the conventional RL method, PPO, and prompt tuning method, SayCan, in both classical decision-making environment, Overcooked, and simulated household environment, VirtualHome. ii) Benefiting from LLMs' open-voc abulary feature, TWOSOME shows superior generalization ability to unseen tasks. iii) Under our framework, there is no significant loss of the LLMs' original abi lity during online PPO finetuning.
****
Haoying Li,Jixin Zhao,Shangchen Zhou,Huajun Feng,Chongyi Li,Chen Change Loy
Adaptive Window Pruning for Efficient Local Motion Deblurring
Local motion blur commonly occurs in real-world photography due to the mixing be tween moving objects and stationary backgrounds during exposure. Existing image deblurring methods predominantly focus on global deblurring, inadvertently affec ting the sharpness of backgrounds in locally blurred images and wasting unnecess ary computation on sharp pixels, especially for high-resolution images.
This paper aims to adaptively and efficiently restore high-resolution locally bl urred images. We propose a local motion deblurring vision Transformer (LMD-ViT) built on adaptive window pruning Transformer blocks (AdaWPT). To focus deblurrin g on local regions and reduce computation, AdaWPT prunes unnecessary windows, on ly allowing the active windows to be involved in the deblurring processes. The p runing operation relies on the blurriness confidence predicted by a confidence p redictor that is trained end-to-end using a reconstruction loss with Gumbel-Soft max re-parameterization and a pruning loss guided by annotated blur masks. Our m

ethod removes local motion blur effectively without distorting sharp regions, demonstrated by its exceptional perceptual and quantitative improvements (+0.28dB) compared to state-of-the-art methods. In addition, our approach substantially reduces FLOPs by 66% and achieves more than a twofold increase in inference speed compared to Transformer-based deblurring methods. We will make our code and annotated blur masks publicly available.
****
Zhikai Chen,Haitao Mao,Hongzhi Wen,Haoyu Han,Wei Jin,Haiyang Zhang,Hui Liu,Jiliang Tang
Label-free Node Classification on Graphs with Large Language Models (LLMs)
In recent years, there have been remarkable advancements in node classification achieved by Graph Neural Networks (GNNs). However, they necessitate abundant high-quality labels to ensure promising performance. In contrast, Large Language Models (LLMs) exhibit impressive zero-shot proficiency on text-attributed graphs. Yet, they face challenges in efficiently processing structural data and suffer from high inference costs. In light of these observations, this work introduces a label-free node classification on graphs with LLMs pipeline, LLM-GNN. It amalgamates the strengths of both GNNs and LLMs while mitigating their limitations. Specifically, LLMs are leveraged to annotate a small portion of nodes and then GNNs are trained on LLMs' annotations to make predictions for the remaining large portion of nodes. The implementation of LLM-GNN faces a unique challenge: how can we actively select nodes for LLMs to annotate and consequently enhance the GNN training? How can we leverage LLMs to obtain annotations of high quality, representativeness, and diversity, thereby enhancing GNN performance with less cost? To tackle this challenge, we develop an annotation quality heuristic and leverage the confidence scores derived from LLMs to advanced node selection. Comprehensive experimental results validate the effectiveness of LLM-GNN. In particular, LLM-GNN can achieve an accuracy of 74.9\% on a vast-scale dataset \products with a cost less than 1 dollar.
****
Chuyu Zhang,Hui Ren,Xuming He
P^{2}OT: Progressive Partial Optimal Transport for Deep Imbalanced Clustering
Deep clustering, which learns representation and semantic clustering without labels information, poses a great challenge for deep learning-based approaches. Despite significant progress in recent years, most existing methods focus on uniformly distributed datasets, significantly limiting the practical applicability of their methods. In this paper, we first introduce a more practical problem setting named deep imbalanced clustering, where the underlying classes exhibit an imbalance distribution. To tackle this problem, we propose a novel pseudo-labeling-based learning framework. Our framework formulates pseudo-label generation as a progressive partial optimal transport problem, which progressively transports each sample to imbalanced clusters under prior distribution constraints, thus generating imbalance-aware pseudo-labels and learning from high-confident samples.
In addition, we transform the initial formulation into an unbalanced optimal transport problem with augmented constraints, which can be solved efficiently by a fast matrix scaling algorithm. Experiments on various datasets, including a human-curated long-tailed CIFAR100, challenging ImageNet-R, and large-scale subsets of fine-grained iNaturalist2018 datasets, demonstrate the superiority of our method.
****
Qian qian Dong,Zhiying Huang,Qiao Tian,Chen Xu,Tom Ko,yunlong zhao,Siyuan Feng,Tang Li,Kexin Wang,Xuxin Cheng,Fengpeng Yue,Ye Bai,Xi Chen,Lu Lu,Zejun MA,Yuping Wang,Mingxuan Wang,Yuxuan Wang
PolyVoice: Language Models for Speech to Speech Translation
With the huge success of GPT models in natural language processing, there is a growing interest in applying language modeling approaches to speech tasks.
Currently, the dominant architecture in speech-to-speech translation (S2ST) remains the encoder-decoder paradigm, creating a need to investigate the impact of language modeling approaches in this area.
In this study, we introduce PolyVoice, a language model-based framework designed

for S2ST systems. Our framework comprises three decoder-only language models: a translation language model, a duration language model, and a speech synthesis language model.
These language models employ different types of prompts to extract learned information effectively. By utilizing unsupervised semantic units, our framework can transfer semantic information across these models, making it applicable even to unwritten languages.
We evaluate our system on Chinese $\rightarrow$ English and English $\rightarrow$ Spanish language pairs. Experimental results demonstrate that \method outperforms the state-of-the-art encoder-decoder model, producing voice-cloned speech with high translation and audio quality.
Speech samples are available at https://polyvoice.github.io.
****
Ming Yang Zhou,Zichao Yan,Elliot Layne,Nikolay Malkin,Dinghuai Zhang,Moksh Jain,Mathieu Blanchette,Yoshua Bengio
PhyloGFN: Phylogenetic inference with generative flow networks
Phylogenetics is a branch of computational biology that studies the evolutionary relationships among biological entities. Its long history and numerous applications notwithstanding, inference of phylogenetic trees from sequence data remains challenging: the high complexity of tree space poses a significant obstacle for the current combinatorial and probabilistic techniques. In this paper, we adopt the framework of generative flow networks (GFlowNets) to tackle two core problems in phylogenetics: parsimony-based and Bayesian phylogenetic inference. Because GFlowNets are well-suited for sampling complex combinatorial structures, they are a natural choice for exploring and sampling from the multimodal posterior distribution over tree topologies and evolutionary distances. We demonstrate that our amortized posterior sampler, PhyloGFN, produces diverse and high-quality evolutionary hypotheses on real benchmark datasets. PhyloGFN is competitive with prior works in marginal likelihood estimation and achieves a closer fit to the target distribution than state-of-the-art variational inference methods.
****
Shanqi Liu,Dong Xing,Pengjie Gu,Xinrun Wang,Bo An,Yong Liu
Solving Homogeneous and Heterogeneous Cooperative Tasks with Greedy Sequential Execution
Cooperative multi-agent reinforcement learning (MARL) is extensively used for solving complex cooperative tasks, and value decomposition methods are a prevalent approach for this domain. However, these methods have not been successful in addressing both homogeneous and heterogeneous tasks simultaneously which is a crucial aspect for the practical application of cooperative agents.
On one hand, value decomposition methods demonstrate superior performance in homogeneous tasks. Nevertheless, they tend to produce agents with similar policies, which is unsuitable for heterogeneous tasks. On the other hand, solutions based on personalized observation or assigned roles are well-suited for heterogeneous tasks. However, they often lead to a trade-off situation where the agent's performance in homogeneous scenarios is negatively affected due to the aggregation of distinct policies. An alternative approach is to adopt sequential execution policies, which offer a flexible form for learning both types of tasks. However, learning sequential execution policies poses challenges in terms of credit assignment, and the limited information about subsequently executed agents can lead to sub-optimal solutions, which is known as the relative over-generalization problem. To tackle these issues, this paper proposes Greedy Sequential Execution (GSE) as a solution to learn the optimal policy that covers both scenarios. In the proposed GSE framework, we introduce an individual utility function into the framework of value decomposition to consider the complex interactions between agents.
This function is capable of representing both the homogeneous and heterogeneous optimal policies. Furthermore, we utilize greedy marginal contribution calculated by the utility function as the credit value of the sequential execution policy to address the credit assignment and relative over-generalization problem. We evaluated GSE in both homogeneous and heterogeneous scenarios. The results demons

trate that GSE achieves significant improvement in performance across multiple domains, especially in scenarios involving both homogeneous and heterogeneous tasks.
****
Weiming Zhuang,Lingjuan Lyu
FedWon: Triumphing Multi-domain Federated Learning Without Normalization
Federated learning (FL) enhances data privacy with collaborative in-situ training on decentralized clients. Nevertheless, FL encounters challenges due to non-independent and identically distributed (non-i.i.d) data, leading to potential performance degradation and hindered convergence. While prior studies predominantly addressed the issue of skewed label distribution, our research addresses a crucial yet frequently overlooked problem known as multi-domain FL. In this scenario, clients' data originate from diverse domains with distinct feature distributions, instead of label distributions. To address the multi-domain problem in FL, we propose a novel method called Federated Learning Without Normalizations (FedWon). FedWon draws inspiration from the observation that batch normalization (BN) faces challenges in effectively modeling the statistics of multiple domains, while existing normalization techniques possess their own limitations. In order to address these issues, FedWon eliminates the normalization layers in FL and reparameterizes convolution layers with scaled weight standardization. Through extensive experimentation on five datasets and five models, our comprehensive experimental results demonstrate that FedWon surpasses both FedAvg and the current state-of-the-art method (FedBN) across all experimental setups, achieving notable accuracy improvements of more than 10% in certain domains. Furthermore, FedWon is versatile for both cross-silo and cross-device FL, exhibiting robust domain generalization capability, showcasing strong performance even with a batch size as small as 1, thereby catering to resource-constrained devices. Additionally, FedWon can also effectively tackle the challenge of skewed label distribution.
****
Giorgio Mariani,Irene Tallini,Emilian Postolache,Michele Mancusi,Luca Cosmo,Emanuele Rodolà
Multi-Source Diffusion Models for Simultaneous Music Generation and Separation
In this work, we define a diffusion-based generative model capable of both music generation and source separation by learning the score of the joint probability density of sources sharing a context. Alongside the classic total inference tasks (i.e., generating a mixture, separating the sources), we also introduce and experiment on the partial generation task of source imputation, where we generate a subset of the sources given the others (e.g., play a piano track that goes well with the drums). Additionally, we introduce a novel inference method for the separation task based on Dirac likelihood functions. We train our model on Slakh2100, a standard dataset for musical source separation, provide qualitative results in the generation settings, and showcase competitive quantitative results in the source separation setting. Our method is the first example of a single model that can handle both generation and separation tasks, thus representing a step toward general audio models.
****
Yinbin Han,Meisam Razaviyayn,Renyuan Xu
Neural Network-Based Score Estimation in Diffusion Models: Optimization and Generalization
Diffusion models have emerged as a powerful tool rivaling GANs in generating high-quality samples with improved fidelity, flexibility, and robustness. A key component of these models is to learn the score function through score matching. Despite empirical success on various tasks, it remains unclear whether gradient-based algorithms can learn the score function with a provable accuracy. As a first step toward answering this question, this paper establishes a mathematical framework for analyzing score estimation using neural networks trained by gradient descent. Our analysis covers both the optimization and the generalization aspects of the learning procedure. In particular, we propose a parametric form to formulate the denoising score-matching problem as a regression with noisy labels. Compared to the standard supervised learning setup, the score-matching problem int

roduces distinct challenges, including unbounded input, vector-valued output, and an additional time variable, preventing existing techniques from being applied directly. In this paper, we show that with proper designs, the evolution of neural networks during training can be accurately modeled by a series of kernel regression tasks. Furthermore, by applying an early-stopping rule for gradient descent and leveraging recent developments in neural tangent kernels, we establish the first generalization error (sample complexity) bounds for learning the score function with neural networks, despite the presence of noise in the observations. Our analysis is grounded in a novel parametric form of the neural network and an innovative connection between score matching and regression analysis, facilitating the application of advanced statistical and optimization techniques.
****
Federico Errica,Mathias Niepert
Tractable Probabilistic Graph Representation Learning with Graph-Induced Sum-Product Networks
We introduce Graph-Induced Sum-Product Networks (GSPNs), a new probabilistic framework for graph representation learning that can tractably answer probabilistic queries. Inspired by the computational trees induced by vertices in the context of message-passing neural networks, we build hierarchies of sum-product networks (SPNs) where the parameters of a parent SPN are learnable transformations of the a-posterior mixing probabilities of its children's sum units. Due to weight sharing and the tree-shaped computation graphs of GSPNs, we obtain the efficiency and efficacy of deep graph networks with the additional advantages of a probabilistic model. We show the model's competitiveness on scarce supervision scenarios, under missing data, and for graph classification in comparison to popular neural models. We complement the experiments with qualitative analyses on hyper-parameters and the model's ability to answer probabilistic queries.
****
Luotian Yuan,Yemin Yu,Ying Wei,Yongwei Wang,Zhihua Wang,Fei Wu
Active Retrosynthetic Planning Aware of Route Quality
Retrosynthetic planning is a sequential decision-making process of identifying synthetic routes from the available building block materials to reach a desired target molecule.
Though existing planning approaches show promisingly high solving rates and low costs, the trivial route cost evaluation via pre-trained forward reaction prediction models certainly falls short of real-world chemical practice.
An alternative option is to annotate the actual cost of a route, such as yield, through chemical experiments or input from chemists, while
this often leads to substantial query costs.
In order to strike the balance between query costs and route quality evaluation, we propose an Active Retrosynthetic Planning (ARP) framework that remains compatible with the established retrosynthetic planners.
On one hand, the proposed ARP trains an actor that decides whether to query the cost of a reaction; on the other hand, it resorts to a critic to estimate the value of a molecule with its preceding reaction cost as input.
Those molecules with low reaction costs are preferred to expand first.
We apply our framework to different existing approaches on both the benchmark and an expert dataset and demonstrate that it outperforms the existing state-of-the-art approach by 6.2\% in route quality while reducing the query cost by 12.8\%.
In addition,
ARP consistently plans
high-quality routes with either abundant or sparse annotations.
****
Ling Yang,Ye Tian,Minkai Xu,Zhongyi Liu,Shenda Hong,Wei Qu,Wentao Zhang,Bin CUI,Muhan Zhang,Jure Leskovec
VQGraph: Rethinking Graph Representation Space for Bridging GNNs and MLPs
GNN-to-MLP distillation aims to utilize knowledge distillation (KD) to learn computationally-efficient multi-layer perceptron (student MLP) on graph data by mimicking the output representations of teacher GNN. Existing methods mainly make t

he MLP to mimic the GNN predictions over a few class labels. However, the class space may not be expressive enough for covering numerous diverse local graph str uctures, thus limiting the performance of knowledge transfer from GNN to MLP. To address this issue, we propose to learn a new powerful graph representation spa ce by directly labeling nodes' diverse local structures for GNN-to-MLP distillat ion. Specifically, we propose a variant of VQ-VAE to learn a structure-aware tok enizer on graph data that can encode each node's local substructure as a discret e code. The discrete codes constitute a codebook as a new graph representation s pace that is able to identify different local graph structures of nodes with the corresponding code indices. Then, based on the learned codebook, we propose a n ew distillation target, namely soft code assignments, to directly transfer the s tructural knowledge of each node from GNN to MLP. The resulting framework VQGrap h achieves new state-of-the-art performance on GNN-to-MLP distillation in both t ransductive and inductive settings across seven graph datasets. We show that VQG raph with better performance infers faster than GNNs by 828×, and also achieves accuracy improvement over GNNs and stand-alone MLPs by 3.90% and 28.05% on avera ge, respectively. Our code is available at https://github.com/YangLing0818/VQGra ph
****
Dong Bok Lee,Seanie Lee,Joonho Ko,Kenji Kawaguchi,Juho Lee,Sung Ju Hwang
Self-Supervised Dataset Distillation for Transfer Learning
Dataset distillation methods have achieved remarkable success in distilling a la rge dataset into a small set of representative samples. However, they are not de signed to produce a distilled dataset that can be effectively used for facilitat ing self-supervised pre-training. To this end, we propose a novel problem of dis tilling an unlabeled dataset into a set of small synthetic samples for efficient self-supervised learning (SSL). We first prove that a gradient of synthetic sam ples with respect to a SSL objective in naive bilevel optimization is \textit{bi ased} due to the randomness originating from data augmentations or masking. To a ddress this issue, we propose to minimize the mean squared error (MSE) between a model's representations of the synthetic examples and their corresponding learn able target feature representations for the inner objective, which does not intr oduce any randomness. Our primary motivation is that the model obtained by the p roposed inner optimization can mimic the \textit{self-supervised target model}. To achieve this, we also introduce the MSE between representations of the inner model and the self-supervised target model on the original full dataset for oute r optimization. Lastly, assuming that a feature extractor is fixed, we only opti mize a linear head on top of the feature extractor, which allows us to reduce th e computational cost and obtain a closed-form solution of the head with kernel ridge regression. We empirically validate the effectiveness of our method on var ious applications involving transfer learning.
****
Lorenz Richter,Julius Berner
Improved sampling via learned diffusions
Recently, a series of papers proposed deep learning-based approaches to sample f rom unnormalized target densities using controlled diffusion processes. In this work, we identify these approaches as special cases of the Schrödinger bridge pr oblem, seeking the most likely stochastic evolution between a given prior distri bution and the specified target, and propose the perspective from measures on pa th space as a unifying framework. The optimal controls of such entropy-constrain ed optimal transport problems can then be described by systems of partial differ ential equations and corresponding backward stochastic differential equations. B uilding on these optimality conditions and exploiting the path measure perspecti ve, we obtain variational formulations of the respective approaches and recover the objectives which can be approached via gradient descent. Our formulations al low to introduce losses different from the typically employed reverse Kullback-L eibler divergence that is known to suffer from mode collapse. In particular, we propose the so-called $\textit{log-variance loss}$, which exhibits favorable num erical properties and leads to significantly improved performance across all con sidered approaches.

****
Bowen Yin,Xuying Zhang,Zhong-Yu Li,Li Liu,Ming-Ming Cheng,Qibin Hou
DFormer: Rethinking RGBD Representation Learning for Semantic Segmentation
We present DFormer, a novel RGB-D pretraining framework to learn transferable re presentations for RGB-D segmentation tasks. DFormer has two new key innovations: 1) Unlike previous works that encode RGB-D information with RGB pretrained back bone, we pretrain the backbone using image-depth pairs from ImageNet-1K, and thu s the DFormer is endowed with the capacity to encode RGB-D representations; 2) D Former comprises a sequence of RGB-D blocks, which are tailored for encoding bot h RGB and depth information through a novel building block design. DFormer avoid s the mismatched encoding of the 3D geometry relationships in depth maps by RGB pretrained backbones, which widely lies in existing methods but has not been res olved. We finetune the pretrained DFormer on two popular RGB-D tasks, i.e., RGB-D semantic segmentation and RGB-D salient object detection, with a lightweight d ecoder head. Experimental results show that our DFormer achieves new state-of-th e-art performance on these two tasks with less than half of the computational co st of the current best methods on two RGB-D semantic segmentation datasets and f ive RGB-D salient object detection datasets. Code will be made publicly availabl e.
****
Yuhao Mao,Mark Niklas Mueller,Marc Fischer,Martin Vechev
Understanding Certified Training with Interval Bound Propagation
As robustness verification methods are becoming more precise, training certifiab ly robust neural networks is becoming ever more relevant. To this end, certified training methods compute and then optimize an upper bound on the worst-case los s over a robustness specification. Curiously, training methods based on the impr ecise interval bound propagation (IBP) consistently outperform those leveraging more precise bounds. Still, we lack a theoretical understanding of the mechanism s making IBP so successful. In this work, we investigate these mechanisms by lev eraging a novel metric measuring the tightness of IBP bounds. We first show theo retically that, for deep linear models (DLNs), tightness decreases with width an d depth at initialization, but improves with IBP training. We, then, derive suff icient and necessary conditions on weight matrices for IBP bounds to become exac t and demonstrate that these impose strong regularization, providing an explanat ion for the observed robustness-accuracy trade-off. Finally, we show how these r esults on DLNs transfer to ReLU networks, before conducting an extensive empiric al study, (i) confirming this transferability and yielding state-of-the-art cert ified accuracy, (ii) finding that while all IBP-based training methods lead to h igh tightness, this increase is dominated by the size of the propagated input re gions rather than the robustness specification, and finally (iii) observing that non-IBP-based methods do not increase tightness. Together, these results help e xplain the success of recent certified training methods and may guide the develo pment of new ones.
****
Lijun Yu,Jose Lezama,Nitesh Bharadwaj Gundavarapu,Luca Versari,Kihyuk Sohn,David Minnen,Yong Cheng,Agrim Gupta,Xiuye Gu,Alexander G Hauptmann,Boqing Gong,Ming-H suan Yang,Irfan Essa,David A Ross,Lu Jiang
Language Model Beats Diffusion - Tokenizer is key to visual generation
While Large Language Models (LLMs) are the dominant models for generative tasks in language, they do not perform as well as diffusion models on image and video generation. To effectively use LLMs for visual generation, one crucial component is the visual tokenizer that maps pixel-space inputs to discrete tokens appropr iate for LLM learning. In this paper, we introduce \modelname{}, a video tokeniz er designed to generate concise and expressive tokens for both videos and images using a common token vocabulary. Equipped with this new tokenizer, we show that LLMs outperform diffusion models on standard image and video generation benchma rks including ImageNet and Kinetics. In addition, we demonstrate that our tokeni zer surpasses the previously top-performing video tokenizer on two more tasks: (1) video compression comparable to the next-generation video codec (VCC) accordi ng to human evaluations, and (2) learning effective representations for action r

ecognition tasks.
****
Borui Zhang,Wenzhao Zheng,Jie Zhou,Jiwen Lu
Path Choice Matters for Clear Attributions in Path Methods
Rigorousness and clarity are both essential for interpretations of DNNs to engender human trust. Path methods are commonly employed to generate rigorous attributions that satisfy three axioms. However, the meaning of attributions remains ambiguous due to distinct path choices. To address the ambiguity, we introduce Concentration Principle, which centrally allocates high attributions to indispensable features, thereby endowing aesthetic and sparsity. We then present SAMP, a model-agnostic interpreter, which efficiently searches the near-optimal path from a pre-defined set of manipulation paths. Moreover, we propose the infinitesimal constraint (IC) and momentum strategy (MS) to improve the rigorousness and optimality. Visualizations show that SAMP can precisely reveal DNNs by pinpointing salient image pixels.
We also perform quantitative experiments and observe that our method significantly outperforms the counterparts.
****
Min Lin
Automatic Functional Differentiation in JAX
We extend JAX with the capability to automatically differentiate higher-order functions (functionals and operators). By representing functions as infinite dimensional generalization of arrays, we seamlessly use JAX's existing primitive system to implement higher-order functions. We present a set of primitive operators that serve as foundational building blocks for constructing several key types of functionals. For every introduced primitive operator, we derive and implement both linearization and transposition rules, aligning with JAX's internal protocols for forward and reverse mode automatic differentiation. This enhancement allows for functional differentiation in the same syntax traditionally use for functions. The resulting functional gradients are themselves functions ready to be invoked in python. We showcase this tool's efficacy and simplicity through applications where functional derivatives are indispensable.
****
Hannah Lawrence,Mitchell Tong Harris
Learning Polynomial Problems with $SL(2, \mathbb{R})$-Equivariance
Optimizing and certifying the positivity of polynomials are fundamental primitives across mathematics and engineering applications, from dynamical systems to operations research. However, solving these problems in practice requires large semidefinite programs, with poor scaling in dimension and degree. In this work, we demonstrate for the first time that neural networks can effectively solve such problems in a data-driven fashion, achieving tenfold speedups while retaining high accuracy. Moreover, we observe that these polynomial learning problems are equivariant to the non-compact group $SL(2,\mathbb{R})$, which consists of area-preserving linear transformations. We therefore adapt our learning pipelines to accommodate this structure, including data augmentation, a new $SL(2,\mathbb{R})$-equivariant architecture, and an architecture equivariant with respect to its maximal compact subgroup, $SO(2, \mathbb{R})$. Surprisingly, the most successful approaches in practice do not enforce equivariance to the entire group, which we prove arises from an unusual lack of architecture universality for $SL(2,\mathbb{R})$ in particular. A consequence of this result, which is of independent interest, is that there exists an equivariant function for which there is no sequence of equivariant polynomials multiplied by arbitrary invariants that approximates the original function. This is a rare example of a symmetric problem where data augmentation outperforms a fully equivariant architecture, and provides interesting lessons in both theory and practice for other problems with non-compact symmetries.
****
Thanh Tung Le,Khai Nguyen,shanlin sun,Kun Han,Nhat Ho,Xiaohui Xie
Diffeomorphic Mesh Deformation via Efficient Optimal Transport for Cortical Surface Reconstruction

Mesh deformation plays a pivotal role in many 3D vision tasks including dynamic simulations, rendering, and reconstruction. However, defining an efficient discrepancy between predicted and target meshes remains an open problem. A prevalent approach in current deep learning is the set-based approach which measures the discrepancy between two surfaces by comparing two randomly sampled point-clouds from the two meshes with Chamfer pseudo-distance. Nevertheless, the set-based approach still has limitations such as lacking a theoretical guarantee for choosing the number of points in sampled point-clouds, and the pseudo-metricity and the quadratic complexity of the Chamfer divergence. To address these issues, we propose a novel metric for learning mesh deformation. The metric is defined by sliced Wasserstein distance on meshes represented as probability measures that generalize the set-based approach. By leveraging probability measure space, we gain flexibility in encoding meshes using diverse forms of probability measures, such as continuous, empirical, and discrete measures via \textit{varifold} representation. After having encoded probability measures, we can compare meshes by using the sliced Wasserstein distance which is an effective optimal transport distance with linear computational complexity and can provide a fast statistical rate for approximating the surface of meshes. To the end, we employ a neural ordinary differential equation (ODE) to deform the input surface into the target shape by modeling the trajectories of the points on the surface. Our experiments on cortical surface reconstruction demonstrate that our approach surpasses other competing methods in multiple datasets and metrics.
****
Guanhua Wang,Heyang Qin,Sam Ade Jacobs,Xiaoxia Wu,Connor Holmes,Zhewei Yao,Samyam Rajbhandari,Olatunji Ruwase,Feng Yan,Lei Yang,Yuxiong He
ZeRO++: Extremely Efficient Collective Communication for Large Model Training
Zero Redundancy Optimizer (ZeRO) has been used to train a wide range of large language models on massive GPU clusters due to its ease of use, efficiency, and good scalability. However, when training on low-bandwidth clusters, and/or when small batch size per GPU is used, ZeRO’s effective throughput is limited due to communication overheads. To alleviate this limitation, this paper introduces ZeRO++ composing of three communication volume reduction techniques (lowprecision all-gather, data remapping, and low-precision gradient averaging) to significantly reduce the communication volume up to 4x that enables up to 2.16x better throughput at 384 GPU scale. Our results also show ZeRO++ can speedup the RLHF by 3.3x compared to vanilla ZeRO. To verify the convergence of ZeRO++, we test up to 13B model for pretraining with 8/6-bits all gather and up to 30B model for finetuning with 4/2-bits all gather, and demonstrate on-par accuracy as original ZeRO (aka standard training). As a byproduct, the model trained with ZeRO++ is naturally weight-quantized, which can be directly used for inference without post-training quantization or quantization-aware training.
****
Jonas Seng,Matej Ze■evi■,Devendra Singh Dhami,Kristian Kersting
Learning Large DAGs is Harder than you Think: Many Losses are Minimal for the Wrong DAG
Structure learning is a crucial task in science, especially in fields such as medicine and biology, where the wrong identification of (in)dependencies among random variables can have significant implications. The primary objective of structure learning is to learn a Directed Acyclic Graph (DAG) that represents the underlying probability distribution of the data. Many prominent DAG learners rely on least square losses or log-likelihood losses for optimization. It is well-known from regression models that least square losses are heavily influenced by the scale of the variables. Recently it has been demonstrated that the scale of data also affects performance of structure learning algorithms, though with a strong focus on linear 2-node systems and simulated data. Moving beyond these results, we provide conditions under which square-based losses are minimal for wrong DAGs in d-dimensional cases. Furthermore, we also show that scale can impair performance of structure learners if relations among variables are non-linear for both square based and log-likelihood based losses. We confirm our theoretical findings through extensive experiments on synthetic and real-world data.

****
YU-JU TSAI,Yu-Lun Liu,Lu Qi,Kelvin C.K. Chan,Ming-Hsuan Yang
Dual Associated Encoder for Face Restoration
Restoring facial details from low-quality (LQ) images has remained challenging due to the nature of the problem caused by various degradations in the wild.
The codebook prior has been proposed to address the ill-posed problems by leveraging an autoencoder and learned codebook of high-quality (HQ) features, achieving remarkable quality.
However, existing approaches in this paradigm frequently depend on a single encoder pre-trained on HQ data for restoring HQ images, disregarding the domain gap and distinct feature representations between LQ and HQ images.
As a result, encoding LQ inputs with the same encoder could be insufficient, resulting in imprecise feature representation and leading to suboptimal performance.
To tackle this problem, we propose a novel dual-branch framework named $\textit{DAEFR}$. Our method introduces an auxiliary LQ branch that extracts domain-specific information from the LQ inputs.
Additionally, we incorporate association training to promote effective synergy between the two branches, enhancing code prediction and restoration quality.
We evaluate the effectiveness of DAEFR on both synthetic and real-world datasets, demonstrating its superior performance in restoring facial details.
Project page: https://liagm.github.io/DAEFR/
****
Nan Ding,Tomer Levinboim,Jialin Wu,Sebastian Goodman,Radu Soricut
CausalLM is not optimal for in-context learning
Recent empirical evidence indicates that transformer based in-context learning performs better when using a prefix language model (prefixLM), in which in-context samples can all attend to each other, compared to causal language models (causalLM), which use auto-regressive attention that prohibits in-context samples to attend to future samples. While this result is intuitive, it is not understood from a theoretical perspective. In this paper we take a theoretical approach and analyze the convergence behavior of prefixLM and causalLM under a certain parameter construction. Our analysis shows that both LM types converge to their stationary points at a linear rate, but that while prefixLM converges to the optimal solution of linear regression, causalLM convergence dynamics follows that of an online gradient descent algorithm, which is not guaranteed to be optimal even as the number of samples grows infinitely. We supplement our theoretical claims with empirical experiments over synthetic and real tasks and using various types of transformers. Our experiments verify that causalLM consistently underperforms prefixLM in all settings.
****
Junlong Li,Shichao Sun,Weizhe Yuan,Run-Ze Fan,hai zhao,Pengfei Liu
Generative Judge for Evaluating Alignment
The rapid development of Large Language Models (LLMs) has substantially expanded the range of tasks they can address. In the field of Natural Language Processing (NLP), researchers have shifted their focus from conventional NLP tasks (e.g., sequence tagging and parsing) towards tasks that revolve around aligning with human needs (e.g., brainstorming and email writing). This shift in task distribution imposes new requirements on evaluating these aligned models regarding *generality* (i.e., assessing performance across diverse scenarios), *flexibility* (i.e., examining under different protocols), and *interpretability* (i.e., scrutinizing models with explanations). In this paper, we propose a generative judge with 13B parameters, **Auto-J**, designed to address these challenges. Our model is trained on user queries and LLM-generated responses under massive real-world scenarios and accommodates diverse evaluation protocols (e.g., pairwise response comparison and single-response evaluation) with well-structured natural language critiques. To demonstrate the efficacy of our approach, we construct a new testbed covering 58 different scenarios. Experimentally, **Auto-J** outperforms a series of strong competitors, including both open-source and closed-source models, by a large margin. We also provide detailed analysis and case studies to further

reveal the potential of our method and make a variety of resources public at https://github.com/GAIR-NLP/auto-j.
****
Beatrice Bevilacqua,Moshe Eliasof,Eli Meirom,Bruno Ribeiro,Haggai Maron
Efficient Subgraph GNNs by Learning Effective Selection Policies
Subgraph GNNs are provably expressive neural architectures that learn graph representations from sets of subgraphs. Unfortunately, their applicability is hampered by the computational complexity associated with performing message passing on many subgraphs. In this paper, we consider the problem of learning to select a small subset of the large set of possible subgraphs in a data-driven fashion. We first motivate the problem by proving that there are families of WL-indistinguishable graphs for which there exist efficient subgraph selection policies: small subsets of subgraphs that can already identify all the graphs within the family. We then propose a new approach, called _Policy-Learn_, that learns how to select subgraphs in an iterative manner. We prove that, unlike popular random policies and prior work addressing the same problem, our architecture is able to learn the efficient policies mentioned above. Our experimental results demonstrate that _Policy-Learn_ outperforms existing baselines across a wide range of datasets.
****
Chongyu Fan,Jiancheng Liu,Yihua Zhang,Eric Wong,Dennis Wei,Sijia Liu
SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation
With evolving data regulations, machine unlearning (MU) has become an important tool for fostering trust and safety in today's AI models. However, existing MU methods focusing on data and/or weight perspectives often suffer limitations in unlearning accuracy, stability, and cross-domain applicability. To address these challenges, we introduce the concept of 'weight saliency' for MU, drawing parallels with input saliency in model explanation. This innovation directs MU's attention toward specific model weights rather than the entire model, improving effectiveness and efficiency. The resultant method that we call saliency unlearning (SalUn) narrows the performance gap with 'exact' unlearning (model retraining from scratch after removing the forgetting data points). To the best of our knowledge, SalUn is the first principled MU approach that can effectively erase the influence of forgetting data, classes, or concepts in both image classification and generation tasks. As highlighted below, For example, SalUn yields a stability advantage in high-variance random data forgetting, e.g., with a 0.2% gap compared to exact unlearning on the CIFAR-10 dataset. Moreover, in preventing conditional diffusion models from generating harmful images, SalUn achieves nearly 100% unlearning accuracy, outperforming current state-of-the-art baselines like Erased Stable Diffusion and Forget-Me-Not. Codes are available at https://github.com/OPTML-Group/Unlearn-Saliency.

WARNING: This paper contains model outputs that may be offensive in nature.
****
Niloofar Mireshghallah,Hyunwoo Kim,Xuhui Zhou,Yulia Tsvetkov,Maarten Sap,Reza Shokri,Yejin Choi
Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory
Existing efforts on quantifying privacy implications for large language models (LLMs) solely focus on measuring leakage of training data. In this work, we shed light on the often-overlooked interactive settings where an LLM receives information from multiple sources and generates an output to be shared with other entities, creating the potential of exposing sensitive input data in inappropriate contexts. In these scenarios, humans nat- urally uphold privacy by choosing whether or not to disclose information depending on the context. We ask the question “Can LLMs demonstrate an equivalent discernment and reasoning capability when considering privacy in context?” We propose CONFAIDE, a benchmark grounded in the theory of contextual integrity and designed to identify critical weaknesses in the privacy reasoning capabilities of instruction-tuned LLMs. CONFAIDE consists of

four tiers, gradually increasing in complexity, with the final tier evaluating contextual privacy reasoning and theory of mind capabilities. Our experiments show that even commercial models such as GPT-4 and ChatGPT reveal private information in contexts that humans would not, 39% and 57% of the time, respectively, highlighting the urgent need for a new direction of privacy-preserving approaches as we demonstrate a larger underlying problem stemmed in the models' lack of reasoning capabilities.
****
Zhaomin Wu,Junyi Hou,Bingsheng He
VertiBench: Advancing Feature Distribution Diversity in Vertical Federated Learning Benchmarks
Vertical Federated Learning (VFL) is a crucial paradigm for training machine learning models on feature-partitioned, distributed data. However, due to privacy restrictions, few public real-world VFL datasets exist for algorithm evaluation, and these represent a limited array of feature distributions. Existing benchmarks often resort to synthetic datasets, derived from arbitrary feature splits from a global set, which only capture a subset of feature distributions, leading to inadequate algorithm performance assessment. This paper addresses these shortcomings by introducing two key factors affecting VFL performance - feature importance and feature correlation - and proposing associated evaluation metrics and dataset splitting methods. Additionally, we introduce a real VFL dataset to address the deficit in image-image VFL scenarios. Our comprehensive evaluation of cutting-edge VFL algorithms provides valuable insights for future research in the field.
****
Ted Moskovitz,Aaditya K Singh,DJ Strouse,Tuomas Sandholm,Ruslan Salakhutdinov,Anca Dragan,Stephen Marcus McAleer
Confronting Reward Model Overoptimization with Constrained RLHF
Large language models are typically aligned with human preferences by optimizing reward models (RMs) fitted to human feedback. However, human preferences are multi-faceted, and it is increasingly common to derive reward from a composition of simpler reward models which each capture a different aspect of language quality. This itself presents a challenge, as it is difficult to appropriately weight these component RMs when combining them. Compounding this difficulty, because any RM is only a proxy for human evaluation, this process is vulnerable to *overoptimization*, wherein past a certain point, accumulating higher reward is associated with worse human ratings. In this paper, we perform the first study on overoptimization in composite RMs, showing that correlation between component RMs has a significant effect on the locations of these points. We then introduce an approach to solve this issue using constrained reinforcement learning as a means of preventing the agent from exceeding each RM's threshold of usefulness. Our method addresses the problem of weighting component RMs by learning dynamic weights, naturally given by the Lagrange multipliers. As a result, each RM stays within the range at which it is an effective proxy, improving evaluation performance. Finally, we introduce an adaptive method using gradient-free optimization to identify and optimize towards these points during a single run.
****
Kaijie Zhu,Jiaao Chen,Jindong Wang,Neil Zhenqiang Gong,Diyi Yang,Xing Xie
DyVal: Dynamic Evaluation of Large Language Models for Reasoning Tasks
Large language models (LLMs) have achieved remarkable performance in various evaluation benchmarks. However, concerns are raised about potential data contamination in their considerable volume of training corpus. Moreover, the static nature and fixed complexity of current benchmarks may inadequately gauge the advancing capabilities of LLMs.
In this paper, we introduce DyVal, a general and flexible protocol for dynamic evaluation of LLMs. Based on our framework, we build graph-informed DyVal by leveraging the structural advantage of directed acyclic graphs to dynamically generate evaluation samples with controllable complexities. DyVal generates challenging evaluation sets on reasoning tasks including mathematics, logical reasoning, and algorithm problems. We evaluate various LLMs ranging from Flan-T5-large to GP

T-3.5-Turbo and GPT-4. Experiments show that LLMs perform worse in DyVal-generated evaluation samples with different complexities, highlighting the significance of dynamic evaluation.
We also analyze the failure cases and results of different prompting methods.
Moreover, DyVal-generated samples are not only evaluation sets, but also helpful data for fine-tuning to improve the performance of LLMs on existing benchmarks.
We hope that DyVal can shed light on future evaluation research of LLMs. Code is available at: https://github.com/microsoft/promptbench.
****
Miao Xiong,Zhiyuan Hu,Xinyang Lu,YIFEI LI,Jie Fu,Junxian He,Bryan Hooi
Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs
Empowering large language models (LLMs) to accurately express confidence in their answers is essential for reliable and trustworthy decision-making. Previous confidence elicitation methods, which primarily rely on *white-box access* to internal model information or model fine-tuning, have become less suitable for LLMs, especially closed-source commercial APIs. This leads to a growing need to explore the untapped area of *black-box* approaches for LLM uncertainty estimation. To better break down the problem, we define a systematic framework with three components: *prompting* strategies for eliciting verbalized confidence, *sampling* methods for generating multiple responses, and *aggregation* techniques for computing consistency. We then benchmark these methods on two key tasks—confidence calibration and failure prediction—across five types of datasets (e.g., commonsense and arithmetic reasoning) and five widely-used LLMs including GPT-4 and LLaMA 2 Chat. Our analysis uncovers several key insights: 1) LLMs, when verbalizing their confidence, tend to be *overconfident*, potentially imitating human patterns of expressing confidence. 2) As model capability scales up, both calibration and failure prediction performance improve, yet still far from ideal performance.

3) Employing our proposed strategies, such as human-inspired prompts, consistency among multiple responses, and better aggregation strategies can help mitigate this overconfidence from various perspectives.
4) Comparisons with white-box methods indicate that while white-box methods perform better, the gap is narrow, e.g., 0.522 to 0.605 in AUROC. Despite these advancements, none of these techniques consistently outperform others, and all investigated methods struggle in challenging tasks, such as those requiring professional knowledge, indicating significant scope for improvement. We believe this study can serve as a strong baseline and provide insights for eliciting confidence in black-box LLMs. The code is publicly available at https://github.com/MiaoXiong2320/llm-uncertainty.
****
Junwei Su,Difan Zou,Chuan Wu
PRES: Toward Scalable Memory-Based Dynamic Graph Neural Networks
Memory-based Dynamic Graph Neural Networks (MDGNNs) are a family of dynamic graph neural networks that leverage a memory module to extract, distill, and memorize long-term temporal dependencies, leading to superior performance compared to memory-less counterparts. However, training MDGNNs faces the challenge of handling entangled temporal and structural dependencies, requiring sequential and chronological processing of data sequences to capture accurate temporal patterns. During the batch training, the temporal data points within the same batch will be processed in parallel, while their temporal dependencies are neglected. This issue is referred to as temporal discontinuity and restricts the effective temporal batch size, limiting data parallelism and reducing MDGNNs' flexibility in industrial applications. This paper studies the efficient training of MDGNNs at scale, focusing on the temporal discontinuity in training MDGNNs with large temporal batch sizes. We first conduct a theoretical study on the impact of temporal batch

size on the convergence of MDGNN training. Based on the analysis, we propose PRES, an iterative prediction-correction scheme combined with a memory coherence learning objective to mitigate the effect of temporal discontinuity, enabling MDGN

Ns to be trained with significantly larger temporal batches without sacrificing generalization performance. Experimental results demonstrate that our approach e nables up to a 4 $\times$ larger temporal batch (3.4$\times$ speed-up) during MD GNN training.
****
Zhenheng Tang,Yonggang Zhang,Shaohuai Shi,Xinmei Tian,Tongliang Liu,Bo Han,Xiaow en Chu
FedImpro: Measuring and Improving Client Update in Federated Learning
Federated Learning (FL) models often experience client drift caused by heterogen eous data, where the distribution of data differs across clients. To address thi s issue, advanced research primarily focuses on manipulating the existing gradie nts to achieve more consistent client models. In this paper, we present an alter native perspective on client drift and aim to mitigate it by generating improved local models. First, we analyze the generalization contribution of local traini ng and conclude that this generalization contribution is bounded by the conditio nal Wasserstein distance between the data distribution of different clients. The n, we propose FedImpro, to construct similar conditional distributions for local training. Specifically, FedImpro decouples the model into high-level and low-le vel components, and trains the high-level portion on reconstructed feature distr ibutions. This approach enhances the generalization contribution and reduces the dissimilarity of gradients in FL. Experimental results show that FedImpro can h elp FL defend against data heterogeneity and enhance the generalization performa nce of the model.
****
Mert Yuksekgonul,Varun Chandrasekaran,Erik Jones,Suriya Gunasekar,Ranjita Naik,H amid Palangi,Ece Kamar,Besmira Nushi
Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Languag e Models
We investigate the internal behavior of Transformer-based Large Language Models (LLMs) when they generate factually incorrect text. We propose modeling factual queries as constraint satisfaction problems and use this framework to investigat e how the LLM interacts internally with factual constraints. We find a strong po sitive relationship between the LLM's attention to constraint tokens and the fac tual accuracy of generations. We curate a suite of 10 datasets containing over 4 0,000 prompts to study the task of predicting factual errors with the Llama-2 fa mily across all scales (7B, 13B, 70B). We propose SAT Probe, a method probing at tention patterns, that can predict factual errors and fine-grained constraint sa tisfaction, and allow early error identification. The approach and findings take another step towards using the mechanistic understanding of LLMs to enhance the ir reliability.
****
Yiming Xie,Varun Jampani,Lei Zhong,Deqing Sun,Huaizu Jiang
OmniControl: Control Any Joint at Any Time for Human Motion Generation
We present a novel approach named OmniControl for incorporating flexible spatial control signals into a text-conditioned human motion generation model based on the diffusion process. Unlike previous methods that can only control the pelvis trajectory, OmniControl can incorporate flexible spatial control signals over di fferent joints at different times with only one model. Specifically, we propose analytic spatial guidance that ensures the generated motion can tightly conform to the input control signals. At the same time, realism guidance is introduced t o refine all the joints to generate more coherent motion. Both the spatial and r ealism guidance are essential and they are highly complementary for balancing co ntrol accuracy and motion realism. By combining them, OmniControl generates moti ons that are realistic, coherent, and consistent with the spatial constraints. E xperiments on HumanML3D and KIT-ML datasets show that OmniControl not only achie ves significant improvement over state-of-the-art methods on pelvis control but also shows promising results when incorporating the constraints over other joint s. Project page: https://neu-vi.github.io/omnicontrol/.
****
Thomas Laurent,James von Brecht,Xavier Bresson

Feature Collapse
We formalize and study a phenomenon called *feature collapse* that makes precise the intuitive idea that entities playing a similar role in a learning task receive similar representations. As feature collapse requires a notion of task, we leverage a synthetic task in which a learner must classify `sentences' constituted of L tokens. We start by showing experimentally that feature collapse goes hand in hand with generalization. We then prove that, in the large sample limit, distinct tokens that play identical roles in the task receive identical local feature representations in the first layer of the network. This analysis shows that a neural network trained on this task provably learns interpretable and meaningful representations in its first layer.
****
Pablo Barcelo,Alexander Kozachinskiy,Anthony Widjaja Lin,Vladimir Podolskii
Logical Languages Accepted by Transformer Encoders with Hard Attention
We contribute to the study of formal languages that can be recognized by transformer encoders. We focus on two self-attention mechanisms: (1) UHAT (Unique Hard Attention Transformers) and (2) AHAT (Average Hard Attention Transformers). UHAT encoders are known to recognize only languages inside the circuit complexity class ${\sf AC}^0$, i.e., accepted by a family of poly-sized and depth-bounded boolean circuits with unbounded fan-ins. On the other hand, AHAT encoders can recognize languages outside ${\sf AC}^0$), but their expressive power still lies within the bigger circuit complexity class ${\sf TC}^0$, i.e., ${\sf AC}^0$-circuits extended by majority gates.
We first show a negative result that there is an ${\sf AC}^0$-language that cannot be recognized by an UHAT encoder. On the positive side, we show that UHAT encoders can recognize a rich fragment of ${\sf AC}^0$-languages, namely, all languages definable in first-order logic with arbitrary unary numerical predicates. This logic, includes, for example, all regular languages from ${\sf AC}^0$. We then show that AHAT encoders can recognize all languages of our logic even when we enrich it with counting terms. Using these results, we obtain a characterization of which counting properties are expressible by UHAT and AHAT, in relation to regular languages.
****
Pablo Pernias,Dominic Rampas,Mats Leon Richter,Christopher Pal,Marc Aubreville
Würstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models
We introduce Würstchen, a novel architecture for text-to-image synthesis that combines competitive performance with unprecedented cost-effectiveness for large-scale text-to-image diffusion models.
A key contribution of our work is to develop a latent diffusion technique in which we learn a detailed but extremely compact semantic image representation used to guide the diffusion process. This highly compressed representation of an image provides much more detailed guidance compared to latent representations of language and this significantly reduces the computational requirements to achieve state-of-the-art results. Our approach also improves the quality of text-conditioned image generation based on our user preference study.
The training requirements of our approach consists of 24,602 A100-GPU hours - compared to Stable Diffusion 2.1's 200,000 GPU hours.
Our approach also requires less training data to achieve these results. Furthermore, our compact latent representations allows us to perform inference over twice as fast, slashing the usual costs and carbon footprint of a state-of-the-art (SOTA) diffusion model significantly, without compromising the end performance. In a broader comparison against SOTA models our approach is substantially more efficient and compares favourably in terms of image quality.
We believe that this work motivates more emphasis on the prioritization of both performance and computational accessibility.
****
Federico Bianchi,Mirac Suzgun,Giuseppe Attanasio,Paul Rottger,Dan Jurafsky,Tatsunori Hashimoto,James Zou
Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models

that Follow Instructions
Training large language models to follow instructions makes them perform better on a wide range of tasks and generally become more helpful. However, a perfectly helpful model will follow even the most malicious instructions and readily generate harmful content.
In this paper, we raise concerns over the safety of models that only emphasize helpfulness, not harmlessness, in their instruction-tuning.
We show that several popular instruction-tuned models are highly unsafe. Moreover, we show that adding just 3\% safety examples (a few hundred demonstrations) when fine-tuning a model like LLaMA can substantially improve its safety. Our safety-tuning does not make models significantly less capable or helpful as measured by standard benchmarks. However, we do find exaggerated safety behaviours, where too much safety-tuning makes models refuse perfectly safe prompts if they superficially resemble unsafe ones. As a whole, our results illustrate trade-offs in training LLMs to be helpful and training them to be safe.
****
Daniel Y Fu,Hermann Kumbong,Eric Nguyen,Christopher Re
FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores
Convolution models with long filters have demonstrated state-of-the-art reasoning abilities in many long-sequence tasks but lag behind the most optimized Transformers in wall-clock time.
A major bottleneck is the Fast Fourier Transform (FFT)---which allows long convolutions to run in $O(N\log N)$ time in sequence length N but has poor hardware utilization.
In this paper, we study how to optimize the FFT convolution.
We find two key bottlenecks: the FFT does not effectively use specialized matrix multiply units, and it incurs expensive I/O between layers of the memory hierarchy.
In response, we propose FlashFFTConv.
FlashFFTConv uses a matrix decomposition that computes the FFT using matrix multiply units and enables kernel fusion for long sequences, reducing I/O.
We also present two sparse convolution algorithms---1) partial convolutions and 2) frequency-sparse convolutions---which can be implemented simply by skipping blocks in the matrix decomposition, enabling further opportunities for memory and compute savings.
FlashFFTConv speeds up exact FFT convolutions by up to 8.7$\times$ over PyTorch and achieves up to 4.4$\times$ speedup end-to-end.
Given the same compute budget, FlashFFTConv allows Hyena-GPT-s to achieve 2.3 points better perplexity and M2-BERT-base to achieve 3.3 points higher GLUE score---matching models with twice the parameter count.
FlashFFTConv also achieves 96.1% accuracy on Path-512, a high-resolution vision task where no model had previously achieved better than 50%.
Furthermore, partial convolutions enable longer-sequence models---yielding the first DNA model that can process the longest human genes (2.3M base pairs)---and frequency-sparse convolutions speed up pretrained models while maintaining or improving model quality.
****
Aiwei Liu,Leyi Pan,Xuming Hu,Shuang Li,Lijie Wen,Irwin King,Philip S. Yu
An Unforgeable Publicly Verifiable Watermark for Large Language Models
Recently, text watermarking algorithms for large language models (LLMs) have been proposed to mitigate the potential harms of text generated by LLMs, including fake news and copyright issues. However, current watermark detection algorithms require the secret key used in the watermark generation process, making them susceptible to security breaches and counterfeiting during public detection.
To address this limitation, we propose an unforgeable publicly verifiable watermark algorithm named UPV that uses two different neural networks for watermark generation and detection, instead of using the same key at both stages. Meanwhile, the token embedding parameters are shared between the generation and detection networks, which makes the detection network achieve a high accuracy very efficiently.

Experiments demonstrate that our algorithm attains high detection accuracy and computational efficiency through neural networks. Subsequent analysis confirms the high complexity involved in forging the watermark from the detection network. Our code is available at https://github.com/THU-BPM/unforgeable_watermark
****
Gunho Park,Baeseong park,Minsub Kim,Sungjae Lee,Jeonghoon Kim,Beomseok Kwon,Se Jung Kwon,Byeongwook Kim,Youngjoo Lee,Dongsoo Lee
LUT-GEMM: Quantized Matrix Multiplication based on LUTs for Efficient Inference in Large-Scale Generative Language Models
Recent advances in self-supervised learning and the Transformer architecture have significantly improved natural language processing (NLP), achieving remarkably low perplexity.
However, the growing size of NLP models introduces a memory wall problem during the generation phase.
To mitigate this issue, recent efforts have focused on quantizing model weights to sub-4-bit precision while preserving full precision for activations, resulting in practical speed-ups during inference on a single GPU.
However, these improvements primarily stem from reduced memory movement, which necessitates a resource-intensive dequantization process rather than actual computational reduction.
In this paper, we introduce LUT-GEMM, an efficient kernel for quantized matrix multiplication, which not only eliminates the resource-intensive dequantization process but also reduces computational costs compared to previous kernels for weight-only quantization.
Furthermore, we proposed group-wise quantization to offer a flexible trade-off between compression ratio and accuracy.
The impact of LUT-GEMM is facilitated by implementing high compression ratios through low-bit quantization and efficient LUT-based operations.
We show experimentally that when applied to the OPT-175B model with 3-bit quantization, LUT-GEMM substantially accelerates token generation latency, achieving a remarkable 2.1x improvement on a single GPU when compared to OPTQ, which relies on the costly dequantization process.
****
Dongyang Liu,Meina Kan,Shiguang Shan,Xilin CHEN
A Simple Romance Between Multi-Exit Vision Transformer and Token Reduction
Vision Transformers (ViTs) are now flourishing in the computer vision area. Despite the remarkable success, ViTs suffer from high computational costs, which greatly hinder their practical usage. Token reduction, which identifies and discards unimportant tokens during forward propagation, has then been proposed to make ViTs more efficient. For token reduction methodologies, a scoring metric is essential to distinguish between important and unimportant tokens. The attention score from the $\mathrm{[CLS]}$ token, which takes the responsibility to aggregate useful information and form the final output, has been established by prior works as an advantageous choice. Nevertheless, whereas the task pressure is applied at the end of the whole model, token reduction generally starts from very early blocks. Given the long distance in between, in the early blocks, $\mathrm{[CLS]}$ token lacks the impetus to gather task-relevant information, causing somewhat arbitrary attention allocation. This phenomenon, in turn, degrades the reliability of token scoring and substantially compromises the effectiveness of token reduction. Inspired by advances in the domain of dynamic neural networks, in this paper, we introduce Multi-Exit Token Reduction (METR), a simple romance between multi-exit architecture and token reduction—two areas previously considered orthogonal. By injecting early task pressure via multi-exit loss, the $\mathrm{[CLS]}$ token is spurred to collect task-related information in even early blocks, thus bolstering the credibility of $\mathrm{[CLS]}$ attention as a token-scoring metric. Additionally, we employ self-distillation to further refine the quality of early supervision. Extensive experiments substantiate both the existence and effectiveness of the newfound chemistry. Comparative assessments also indicate that METR outperforms state-of-the-art token reduction methods on standard benchmarks, especially under aggressive reduction ratios.

****
Arip Asadulaev,Alexander Korotin,Vage Egiazarian,Petr Mokrov,Evgeny Burnaev
Neural Optimal Transport with General Cost Functionals
We introduce a novel neural network-based algorithm to compute optimal transport (OT) plans for general cost functionals. In contrast to common Euclidean costs, i.e., ℓ^1 or ℓ^2, such functionals provide more flexibility and allow using auxiliary information, such as class labels, to construct the required transport map. Existing methods for general cost functionals are discrete and do not provide an out-of-sample estimation. We address the challenge of designing a continuous OT approach for general cost functionals in high-dimensional spaces, such as images. We construct two example functionals: one to map distributions while preserving the class-wise structure and the other one to preserve the given data pairs. Additionally, we provide the theoretical error analysis for our recovered transport plans. Our implementation is available at \url{https://github.com/machinestein/gnot}
****
Aming WU,Cheng Deng
Modulated Phase Diffusor: Content-Oriented Feature Synthesis for Detecting Unknown Objects
To promote the safe deployment of object detectors, a task of unsupervised out-of-distribution object detection (OOD-OD) is recently proposed, aiming to detect unknown objects during training without reliance on any auxiliary OOD data. To alleviate the impact of lacking OOD data, for this task, one feasible solution is to exploit the known in-distribution (ID) data to synthesize proper OOD information for supervision, which strengthens detectors' discrimination. From the frequency perspective, since the phase generally reflects the content of the input, in this paper, we explore leveraging the phase of ID features to generate expected OOD features involving different content. And a method of Modulated Phase Diffusion (MPD) is proposed, containing a shared forward and two different reverse processes. Specifically, after calculating the phase of the extracted features, to prevent the rapid loss of content in the phase, the forward process gradually performs Gaussian Average on the phase instead of adding noise. The averaged phase and original amplitude are combined to obtain the features taken as the input of the reverse process. Next, one OOD branch is defined to synthesize virtual OOD features by continually enlarging the content discrepancy between the OOD features and original ones. Meanwhile, another modulated branch is designed to generate augmented features owning a similar phase as the original features by scaling and shifting the OOD branch. Both original and augmented features are used for training, enhancing the discrimination. Experimental results on OOD-OD, incremental object detection, and open-set object detection demonstrate the superiorities of our method. The source code will be released at https://github.com/AmingWu/MPD.
****
Zhaoyuan Yang,Zhengyang Yu,Zhiwei Xu,Jaskirat Singh,Jing Zhang,Dylan Campbell,Peter Tu,Richard Hartley
IMPUS: Image Morphing with Perceptually-Uniform Sampling Using Diffusion Models
We present a diffusion-based image morphing approach with perceptually-uniform sampling (IMPUS) that produces smooth, direct and realistic interpolations given an image pair. The embeddings of two images may lie on distinct conditioned distributions of a latent diffusion model, especially when they have significant semantic difference. To bridge this gap, we interpolate in the locally linear and continuous text embedding space and Gaussian latent space. We first optimize the endpoint text embeddings and then map the images to the latent space using a probability flow ODE. Unlike existing work that takes an indirect morphing path, we show that the model adaptation yields a direct path and suppresses ghosting artifacts in the interpolated images. To achieve this, we propose a heuristic bottleneck constraint based on a novel relative perceptual path diversity score that automatically controls the bottleneck size and balances the diversity along the path with its directness. We also propose a perceptually-uniform sampling technique that enables visually smooth changes between the interpolated images. Extens

ive experiments validate that our IMPUS can achieve smooth, direct, and realisti
c image morphing and is adaptable to several other generative tasks.
****
Haoyue Dai,Ignavier Ng,Gongxu Luo,Peter Spirtes,Petar Stojanov,Kun Zhang
Gene Regulatory Network Inference in the Presence of Dropouts: a Causal View
Gene regulatory network inference (GRNI) is a challenging problem, particularly owing to the presence of zeros in single-cell RNA sequencing data: some are biological zeros representing no gene expression, while some others are technical zeros arising from the sequencing procedure (aka dropouts), which may bias GRNI by distorting the joint distribution of the measured gene expressions. Existing approaches typically handle dropout error via imputation, which may introduce spurious relations as the true joint distribution is generally unidentifiable. To tackle this issue, we introduce a causal graphical model to characterize the dropout mechanism, namely, Causal Dropout Model. We provide a simple yet effective theoretical result: interestingly, the conditional independence (CI) relations in the data with dropouts, after deleting the samples with zero values (regardless if technical or not) for the conditioned variables, are asymptotically identical to the CI relations in the original data without dropouts. This particular test-wise deletion procedure, in which we perform CI tests on the samples without zeros for the conditioned variables, can be seamlessly integrated with existing structure learning approaches including constraint-based and greedy score-based methods, thus giving rise to a principled framework for GRNI in the presence of dropouts. We further show that the causal dropout model can be validated from data, and many existing statistical models to handle dropouts fit into our model as specific parametric instances. Empirical evaluation on synthetic, curated, and real-world experimental transcriptomic data comprehensively demonstrate the efficacy of our method.
****
Ziteng Wang,Jianfei Chen,Jun Zhu
Efficient Backpropagation with Variance Controlled Adaptive Sampling
Sampling-based algorithms, which eliminate "unimportant" computations during forward and/or backpropagation (BP), offer potential solutions to accelerate neural network training. However, since sampling introduces approximations to training, such algorithms may not consistently maintain accuracy across various tasks. In this work, we introduce a variance-controlled adaptive sampling (VCAS) method designed to minimize the computational load of BP. VCAS computes an unbiased stochastic gradient with fine-grained layerwise importance sampling in data dimension for activation gradient calculation and leverage score sampling in token dimension for weight gradient calculation. To preserve accuracy, we control the additional variance introduced by learning the sample ratio jointly with model parameters during training. We assessed VCAS on multiple fine-tuning and pre-training tasks in both vision and natural language domains. On all the tasks, VCAS can preserve the original training loss trajectory and validation accuracy with an up to 73.87% FLOPs reduction of BP and 49.58% FLOPs reduction of the whole training process. The implementation is available at https://github.com/thu-ml/VCAS.
****
Peng Xu,Wenqi Shao,Mengzhao Chen,Shitao Tang,Kaipeng Zhang,Peng Gao,Fengwei An,Yu Qiao,Ping Luo
BESA: Pruning Large Language Models with Blockwise Parameter-Efficient Sparsity Allocation
Large language models (LLMs) have demonstrated outstanding performance in various tasks, such as text summarization, text question-answering, and etc. While their performance is impressive, the computational footprint due to their vast number of parameters can be prohibitive. Existing solutions such as SparseGPT and Wanda attempt to alleviate this issue through weight pruning. However, their layer-wise approach results in significant perturbation to the model's output and requires meticulous hyperparameter tuning, such as the pruning rate, which can adversely affect overall model performance. To address this, this paper introduces a novel LLM pruning technique dubbed blockwise parameter-efficient sparsity allocation (BESA) by applying a blockwise reconstruction loss. In contrast to the typ

ical layer-wise pruning techniques, BESA is characterized by two distinctive att ributes: i) it targets the overall pruning error with respect to individual tran sformer blocks, and ii) it allocates layer-specific sparsity in a differentiable manner, both of which ensure reduced performance degradation after pruning. Our experiments show that BESA achieves state-of-the-art performance, efficiently p runing LLMs like LLaMA1, and LLaMA2 with 7B to 70B parameters on a single A100 G PU in just five hours. Code is available at [here](https://github.com/LinkAnonym ous/BESA).
****
Jiuding Sun,Chantal Shaib,Byron C Wallace
Evaluating the Zero-shot Robustness of Instruction-tuned Language Models
Instruction fine-tuning has recently emerged as a promising approach for improvi ng the zero-shot capabilities of Large Language Models (LLMs) on new tasks. This technique has shown particular strength in improving the performance of modestl y sized LLMs, sometimes inducing performance competitive with much larger model variants. In this paper, we ask two questions: (1) How sensitive are instruction -tuned models to the particular phrasings of instructions, and, (2) How can we m ake them more robust to such natural language variation? To answer the former, w e collect a set of 319 instructions manually written by NLP practitioners for ov er 80 unique tasks included in widely used benchmarks, and we evaluate the varia nce and average performance of these instructions as compared to instruction phr asings observed during instruction fine-tuning. We find that using novel (unobse rved) but appropriate instruction phrasings consistently degrades model performa nce, sometimes substantially so. Further, such natural instructions yield a wide variance in downstream performance, despite their semantic equivalence. Put ano ther way, instruction-tuned models are not especially robust to instruction re-p hrasings.
We propose a simple method to mitigate this issue by introducing ``soft prompt'' embedding parameters and optimizing these to maximize the similarity between re presentations of semantically equivalent instructions. We show that this method consistently improves the robustness of instruction-tuned models.
****
Quentin Delfosse,Patrick Schramowski,Martin Mundt,Alejandro Molina,Kristian Kers ting
Adaptive Rational Activations to Boost Deep Reinforcement Learning
Latest insights from biology show that intelligence not only emerges from the co nnections between neurons, but that individual neurons shoulder more computation al responsibility than previously anticipated. Specifically, neural plasticity s hould be critical in the context of constantly changing reinforcement learning (RL) environments, yet current approaches still primarily employ static activatio n functions. In this work, we motivate the use of adaptable activation functions in RL and show that rational activation functions are particularly suitable for augmenting plasticity. Inspired by residual networks, we derive a condition und er which rational units are closed under residual connections and formulate a na turally regularised version. The proposed joint-rational activation allows for d esirable degrees of flexibility, yet regularises plasticity to an extent that av oids overfitting by leveraging a mutual set of activation function parameters ac ross layers. We demonstrate that equipping popular algorithms with (joint) ratio nal activations leads to consistent improvements on different games from the Ata ri Learning Environment benchmark, notably making DQN competitive to DDQN and Ra inbow.
****
Satoki Ishikawa,Ryo Karakida
On the Parameterization of Second-Order Optimization Effective towards the Infin ite Width
Second-order optimization has been developed to accelerate the training of deep neural networks and it is being applied to increasingly larger-scale models. In this study, towards training on further larger scales, we identify a specific pa rameterization for second-order optimization that promotes feature learning in a stable manner even if the network width increases significantly. Inspired by a

maximal update parametrization, we consider a one-step update of the gradient and reveal the appropriate scales of hyperparameters including random initialization, learning rates, and damping terms. Our approach covers two major second-order optimization algorithms, K-FAC and Shampoo, and we demonstrate that our parametrization achieves higher generalization performance in feature learning.
In particular, it enables us to transfer the hyperparameters across models with different widths.
****
Ricky T. Q. Chen,Yaron Lipman
Flow Matching on General Geometries
We propose Riemannian Flow Matching (RFM), a simple yet powerful framework for training continuous normalizing flows on manifolds. Existing methods for generative modeling on manifolds either require expensive simulation, are inherently unable to scale to high dimensions, or use approximations for limiting quantities that result in biased training objectives. Riemannian Flow Matching bypasses these limitations and offers several advantages over previous approaches: it is simulation-free on simple geometries, does not require divergence computation, and computes its target vector field in closed-form. The key ingredient behind RFM is the construction of a relatively simple premetric for defining target vector fields, which encompasses the existing Euclidean case. To extend to general geometries, we rely on the use of spectral decompositions to efficiently compute premetrics on the fly. Our method achieves state-of-the-art performance on real-world non-Euclidean datasets, and we demonstrate tractable training on general geometries, including triangular meshes with highly non-trivial curvature and boundaries.
****
Reza Esfandiarpoor,Stephen Bach
Follow-Up Differential Descriptions: Language Models Resolve Ambiguities for Image Classification
A promising approach for improving the performance of vision-language models like CLIP for image classification is to extend the class descriptions (i.e., prompts) with related attributes, e.g., using brown sparrow instead of sparrow. However, current zero-shot methods select a subset of attributes regardless of commonalities between the target classes, potentially providing no useful information that would have helped to distinguish between them. For instance, they may use color instead of bill shape to distinguish between sparrows and wrens, which are both brown. We propose Follow-up Differential Descriptions (FuDD), a zero-shot approach that tailors the class descriptions to each dataset and leads to additional attributes that better differentiate the target classes. FuDD first identifies the ambiguous classes for each image, and then uses a Large Language Model (LLM) to generate new class descriptions that differentiate between them. The new class descriptions resolve the initial ambiguity and help predict the correct label. In our experiments, FuDD consistently outperforms generic description ensembles and naive LLM-generated descriptions on 12 datasets. We show that differential descriptions are an effective tool to resolve class ambiguities, which otherwise significantly degrade the performance. We also show that high quality natural language class descriptions produced by FuDD result in comparable performance to few-shot adaptation methods.
****
Ruizhe Liu,Qian Luo,Yanchao Yang
InfoCon: Concept Discovery with Generative and Discriminative Informativeness
We focus on the self-supervised discovery of manipulation concepts that can be adapted and reassembled to address various robotic tasks. We propose that the decision to conceptualize a physical procedure should not depend on how we name it (semantics) but rather on the significance of the informativeness in its representation regarding the low-level physical state and state changes. We model manipulation concepts -- discrete symbols -- as generative and discriminative goals and derive metrics that can autonomously link them to meaningful sub-trajectories from noisy, unlabeled demonstrations. Specifically, we employ a trainable codebook containing encodings --symbols -- capable of synthesizing the end-state of a

sub-trajectory given the current state (generative informativeness). Moreover, the encoding corresponding to a particular sub-trajectory should differentiate the state within and outside it and confidently predict the subsequent action based on the gradient of its discriminative score (discriminative informativeness). These metrics, which do not rely on human annotation, can be seamlessly integrated into a VQ-VAE framework, enabling the partitioning of demonstrations into semantically consistent sub-trajectories, fulfilling the purpose of discovering manipulation concepts and the corresponding (sub)-goal states. We evaluate the effectiveness of the learned concepts by training policies that utilize them as guidance, demonstrating superior performance compared to other baselines. Additionally, our discovered manipulation concepts compare favorably to human-annotated ones, while saving much manual effort. The code and trained models will be made public.
****
Zecheng Hao,Xinyu Shi,Zihan Huang,Tong Bu,Zhaofei Yu,Tiejun Huang
A Progressive Training Framework for Spiking Neural Networks with Learnable Multi-hierarchical Model
Spiking Neural Networks (SNNs) have garnered considerable attention due to their energy efficiency and unique biological characteristics. However, the widely adopted Leaky Integrate-and-Fire (LIF) model, as the mainstream neuron model in current SNN research, has been revealed to exhibit significant deficiencies in deep-layer gradient calculation and capturing global information on the time dimension. In this paper, we propose the Learnable Multi-hierarchical (LM-H) model to address these issues by dynamically regulating its membrane-related factors. We point out that the LM-H model fully encompasses the information representation range of the LIF model while offering the flexibility to adjust the extraction ratio between historical and current information. Additionally, we theoretically demonstrate the effectiveness of the LM-H model and the functionality of its internal parameters, and propose a progressive training algorithm tailored specifically for the LM-H model. Furthermore, we devise an efficient training framework for our novel advanced model, encompassing hybrid training and time-slicing online training. Through extensive experiments on various datasets, we validate the remarkable superiority of our model and training algorithm compared to previous state-of-the-art approaches. Code is available at https://github.com/hzc1208/STBP_LMH.
****
Yikun Ban,Ishika Agarwal,Ziwei Wu,Yada Zhu,Kommy Weldemariam,Hanghang Tong,Jingrui He
Neural Active Learning Beyond Bandits
We study both stream-based and pool-based active learning with neural network approximations. A recent line of works proposed bandit-based approaches that transformed active learning into a bandit problem, achieving both theoretical and empirical success. However, the performance and computational costs of these methods may be susceptible to the number of classes, denoted as K, due to this transformation. Therefore, this paper seeks to answer the question: "How can we mitigate the adverse impacts of K while retaining the advantages of principled exploration and provable performance guarantees in active learning?" To tackle this challenge, we propose two algorithms based on the newly designed exploitation and exploration neural networks for stream-based and pool-based active learning. Subsequently, we provide theoretical performance guarantees for both algorithms in a non-parametric setting, demonstrating a slower error-growth rate concerning K for the proposed approaches. We use extensive experiments to evaluate the proposed algorithms, which consistently outperform state-of-the-art baselines.
****
Junhyung Lyle Kim,Taha Toghani,Cesar A Uribe,Anastasios Kyrillidis
Adaptive Federated Learning with Auto-Tuned Clients
Federated learning (FL) is a distributed machine learning framework where the global model of a central server is trained via multiple collaborative steps by participating clients without sharing their data. While being a flexible framework, where the distribution of local data, participation rate, and computing power

of each client can greatly vary, such flexibility gives rise to many new challenges, especially in the hyperparameter tuning on the client side. We propose Δ-SGD, a simple step size rule for SGD that enables each client to use its own step size by adapting to the local smoothness of the function each client is optimizing. We provide theoretical and empirical results where the benefit of the client adaptivity is shown in various FL scenarios.
****
Samuel Pinilla,Jeyan Thiyagalingam
Global Optimality for Non-linear Constrained Restoration Problems via Invexity
Signal restoration is an important constrained optimization problem with significant applications in various domains. Although non-convex constrained optimization problems have been shown to perform better than convex counterparts in terms of reconstruction quality, convex constrained optimization problems have been preferably for its global optima guarantees. Despite the success of non-convex methods in a large number of applications, it is not an overstatement to say that there is little or no hope for non-convex problems to ensure global optima. In this paper, for the first time, we develop invex constrained optimization theory to mitigate the loss of guarantees for global optima in non-convex constrained inverse problems, where the invex function is a mapping where any critical point is a global minimizer. We also develop relevant theories to extend the global optima guarantee to a set of quasi-invex functions - the largest optimizable mappings. More specifically, we propose a family of invex/quasi-invex of functions for handling constrained inverse problems using the non-convex setting along with guarantees for their global optima. Our experimental evaluation shows that the proposed approach is very promising and can aid in extending existing convex optimization algorithms, such as the alternating direction method of multipliers, and accelerated proximal gradient methods.
****
Tatsunori Taniai,Ryo Igarashi,Yuta Suzuki,Naoya Chiba,Kotaro Saito,Yoshitaka Ushiku,Kanta Ono
Crystalformer: Infinitely Connected Attention for Periodic Structure Encoding
Predicting physical properties of materials from their crystal structures is a fundamental problem in materials science. In peripheral areas such as the prediction of molecular properties, fully connected attention networks have been shown to be successful. However, unlike these finite atom arrangements, crystal structures are infinitely repeating, periodic arrangements of atoms, whose fully connected attention results in *infinitely connected attention*. In this work, we show that this infinitely connected attention can lead to a computationally tractable formulation, interpreted as *neural potential summation*, that performs infinite interatomic potential summations in a deeply learned feature space. We then propose a simple yet effective Transformer-based encoder architecture for crystal structures called *Crystalformer*. Compared to an existing Transformer-based model, the proposed model requires only 29.4% of the number of parameters, with minimal modifications to the original Transformer architecture. Despite the architectural simplicity, the proposed method outperforms state-of-the-art methods for various property regression tasks on the Materials Project and JARVIS-DFT datasets.
****
Rui Zheng,Wei Shen,Yuan Hua,Wenbin Lai,Shihan Dou,Yuhao Zhou,Zhiheng Xi,Xiao Wang,Haoran Huang,Tao Gui,Qi Zhang,Xuanjing Huang
Improving Generalization of Alignment with Human Preferences through Group Invariant Learning
The success of AI assistants based on language models (LLMs) hinges crucially on Reinforcement Learning from Human Feedback (RLHF), which enables the generation of responses more aligned with human preferences.
As universal AI assistants, there's a growing expectation for them to perform consistently across various domains.
However, previous work shows that Reinforcement Learning (RL) often exploits shortcuts to attain high rewards and overlooks challenging samples.
This focus on quick reward gains undermines both the stability in training and t

he model's ability to generalize to new, unseen data.
In this work, we propose a novel approach that can learn a consistent policy via RL across various data groups or domains.
Given the challenges associated with acquiring group annotations, our method automatically classifies data into different groups, deliberately maximizing performance variance.
Then, we optimize the policy to perform well on challenging groups.
Lastly, leveraging the established groups, our approach adaptively adjusts the exploration space, allocating more learning capacity to more challenging data and preventing the model from over-optimizing on simpler data. Experimental results indicate that our approach significantly enhances training stability and model generalization.
****
Sanghyuk Chun
Improved Probabilistic Image-Text Representations
Image-Text Matching (ITM) task, a fundamental vision-language (VL) task, suffers from the inherent ambiguity arising from multiplicity and imperfect annotations. Deterministic functions are not sufficiently powerful to capture ambiguity, prompting the exploration of probabilistic embeddings to tackle the challenge. However, the existing probabilistic ITM approach encounters two key shortcomings; the burden of heavy computations due to the Monte Carlo approximation, and the loss saturation issue in the face of abundant false negatives. To overcome the issues, this paper presents an improved Probabilistic Cross-Modal Embeddings (named PCME++) by introducing a new probabilistic distance with a closed-form solution. In addition, two optimization techniques are proposed to enhance PCME++ further: first, the incorporation of pseudo-positives to prevent the negative effect under massive false negatives; second, mixed sample data augmentation for probabilistic matching. Experimental results on MS-COCO Caption and two extended benchmarks, CxC and ECCV Caption, demonstrate the effectiveness of PCME++ compared to state-of-the-art ITM methods. The robustness of PCME++ is also evaluated under noisy image-text correspondences. In addition, the potential applicability of PCME++ in automatic prompt-filtering for zero-shot classification is shown. The code is available at https://github.com/naver-ai/pcmepp.
****
Jianfei Yang,Hanjie Qian,Yuecong Xu,Kai Wang,Lihua Xie
Can We Evaluate Domain Adaptation Models Without Target-Domain Labels?
Unsupervised domain adaptation (UDA) involves adapting a model trained on a label-rich source domain to an unlabeled target domain. However, in real-world scenarios, the absence of target-domain labels makes it challenging to evaluate the performance of UDA models. Furthermore, prevailing UDA methods relying on adversarial training and self-training could lead to model degeneration and negative transfer, further exacerbating the evaluation problem. In this paper, we propose a novel metric called the Transfer Score to address these issues. The proposed metric enables the unsupervised evaluation of UDA models by assessing the spatial uniformity of the classifier via model parameters, as well as the transferability and discriminability of deep representations. Based on the metric, we achieve three novel objectives without target-domain labels: (1) selecting the best UDA method from a range of available options, (2) optimizing hyperparameters of UDA models to prevent model degeneration, and (3) identifying which checkpoint of UDA model performs optimally. Our work bridges the gap between data-level UDA research and practical UDA scenarios, enabling a realistic assessment of UDA model performance. We validate the effectiveness of our metric through extensive empirical studies on UDA datasets of different scales and imbalanced distributions. The results demonstrate that our metric robustly achieves the aforementioned goals.
****
Sam Toyer,Olivia Watkins,Ethan Adrian Mendes,Justin Svegliato,Luke Bailey,Tiffany Wang,Isaac Ong,Karim Elmaaroufi,Pieter Abbeel,Trevor Darrell,Alan Ritter,Stuart Russell
Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game

While Large Language Models (LLMs) are increasingly being used in real-world applications, they remain vulnerable to *prompt injection attacks*: malicious third party prompts that subvert the intent of the system designer. To help researchers study this problem, we present a dataset of over 126,000 prompt injection attacks and 46,000 prompt-based "defenses" against prompt injection, all created by players of an online game called Tensor Trust. To the best of our knowledge, this is the first dataset that includes both human-generated attacks and defenses for instruction-following LLMs. The attacks in our dataset have easily interpretable structure, and shed light on the weaknesses of LLMs. We also use the dataset to create a benchmark for resistance to two types of prompt injection, which we refer to as *prompt extraction* and *prompt hijacking*. Our benchmark results show that many models are vulnerable to the attack strategies in the Tensor Trust dataset. Furthermore, we show that some attack strategies from the dataset generalize to deployed LLM-based applications, even though they have a very different set of constraints to the game. We release data and code at tensortrust.ai/paper
****
Rui Li,Guoyin Wang,Jiwei Li
Are Human-generated Demonstrations Necessary for In-context Learning?
Despite the promising few-shot ability of large language models (LLMs), the standard paradigm of In-context Learning (ICL) suffers the disadvantages of susceptibility to selected demonstrations and the intricacy to generate these demonstrations. In this paper, we raise the fundamental question that whether human-generated demonstrations are necessary for ICL. To answer this question, we propose self-contemplation prompting strategy (SEC), a paradigm free from human-crafted demonstrations. The key point of SEC is that, instead of using hand-crafted examples as demonstrations in ICL, SEC asks LLMs to first create demonstrations on their own, based on which the final output is generated. SEC is a flexible framework and can be adapted to both the vanilla ICL and the chain-of-thought (CoT), but with greater ease: as the manual-generation process of both examples and rationale can be saved. Extensive experiments in arithmetic reasoning, commonsense reasoning, multi-task language understanding, and code generation benchmarks, show that SEC, which does not require hand-crafted demonstrations, significantly outperforms the zero-shot learning strategy, and achieves comparable results to ICL with hand-crafted demonstrations. This demonstrates that, for many tasks, contemporary LLMs possess a sufficient level of competence to exclusively depend on their own capacity for decision making, removing the need for external training data.
****
Jianliang He,Han Zhong,Zhuoran Yang
Sample-efficient Learning of Infinite-horizon Average-reward MDPs with General Function Approximation
We study infinite-horizon average-reward Markov decision processes (AMDPs) in the context of general function approximation. Specifically, we propose a novel algorithmic framework named Local-fitted Optimization with OPtimism (LOOP), which incorporates both model-based and value-based incarnations. In particular, LOOP features a novel construction of confidence sets and a low-switching policy updating scheme, which are tailored to the average-reward and function approximation setting. Moreover, for AMDPs, we propose a novel complexity measure --- average-reward generalized eluder coefficient (AGEC) --- which captures the challenge of exploration in AMDPs with general function approximation. Such a complexity measure encompasses almost all previously known tractable AMDP models, such as linear AMDPs and linear mixture AMDPs, and also includes newly identified cases such as kernel AMDPs and AMDPs with Bellman eluder dimensions. Using AGEC, we prove that LOOP achieves a sublinear $\tilde{\mathcal{O}}(\mathrm{poly}(d, \mathrm{sp}(V^*)) \sqrt{T\beta})$ regret, where d and β correspond to AGEC and log-covering number of the hypothesis class respectively, $\mathrm{sp}(V^*)$ is the span of the optimal state bias function, T denotes the number of steps, and $\tilde{\mathcal{O}} (\cdot) $ omits logarithmic factors. When specialized to concrete AMDP models, our regret bounds are comparable to those established by

the existing algorithms designed specifically for these special cases. To the best of our knowledge, this paper presents the first comprehensive theoretical framework capable of handling nearly all AMDPs.
****
Jack Merullo,Carsten Eickhoff,Ellie Pavlick
Circuit Component Reuse Across Tasks in Transformer Language Models
Recent work in mechanistic interpretability has shown that behaviors in language models can be successfully reverse-engineered through circuit analysis. A common criticism, however, is that each circuit is task-specific, and thus such analysis cannot contribute to understanding the models at a higher level. In this work, we present evidence that insights (both low-level findings about specific heads and higher-level findings about general algorithms) can indeed generalize across tasks. Specifically, we study the circuit discovered in (Wang, 2022) for the Indirect Object Identification (IOI) task and 1.) show that it reproduces on a larger GPT2 model, and 2.) that it is mostly reused to solve a seemingly different task: Colored Objects (Ippolito & Callison-Burch, 2023). We provide evidence that the process underlying both tasks is functionally very similar, and contains about a 78% overlap in in-circuit attention heads. We further present a proof-of-concept intervention experiment, in which we adjust four attention heads in middle layers in order to ‘repair’ the Colored Objects circuit and make it behave like the IOI circuit. In doing so, we boost accuracy from 49.6% to 93.7% on the Colored Objects task and explain most sources of error. The intervention affects downstream attention heads in specific ways predicted by their interactions in the IOI circuit, indicating that this subcircuit behavior is invariant to the different task inputs. Overall, our results provide evidence that it may yet be possible to explain large language models' behavior in terms of a relatively small number of interpretable task-general algorithmic building blocks and computational components.
****
Arnab Kumar Mondal,Siba Smarak Panigrahi,Sai Rajeswar,Kaleem Siddiqi,Siamak Ravanbakhsh
Efficient Dynamics Modeling in Interactive Environments with Koopman Theory
The accurate modeling of dynamics in interactive environments is critical for successful long-range prediction. Such a capability could advance Reinforcement Learning (RL) and Planning algorithms, but achieving it is challenging. Inaccuracies in model estimates can compound, resulting in increased errors over long horizons.
We approach this problem from the lens of Koopman theory, where the nonlinear dynamics of the environment can be linearized in a high-dimensional latent space. This allows us to efficiently parallelize the sequential problem of long-range prediction using convolution while accounting for the agent's action at every time step.
Our approach also enables stability analysis and better control over gradients through time. Taken together, these advantages result in significant improvement over the existing approaches, both in the efficiency and the accuracy of modeling dynamics over extended horizons. We also show that this model can be easily incorporated into dynamics modeling for model-based planning and model-free RL and report promising experimental results.
****
Aleksei Ustimenko,Aleksandr Beznosikov
Ito Diffusion Approximation of Universal Ito Chains for Sampling, Optimization and Boosting
In this work, we consider rather general and broad class of Markov chains, Ito chains, that look like Euler-Maryama discretization of some Stochastic Differential Equation. The chain we study is a unified framework for theoretical analysis. It comes with almost arbitrary isotropic and state-dependent noise instead of normal and state-independent one as in most related papers. Moreover, in our chain the drift and diffusion coefficient can be inexact in order to cover wide range of applications as Stochastic Gradient Langevin Dynamics, sampling, Stochastic Gradient Descent or Stochastic Gradient Boosting. We prove the bound in $\math

cal{W}_{2}$-distance between the laws of our Ito chain and corresponding differe
ntial equation. These results improve or cover most of the known estimates. And
for some particular cases, our analysis is the first.
****
Xinzhe Yuan,William de Vazelhes,Bin Gu,Huan Xiong
New Insight of Variance reduce in Zero-Order Hard-Thresholding: Mitigating Gradi
ent Error and Expansivity Contradictions
Hard-thresholding is an important type of algorithm in machine learning that is
used to solve ℓ_0 constrained optimization problems. However, the true gra
dient of the objective function can be difficult to access in certain scenarios,
 which normally can be approximated by zeroth-order (ZO) methods. SZOHT algorith
m is the only algorithm tackling ℓ_0 sparsity constraints with zeroth-order
 gradients so far. Unfortunately, SZOHT has a notable limitation on the number
 of random directions due to the inherent conflict between the deviation of ZO g
radients and the expansivity of the hard-thresholding operator.
This paper approaches this problem by considering the role of variance and provi
des a new insight into variance reduction: mitigating the unique conflicts betwe
en ZO gradients and hard-thresholding. Under this perspective, we propose a gen
eralized variance reduced ZO hard-thresholding algorithm as well as the generali
zed convergence analysis under standard assumptions. The theoretical results dem
onstrate the new algorithm eliminates the restrictions on the number of random d
irections, leading to improved convergence rates and broader applicability compa
red with SZOHT. Finally, we illustrate the utility of our method on a portfolio
 optimization problem as well as black-box adversarial attacks.
****
Jihao Andreas Lin,Shreyas Padhy,Javier Antoran,Austin Tripp,Alexander Terenin,Cs
aba Szepesvari,José Miguel Hernández-Lobato,David Janz
Stochastic Gradient Descent for Gaussian Processes Done Right
We study the optimisation problem associated with Gaussian process regression us
ing squared loss.
The most common approach to this problem is to apply an exact solver, such as co
njugate gradient descent, either directly on the problem or on a reduced-order v
ersion of it.
However, stochastic gradient descent has recently gained traction in the Gaussia
n process literature, driven largely by its successes in deep learning. In this
paper, we show that this approach when done right---by which we mean using speci
fic insights from the optimisation and kernel communities---is highly effective.
We thus introduce a particular stochastic dual gradient descent algorithm, conve
niently implementable with a few lines of code using any deep learning framework
.
We explain our design decisions by illustrating their advantage against alternat
ives with ablation studies.
We then show that the new method is highly competitive: our evaluations on stand
ard regression benchmarks and a Bayesian optimisation task set our approach apar
t from conjugate gradients, variational Gaussian process approximations, and a p
rior version of stochastic gradient descent tailored for Gaussian processes.
On a molecular binding affinity prediction task, our method places Gaussian proc
ess regression on par in terms of performance with graph neural networks.
****
Grégoire Mialon,Clémentine Fourrier,Thomas Wolf,Yann LeCun,Thomas Scialom
GAIA: a benchmark for General AI Assistants
We introduce GAIA, a benchmark for General AI Assistants that, if solved, would
represent a milestone in AI research. GAIA proposes real-world questions that re
quire a set of fundamental abilities such as reasoning, multi-modality handling,
 web browsing, and generally tool-use proficiency. Our questions allow simple, f
ast, and factual verification. GAIA questions are conceptually simple for humans
 yet challenging for most advanced AIs: we show that human respondents obtain 92
\% vs. 15\% for GPT-4 equipped with plugins. This notable performance disparity
contrasts with the recent trend of LLMs outperforming humans on tasks requiring
 professional skills in e.g. law or chemistry. GAIA's philosophy departs from th

e current trend in AI benchmarks suggesting to target tasks that are ever more difficult for humans. We posit that the advent of Artificial General Intelligence (AGI) hinges on a system's capability to exhibit similar robustness as the average human does on such questions. Using GAIA's methodology, we devise 466 questions and their answer. We release our questions while retaining answers to 300 of them to power a leader-board \href{https://huggingface.co/xxx}{hereby accessible}.
****
Richard Ngo,Lawrence Chan,Sören Mindermann
The Alignment Problem from a Deep Learning Perspective
AI systems based on deep learning have reached or surpassed human performance in a range of narrow domains. In coming years or decades, artificial general intelligence (AGI) may surpass human capabilities at many critical tasks. In this position paper, we examine the technical difficulty of fine-tuning hypothetical AGI systems based on pretrained deep models to pursue goals that are aligned with human interests. We argue that, if trained like today's most capable models, AGI systems could learn to act deceptively to receive higher reward, learn internally-represented goals which generalize beyond their fine-tuning distributions, and pursue those goals using power-seeking strategies. We review emerging evidence for these properties. AGIs with these properties would be difficult to align and may appear aligned even when they are not.
****
Jin Peng Zhou,Yuhuai Wu,Qiyang Li,Roger Baker Grosse
REFACTOR: Learning to Extract Theorems from Proofs
Human mathematicians are often good at recognizing modular and reusable theorems that make complex mathematical results within reach. In this paper, we propose a novel method called theoREm-from-prooF extrACTOR (REFACTOR) for training neural networks to mimic this ability in formal mathematical theorem proving. We show on a set of unseen proofs, REFACTOR is able to extract 19.6\% of the theorems that humans would use to write the proofs. When applying the model to the existing Metamath library, REFACTOR extracted 16 new theorems. With newly extracted theorems, we show that the existing proofs in the MetaMath database can be refactored. The new theorems are used very frequently after refactoring, with an average usage of 733.5 times, and help shorten the proof lengths. Lastly, we demonstrate that the prover trained on the new-theorem refactored dataset proves more test theorems and outperforms state-of-the-art baselines by frequently leveraging a diverse set of newly extracted theorems. Code can be found at https://github.com/jinpz/refactor.
****
Noa Cohen,Hila Manor,Yuval Bahat,Tomer Michaeli
From Posterior Sampling to Meaningful Diversity in Image Restoration
Image restoration problems are typically ill-posed in the sense that each degraded image can be restored in infinitely many valid ways. To accommodate this, many works generate a diverse set of outputs by attempting to randomly sample from the posterior distribution of natural images given the degraded input. Here we argue that this strategy is commonly of limited practical value because of the heavy tail of the posterior distribution. Consider for example inpainting a missing region of the sky in an image. Since there is a high probability that the missing region contains no object but clouds, any set of samples from the posterior would be entirely dominated by (practically identical) completions of sky. However, arguably, presenting users with only one clear sky completion, along with several alternative solutions such as airships, birds, and balloons, would better outline the set of possibilities. In this paper, we initiate the study of **meaningfully diverse** image restoration. We explore several post-processing approaches that can be combined with any diverse image restoration method to yield semantically meaningful diversity. Moreover, we propose a practical approach for allowing diffusion based image restoration methods to generate meaningfully diverse outputs, while incurring only negligent computational overhead. We conduct extensive user studies to analyze the proposed techniques, and find the strategy of reducing similarity between outputs to be significantly favorable over posterior

sampling. Code and examples are available on the [project's webpage](https://noa-cohen.github.io/MeaningfulDiversityInIR/).
****
Zhong Zheng,Fengyu Gao,Lingzhou Xue,Jing Yang
Federated Q-Learning: Linear Regret Speedup with Low Communication Cost
In this paper, we consider federated reinforcement learning for tabular episodic Markov Decision Processes (MDP) where, under the coordination of a central server, multiple agents collaboratively explore the environment and learn an optimal policy without sharing their raw data. While linear speedup in the number of agents has been achieved for some metrics, such as convergence rate and sample complexity, in similar settings, it is unclear whether it is possible to design a *model-free* algorithm to achieve linear *regret* speedup with low communication cost. We propose two federated Q-Learning algorithms termed as FedQ-Hoeffding and FedQ-Bernstein, respectively, and show that the corresponding total regrets achieve a linear speedup compared with their single-agent counterparts, while the communication cost scales logarithmically in the total number of time steps T. Those results rely on an event-triggered synchronization mechanism between the agents and the server, a novel step size selection when the server aggregates the local estimates of the state-action values to form the global estimates, and a set of new concentration inequalities to bound the sum of non-martingale differences. This is the first work showing that linear regret speedup and logarithmic communication cost can be achieved by model-free algorithms in federated reinforcement learning.
****
Zihao TANG,Zheqi Lv,Shengyu Zhang,Yifan Zhou,Xinyu Duan,Fei Wu,Kun Kuang
AuG-KD: Anchor-Based Mixup Generation for Out-of-Domain Knowledge Distillation
Due to privacy or patent concerns, a growing number of large models are released without granting access to their training data, making transferring their knowledge inefficient and problematic. In response, Data-Free Knowledge Distillation (DFKD) methods have emerged as direct solutions. However, simply adopting models derived from DFKD for real-world applications suffers significant performance degradation, due to the discrepancy between teachers' training data and real-world scenarios (student domain). The degradation stems from the portions of teachers' knowledge that are not applicable to the student domain. They are specific to the teacher domain and would undermine students' performance. Hence, selectively transferring teachers' appropriate knowledge becomes the primary challenge in DFKD. In this work, we propose a simple but effective method AuG-KD. It utilizes an uncertainty-guided and sample-specific anchor to align student-domain data with the teacher domain and leverages a generative method to progressively trade off the learning process between OOD knowledge distillation and domain-specific information learning via mixup learning. Extensive experiments in 3 datasets and 8 settings demonstrate the stability and superiority of our approach.
****
Haoran Xu,Young Jin Kim,Amr Sharaf,Hany Hassan Awadalla
A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models
Generative Large Language Models (LLMs) have achieved remarkable advancements in various NLP tasks. However, these advances have not been reflected in the translation task, especially those with moderate model sizes (i.e., 7B or 13B parameters), which still lag behind conventional supervised encoder-decoder translation models. Previous studies have attempted to improve the translation capabilities of these LLMs, but their gains have been limited. In this study, we propose a novel fine-tuning approach for LLMs that is specifically designed for the translation task, eliminating the need for the abundant parallel data that traditional translation models usually depend on.
Our approach consists of two fine-tuning stages: initial fine-tuning on monolingual data followed by subsequent fine-tuning on a small set of high-quality parallel data. We introduce the LLM developed through this strategy as **A**dvanced **L**anguage **M**odel-based tr**A**nslator (**ALMA**). Based on LLaMA-2 as our underlying model, our results show that the model can achieve an average improv

ement of more than 12 BLEU and 12 COMET over its zero-shot performance across 10
translation directions from the WMT'21 (2 directions) and WMT'22 (8 directions)
test datasets. The performance is significantly better than all prior work and
even superior to the NLLB-54B model \citep{nllb} and GPT-3.5-text-davinci-003, w
ith only 7B or 13B parameters. This method establishes the foundation for a nove
l training paradigm in machine translation.
****
Zixi Wei,Senlin Shu,Yuzhou Cao,Hongxin Wei,Bo An,Lei Feng
Consistent Multi-Class Classification from Multiple Unlabeled Datasets
Weakly supervised learning aims to construct effective predictive models from im
perfectly labeled data. The recent trend of weakly supervised learning has focus
ed on how to learn an accurate classifier from completely unlabeled data, given
little supervised information such as class priors. In this paper, we consider a
newly proposed weakly supervised learning problem called multi-class classifica
tion from multiple unlabeled datasets, where only multiple sets of unlabeled dat
a and their class priors (i.e., the proportions of each class) are provided for
training the classifier. To solve this problem, we first propose a classifier-co
nsistent method (CCM) based on a probability transition matrix. However, CCM can
not guarantee risk consistency and lacks of purified supervision information dur
ing training. Therefore, we further propose a risk-consistent method (RCM) that
progressively purifies supervision information during training by importance wei
ghting. We provide comprehensive theoretical analyses for our methods to demonst
rate the statistical consistency. Experimental results on multiple benchmark dat
asets and various prior matrices demonstrate the superiority of our proposed met
hods.
****
Min Xue,Artur Andrzejak,Marla Leuther
An interpretable error correction method for enhancing code-to-code translation
Transformer-based machine translation models currently dominate the field of mod
el-based program translation. However, these models fail to provide interpretati
ve support for the generated program translations. Moreover, researchers frequen
tly invest substantial time and computational resources in retraining models, ye
t the improvement in translation accuracy is quite limited.
To address these issues, we introduce a novel approach, $k\text{NN-ECD}$, which
combines k-nearest-neighbor search with a key-value error correction datastore
to overwrite the wrong translations of TransCoder-ST. This provides a decision-
making basis for interpreting the corrected translations. Building upon this, we
further propose $k\text{NN-ECS}_{m}$, a methodology that employs a distributed
structure with m sub-datastores connected in series, utilizing m diverse ex
perts for multi-round error correction. Additionally, we put forward a unified n
ame rule, encouraging the datastore to focus more on code logic and structure ra
ther than diverse rare identifiers. Our experimental results show that our appro
ach improves the translation accuracy from 68.9\% to 89.9\% of TransCoder-ST (fo
r translation from Java to Python). This error correction method augments progra
m translation, overcoming the inherent limitations of Transformer-based code tra
nslation models, such as resource-intensive retraining requirements and uninterp
retable outcomes.
****
Lukas Blecher,Guillem Cucurull,Thomas Scialom,Robert Stojnic
Nougat: Neural Optical Understanding for Academic Documents
Scientific knowledge is predominantly stored in books and scientific journals, o
ften in the form of PDFs. However, the PDF format leads to a loss of semantic in
formation, particularly for mathematical expressions. We propose Nougat (Neural
Optical Understanding for Academic Documents), a Visual Transformer model that p
erforms an Optical Character Recognition (OCR) task for processing scientific do
cuments into a markup language, and demonstrate the effectiveness of our model o
n a new dataset of scientific documents. The proposed approach offers a promisin
g solution to enhance the accessibility of scientific knowledge in the digital a
ge, by bridging the gap between human- readable documents and machine-readable t
ext. We release the models and code to accelerate future work on scientific text

recognition.
****
Eliya Segev,Maya Alroy,Ronen Katsir,Noam Wies,Ayana Shenhav,Yael Sapir Ben-Oren, David Zar,Oren Tadmor,Jacob Bitterman,Amnon Shashua,Tal Rosenwein
Align With Purpose: Optimize Desired Properties in CTC Models with a General Plug-and-Play Framework
Connectionist Temporal Classification (CTC) is a widely used criterion for training supervised sequence-to-sequence (seq2seq) models. It learns the alignments between the input and output sequences, by marginalizing over the perfect alignments (that yield the ground truth), at the expense of the imperfect ones.
This dichotomy, and in particular the equal treatment of all perfect alignments, results in a lack of controllability over the predicted alignments.
This controllability is essential for capturing properties that hold significance in real-world applications.
Here we propose Align With Purpose (AWP), a general Plug-and-Play framework for enhancing a desired property in models trained with the CTC criterion. We do that by complementing the CTC loss with an additional loss term that prioritizes alignments according to a desired property. AWP does not require any intervention in the CTC loss function, and allows to differentiate between both perfect and imperfect alignments for a variety of properties. We apply our framework in the domain of Automatic Speech Recognition (ASR) and show its generality in terms of property selection, architectural choice, and scale of training dataset (up to 280,000 hours). To demonstrate the effectiveness of our framework, we apply it to two unrelated properties: token emission time for latency optimization and word error rate (WER). For the former, we report an improvement of up to 590ms in latency optimization with a minor reduction in WER, and for the latter, we report a relative improvement of 4.5\% in WER over the baseline models. To the best of our knowledge, these applications have never been demonstrated to work on this scale of data. Notably, our method can be easily implemented using only a few lines of code and can be extended to other alignment-free loss functions and to domains other than ASR.
****
Sunghwan Hong,Seokju Cho,Seungryong Kim,Stephen Lin
Unifying Feature and Cost Aggregation with Transformers for Semantic and Visual Correspondence
This paper introduces a Transformer-based integrative feature and cost aggregation network designed for dense matching tasks. In the context of dense matching, many works benefit from one of two forms of aggregation: feature aggregation, which pertains to the alignment of similar features, or cost aggregation, a procedure aimed at instilling coherence in the flow estimates across neighboring pixels. In this work, we first show that feature aggregation and cost aggregation exhibit distinct characteristics and reveal the potential for substantial benefits stemming from the judicious use of both aggregation processes. We then introduce a simple yet effective architecture that harnesses self- and cross-attention mechanisms to show that our approach unifies feature aggregation and cost aggregation and effectively harnesses the strengths of both techniques. Within the proposed attention layers, the features and cost volume both complement each other, and the attention layers are interleaved through a coarse-to-fine design to further promote accurate correspondence estimation. Finally at inference, our network produces multi-scale predictions, computes their confidence scores, and selects the most confident flow for final prediction. Our framework is evaluated on standard benchmarks for semantic matching, and also applied to geometric matching, where we show that our approach achieves significant improvements compared to existing methods.
****
Zhoubo Li,Ningyu Zhang,Yunzhi Yao,Mengru Wang,Xi Chen,Huajun Chen
Unveiling the Pitfalls of Knowledge Editing for Large Language Models
As the cost associated with fine-tuning Large Language Models (LLMs) continues to rise, recent research efforts have pivoted towards developing methodologies to edit implicit knowledge embedded within LLMs. Yet, there's still a dark cloud l

ingering overhead -- will knowledge editing trigger butterfly effect? since it i
s still unclear whether knowledge editing might introduce side effects that pose
 potential risks or not. This paper pioneers the investigation into the potentia
l pitfalls associated with knowledge editing for LLMs. To achieve this, we intro
duce new benchmark datasets and propose innovative evaluation metrics. Our resul
ts underline two pivotal concerns: (1) Knowledge Conflict: Editing groups of fac
ts that logically clash can magnify the inherent inconsistencies in LLMs—a facet
 neglected by previous methods. (2) Knowledge Distortion: Altering parameters wi
th the aim of editing factual knowledge can irrevocably warp the innate knowledg
e structure of LLMs. Experimental results vividly demonstrate that knowledge edi
ting might inadvertently cast a shadow of unintended consequences on LLMs, which
 warrant attention and efforts for future works. Code and data are available at
https://github.com/zjunlp/PitfallsKnowledgeEditing.
****
Zishun Yu,Yunzhe Tao,Liyu Chen,Tao Sun,Hongxia Yang
$\mathcal{B}$-Coder: Value-Based Deep Reinforcement Learning for Program Synthes
is
Program synthesis aims to create accurate, executable programs from problem spec
ifications, specifically from natural language descriptions in our context.
Recent studies have leveraged the power of reinforcement learning (RL) in conjun
ction with large language models (LLMs), significantly enhancing code generation
 capabilities. The application of RL focuses on directly optimizing for function
al correctness, offering an advantage over conventional supervised methods.
Despite policy-based RL methods dominating the literature on RL for program synt
hesis, the nature of program synthesis tasks hints at a natural alignment with v
alue-based methods.
This stems from the rich collection of off-policy programs, including those deve
loped by human programmers and also historical samples, coupled with the straigh
tforward verification of generated programs through automated unit testing, mean
ing rewards are easy to obtain.
Diverging from the dominant use of policy-based algorithms, our work explores th
e feasibility of value-based approaches, leading to the development of our $\mat
hcal{B}$-Coder (pronounced Bellman coder).
Yet, training value-based methods presents challenges due to the enormous search
 space inherent to program synthesis.
To this end, we introduce an initialization protocol for RL agents utilizing pre
-trained LMs and a conservative Bellman operator to reduce training complexities
.
Moreover, we demonstrate how to leverage the learned value functions as a dual s
trategy to post-process generated programs.
Our empirical evaluations demonstrated $\mathcal{B}$-Coder's capability in achie
ving state-of-the-art performance when compared to policy-based methods.
Remarkably, this achievement is reached with minimal reward engineering effort,
highlighting the effectiveness of value-based RL, independent of reward designs.
****
Shi Fu,Fengxiang He,Xinmei Tian,Dacheng Tao
Convergence of Bayesian Bilevel Optimization
This paper presents the first theoretical guarantee for Bayesian bilevel optimiz
ation (BBO) that we term for the prevalent bilevel framework combining Bayesian
optimization at the outer level to tune hyperparameters, and the inner-level sto
chastic gradient descent (SGD) for training the model. We prove sublinear regret
 bounds suggesting simultaneous convergence of the inner-level model parameters
and outer-level hyperparameters to optimal configurations for generalization cap
ability. A pivotal, technical novelty in the proofs is modeling the excess risk
of the SGD-trained parameters as evaluation noise during Bayesian optimization.
Our theory implies the inner unit horizon, defined as the number of SGD iteratio
ns, shapes the convergence behavior of BBO. This suggests practical guidance on
configuring the inner unit horizon to enhance training efficiency and model perf
ormance.
****

Jose Javier Gonzalez Ortiz,John Guttag,Adrian V Dalca
Magnitude Invariant Parametrizations Improve Hypernetwork Learning
Hypernetworks, neural networks that predict the parameters of another neural net
work, are powerful models that have been successfully used in diverse applicatio
ns from image generation to multi-task learning. Unfortunately, existing hyperne
tworks are often challenging to train. Training typically converges far more slo
wly than for non-hypernetwork models, and the rate of convergence can be very se
nsitive to hyperparameter choices. In this work, we identify a fundamental and p
reviously unidentified problem that contributes to the challenge of training hyp
ernetworks: a magnitude proportionality between the inputs and outputs of the hy
pernetwork. We demonstrate both analytically and empirically that this can lead
to unstable optimization, thereby slowing down convergence, and sometimes even p
reventing any learning. We present a simple solution to this problem using a rev
ised hypernetwork formulation that we call Magnitude Invariant Parametrizations
(MIP). We demonstrate the proposed solution on several hypernetwork tasks, where
 it consistently stabilizes training and achieves faster convergence. Furthermor
e, we perform a comprehensive ablation study including choices of activation fun
ction, normalization strategies, input dimensionality, and hypernetwork architec
ture; and find that MIP improves training in all scenarios. We provide easy-to-u
se code that can turn existing networks into MIP-based hypernetworks.
****
Seyed Saman Saboksayr,Gonzalo Mateos,Mariano Tepper
CoLiDE: Concomitant Linear DAG Estimation
We deal with the combinatorial problem of learning directed acyclic graph (DAG)
structure from observational data adhering to a linear structural equation model
 (SEM). Leveraging advances in differentiable, nonconvex characterizations of ac
yclicity, recent efforts have advocated a continuous constrained optimization pa
radigm to efficiently explore the space of DAGs. Most existing methods employ la
sso-type score functions to guide this search, which (i) require expensive penal
ty parameter retuning when the $\textit{unknown}$ SEM noise variances change acr
oss problem instances; and (ii) implicitly rely on limiting homoscedasticity ass
umptions. In this work, we propose a new convex score function for sparsity-awar
e learning of linear DAGs, which incorporates concomitant estimation of scale an
d thus effectively decouples the sparsity parameter from noise levels. Regulariz
ation via a smooth, nonconvex acyclicity penalty term yields CoLiDE ($\textbf{Co
}$ncomitant $\textbf{Li}$near $\textbf{D}$AG $\textbf{E}$stimation), a regressio
n-based criterion amenable to efficient gradient computation and closed-form est
imation of exogenous noise levels in heteroscedastic scenarios. Our algorithm ou
tperforms state-of-the-art methods without incurring added complexity, especiall
y when the DAGs are larger and the noise level profile is heterogeneous. We also
 find CoLiDE exhibits enhanced stability manifested via reduced standard deviati
ons in several domain-specific metrics, underscoring the robustness of our novel
 linear DAG estimator.
****
Juan Elenter,Luiz F. O. Chamon,Alejandro Ribeiro
Near-Optimal Solutions of Constrained Learning Problems
With the widespread adoption of machine learning systems, the need to curtail th
eir behavior has become increasingly apparent. This is evidenced by recent advan
cements towards developing models that satisfy robustness, safety, and fairness
requirements. These requirements can be imposed (with generalization guarantees)
 by formulating constrained learning problems that can then be tackled by dual a
scent algorithms. Yet, though these algorithms converge in objective value, even
 in non-convex settings, they cannot guarantee that their outcome is feasible. D
oing so requires randomizing over all iterates, which is impractical in virtuall
y any modern applications. Still, final iterates have been observed to perform w
ell in practice. In this work, we address this gap between theory and practice b
y characterizing the constraint violation of Lagrangian minimizers associated wi
th optimal dual variables, despite lack of convexity. To do this, we leverage th
e fact that non-convex, finite-dimensional constrained learning problems can be
seen as parametrizations of convex, functional problems. Our results show that r

ich parametrizations effectively mitigate the issue of feasibility in dual metho ds, shedding light on prior empirical successes of dual learning. We illustrate our findings in fair learning tasks.
****
Dyah Adila,Changho Shin,Linrong Cai,Frederic Sala
Zero-Shot Robustification of Zero-Shot Models
Zero-shot inference is a powerful paradigm that enables the use of large pretrai ned models for downstream classification tasks without further training. However , these models are vulnerable to inherited biases that can impact their performa nce. The traditional solution is fine-tuning, but this undermines the key advant age of pretrained models, which is their ability to be used out-of-the-box. We p ropose RoboShot, a method that improves the robustness of pretrained model embed dings in a fully zero-shot fashion. First, we use language models (LMs) to obtai n useful insights from task descriptions. These insights are embedded and used t o remove harmful and boost useful components in embeddings---without any supervi sion. Theoretically, we provide a simple and tractable model for biases in zero-shot embeddings and give a result characterizing under what conditions our appro ach can boost performance. Empirically, we evaluate RoboShot on nine image and N LP classification tasks and show an average improvement of 15.98% over several z ero-shot baselines. Additionally, we demonstrate that RoboShot is compatible wit h a variety of pretrained and language models and propose a way to further boost performance with a zero-shot adaptation variant.
****
Tai-Yu Pan,Chenyang Ma,Tianle Chen,Cheng Perng Phoo,Katie Z Luo,Yurong You,Mark Campbell,Kilian Q Weinberger,Bharath Hariharan,Wei-Lun Chao
Pre-training LiDAR-based 3D Object Detectors through Colorization
Accurate 3D object detection and understanding for self-driving cars heavily rel ies on LiDAR point clouds, necessitating large amounts of labeled data to train. In this work, we introduce an innovative pre-training approach, Grounded Point Colorization (GPC), to bridge the gap between data and labels by teaching the mo del to colorize LiDAR point clouds, equipping it with valuable semantic cues. To tackle challenges arising from color variations and selection bias, we incorpor ate color as "context" by providing ground-truth colors as hints during coloriza tion.
Experimental results on the KITTI and Waymo datasets demonstrate GPC's remarkabl e effectiveness. Even with limited labeled data, GPC significantly improves fine -tuning performance; notably, on just 20% of the KITTI dataset, GPC outperforms training from scratch with the entire dataset.
In sum, we introduce a fresh perspective on pre-training for 3D object detection , aligning the objective with the model's intended role and ultimately advancing the accuracy and efficiency of 3D object detection for autonomous vehicles.
****
Zhenting Wang,Chen Chen,Lingjuan Lyu,Dimitris N. Metaxas,Shiqing Ma
DIAGNOSIS: Detecting Unauthorized Data Usages in Text-to-image Diffusion Models
Recent text-to-image diffusion models have shown surprising performance in gener ating high-quality images. However, concerns have arisen regarding the unauthori zed data usage during the training or fine-tuning process. One example is when a model trainer collects a set of images created by a particular artist and attem pts to train a model capable of generating similar images without obtaining perm ission and giving credit to the artist. To address this issue, we propose a meth od for detecting such unauthorized data usage by planting the injected memorizat ion into the text-to-image diffusion models trained on the protected dataset. Sp ecifically, we modify the protected images by adding unique contents on these im ages using stealthy image warping functions that are nearly imperceptible to hum ans but can be captured and memorized by diffusion models. By analyzing whether the model has memorized the injected content (i.e., whether the generated images are processed by the injected post-processing function), we can detect models t hat had illegally utilized the unauthorized data. Experiments on Stable Diffusio n and VQ Diffusion with different model training or fine-tuning methods (i.e, Lo RA, DreamBooth, and standard training) demonstrate the effectiveness of our prop

osed method in detecting unauthorized data usages. Code: https://github.com/Zhen
tingWang/DIAGNOSIS.
****
Annie S Chen,Yoonho Lee,Amrith Setlur,Sergey Levine,Chelsea Finn
Project and Probe: Sample-Efficient Adaptation by Interpolating Orthogonal Featu
res
Transfer learning with a small amount of target data is an effective and common
approach to adapting a pre-trained model to distribution shifts. In some situati
ons, target data labels may be expensive to obtain, so we may only have access t
o a limited number of target data points. To make the most of a very small targe
t dataset, we propose a lightweight, sample-efficient approach that learns a div
erse set of features and adapts to a target distribution by interpolating these
features. Our approach, Project and Probe (Pro^2), first learns a linear proje
ction that maps a pre-trained embedding onto orthogonal directions while being p
redictive of labels in the source dataset. The goal of this step is to learn a v
ariety of predictive features, so that at least some of them remain useful after
 distribution shift. Pro^2 then learns a linear classifier on top of these pro
jected features using a small target dataset. Theoretically, we find that Pro$^2
$ results in more sample-efficient generalization by inducing a favorable bias-v
ariance tradeoff. Our experiments on four datasets, with multiple distribution s
hift settings for each, show that Pro^2 improves performance by 5-15% when giv
en limited target data compared to prior methods such as standard linear probing
.
****
Haoyu Lu,Yuqi Huo,Guoxing Yang,Zhiwu Lu,Wei Zhan,Masayoshi Tomizuka,Mingyu Ding
UniAdapter: Unified Parameter-Efficient Transfer Learning for Cross-modal Modeli
ng
Large-scale vision-language pre-trained models have shown promising transferabil
ity to various downstream tasks. As the size of these foundation models and the
number of downstream tasks grow, the standard full fine-tuning paradigm becomes
unsustainable due to heavy computational and storage costs. This paper proposes
UniAdapter, which unifies unimodal and multimodal adapters for parameter-efficie
nt cross-modal adaptation on pre-trained vision-language models. Specifically, a
dapters are distributed to different modalities and their interactions, with the
 total number of tunable parameters reduced by partial weight sharing. The unifi
ed and knowledge-sharing design enables powerful cross-modal representations tha
t can benefit various downstream tasks, requiring only 1.0%-2.0% tunable paramet
ers of the pre-trained model. Extensive experiments on 7 cross-modal downstream
benchmarks (including video-text retrieval, image-text retrieval, VideoQA, VQA a
nd Caption) show that in most cases, UniAdapter not only outperforms the state-o
f-the-arts, but even beats the full fine-tuning strategy. Particularly, on the M
SRVTT retrieval task, UniAdapter achieves 49.7% recall@1 with 2.2% model paramet
ers, outperforming the latest competitors by 2.0%. The code and models are avail
able at https://github.com/RERV/UniAdapter.
****
Vimal Thilak,Chen Huang,Omid Saremi,Laurent Dinh,Hanlin Goh,Preetum Nakkiran,Jos
hua M. Susskind,Etai Littwin
LiDAR: Sensing Linear Probing Performance in Joint Embedding SSL Architectures
Joint embedding (JE) architectures have emerged as a promising avenue for ac-
quiring transferable data representations. A key obstacle to using JE methods,
however, is the inherent challenge of evaluating learned representations without
access to a downstream task, and an annotated dataset. Without efficient and re-
liable evaluation, it is difficult to iterate on architectural and training choi
ces for
JE methods. In this paper, we introduce LiDAR (Linear Discriminant Analysis
Rank), a metric designed to measure the quality of representations within JE arc
hi-
tectures. Our metric addresses several shortcomings of recent approaches based
on feature covariance rank by discriminating between informative and uninforma-
tive features. In essence, LiDAR quantifies the rank of the Linear Discriminant

Analysis (LDA) matrix associated with the surrogate SSL task—a measure that intuitively captures the information content as it pertains to solving the SSL task.
We empirically demonstrate that LiDAR significantly surpasses naive rank based approaches in its predictive power of optimal hyperparameters. Our proposed criterion presents a more robust and intuitive means of assessing the quality of representations within JE architectures, which we hope facilitates broader adoption of these powerful techniques in various domains.
****
Aniket Rajiv Didolkar,Anirudh Goyal,Yoshua Bengio
Cycle Consistency Driven Object Discovery
Developing deep learning models that effectively learn object-centric representations, akin to human cognition, remains a challenging task. Existing approaches facilitate object discovery by representing objects as fixed-size vectors, called ``slots'' or ``object files''. While these approaches have shown promise in certain scenarios, they still exhibit certain limitations. First, they rely on architectural priors which can be unreliable and usually require meticulous engineering to identify the correct objects. Second, there has been a notable gap in investigating the practical utility of these representations in downstream tasks. To address the first limitation, we introduce a method that explicitly optimizes the constraint that each object in a scene should be associated with a distinct slot. We formalize this constraint by introducing consistency objectives which are cyclic in nature. By integrating these consistency objectives into various existing slot-based object-centric methods, we showcase substantial improvements in object-discovery performance. These enhancements consistently hold true across both synthetic and real-world scenes, underscoring the effectiveness and adaptability of the proposed approach. To tackle the second limitation, we apply the learned object-centric representations from the proposed method to two downstream reinforcement learning tasks, demonstrating considerable performance enhancements compared to conventional slot-based and monolithic representation learning methods. Our results suggest that the proposed approach not only improves object discovery, but also provides richer features for downstream tasks.
****
Zihan Zhong,Zhiqiang Tang,Tong He,Haoyang Fang,Chun Yuan
Convolution Meets LoRA: Parameter Efficient Finetuning for Segment Anything Model
The Segment-Anything Model (SAM) stands as a foundational framework for image segmentation. While it exhibits remarkable zero-shot generalization in typical scenarios, its advantage diminishes when applied to specialized domains like medical imagery and remote sensing. To address this limitation, this paper introduces Conv-LoRA, a simple yet effective parameter-efficient fine-tuning approach. By integrating ultra-lightweight convolutional parameters into Low-Rank Adaptation (LoRA), Conv-LoRA can inject image-related inductive biases into the plain ViT encoder, further reinforcing SAM’s local prior assumption. Notably, Conv-LoRA not only preserves SAM’s extensive segmentation knowledge but also revives its capacity of learning high-level image semantics, which is constrained by SAM’s foreground-background segmentation pretraining. Comprehensive experimentation across diverse benchmarks spanning multiple domains underscores Conv-LoRA’s superiority in adapting SAM to real-world semantic segmentation tasks.
****
Yuxiang Tuo,Wangmeng Xiang,Jun-Yan He,Yifeng Geng,Xuansong Xie
AnyText: Multilingual Visual Text Generation and Editing
Diffusion model based Text-to-Image has achieved impressive achievements recently. Although current technology for synthesizing images is highly advanced and capable of generating images with high fidelity, it is still possible to give the show away when focusing on the text area in the generated image, as synthesized text often contains blurred, unreadable, or incorrect characters, making visual text generation one of the most challenging issues in this field. To address this issue, we introduce AnyText, a diffusion-based multilingual visual text genera

tion and editing model, that focuses on rendering accurate and coherent text in the image. AnyText comprises a diffusion pipeline with two primary elements: an auxiliary latent module and a text embedding module. The former uses inputs like text glyph, position, and masked image to generate latent features for text generation or editing. The latter employs an OCR model for encoding stroke data as embeddings, which blend with image caption embeddings from the tokenizer to generate texts that seamlessly integrate with the background. We employed text-control diffusion loss and text perceptual loss for training to further enhance writing accuracy. AnyText can write characters in multiple languages, to the best of our knowledge, this is the first work to address multilingual visual text generation. It is worth mentioning that AnyText can be plugged into existing diffusion models from the community for rendering or editing text accurately. After conducting extensive evaluation experiments, our method has outperformed all other approaches by a significant margin. Additionally, we contribute the first large-scale multilingual text images dataset, AnyWord-3M, containing 3 million image-text pairs with OCR annotations in multiple languages. Based on AnyWord-3M dataset, we propose AnyText-benchmark for the evaluation of visual text generation accuracy and quality. Our project will be open-sourced soon to improve and promote the development of text generation technology.
****
Emmeran Johnson,Ciara Pike-Burke,Patrick Rebeschini
Sample-Efficiency in Multi-Batch Reinforcement Learning: The Need for Dimension-Dependent Adaptivity
We theoretically explore the relationship between sample-efficiency and adaptivity in reinforcement learning. An algorithm is sample-efficient if it uses a number of queries n to the environment that is polynomial in the dimension d of the problem. Adaptivity refers to the frequency at which queries are sent and feedback is processed to update the querying strategy. To investigate this interplay, we employ a learning framework that allows sending queries in K batches, with feedback being processed and queries updated after each batch. This model encompasses the whole adaptivity spectrum, ranging from non-adaptive `offline' ($K=1$) to fully adaptive ($K=n$) scenarios, and regimes in between. For the problems of policy evaluation and best-policy identification under d-dimensional linear function approximation, we establish $\Omega(\log \log d)$ lower bounds on the number of batches K required for sample-efficient algorithms with $n = O(poly(d))$ queries. Our results show that just having adaptivity ($K>1$) does not necessarily guarantee sample-efficiency. Notably, the adaptivity-boundary for sample-efficiency is not between offline reinforcement learning ($K=1$), where sample-efficiency was known to not be possible, and adaptive settings. Instead, the boundary lies between different regimes of adaptivity and depends on the problem dimension.
****
Weiyu Li,Rui Chen,Xuelin Chen,Ping Tan
SweetDreamer: Aligning Geometric Priors in 2D diffusion for Consistent Text-to-3D
It is inherently ambiguous to lift 2D results from pre-trained diffusion models to a 3D world for text-to-3D generation. 2D diffusion models solely learn view-agnostic priors and thus lack 3D knowledge during the lifting, leading to the multi-view inconsistency problem. We find that this problem primarily stems from geometric inconsistency, and avoiding misplaced geometric structures substantially mitigates the problem in the final outputs. Therefore, we improve the consistency by aligning the 2D geometric priors in diffusion models with well-defined 3D shapes during the lifting, addressing the vast majority of the problem. This is achieved by fine-tuning the 2D diffusion model to be viewpoint-aware and to produce view-specific coordinate maps of canonically oriented 3D objects. In our process, only coarse 3D information is used for aligning. This “coarse” alignment not only resolves the multi-view inconsistency in geometries but also retains the ability in 2D diffusion models to generate detailed and diversified high-quality ob-jects unseen in the 3D datasets. Furthermore, our aligned geometric priors (AGP) are generic and can be seamlessly integrated into various state-of-the-art

pipelines, obtaining high generalizability in terms of unseen shapes and visual appearance while greatly alleviating the multi-view inconsistency problem.
****
Long Lian,Baifeng Shi,Adam Yala,Trevor Darrell,Boyi Li
LLM-grounded Video Diffusion Models
Text-conditioned diffusion models have emerged as a promising tool for neural video generation. However, current models still struggle with intricate spatiotemporal prompts and often generate restricted or incorrect motion. To address these limitations, we introduce LLM-grounded Video Diffusion (LVD). Instead of directly generating videos from the text inputs, LVD first leverages a large language model (LLM) to generate dynamic scene layouts based on the text inputs and subsequently uses the generated layouts to guide a diffusion model for video generation. We show that LLMs are able to understand complex spatiotemporal dynamics from text alone and generate layouts that align closely with both the prompts and the object motion patterns typically observed in the real world. We then propose to guide video diffusion models with these layouts by adjusting the attention maps. Our approach is training-free and can be integrated into any video diffusion model that admits classifier guidance. Our results demonstrate that LVD significantly outperforms its base video diffusion model and several strong baseline methods in faithfully generating videos with the desired attributes and motion patterns.
****
Jiawei Ge,Shange Tang,Jianqing Fan,Cong Ma,Chi Jin
Maximum Likelihood Estimation is All You Need for Well-Specified Covariate Shift
A key challenge of modern machine learning systems is to achieve Out-of-Distribution (OOD) generalization---generalizing to target data whose distribution differs from that of source data. Despite its significant importance, the fundamental question of ``what are the most effective algorithms for OOD generalization'' remains open even under the standard setting of covariate shift.
This paper addresses this fundamental question by proving that, surprisingly, classical Maximum Likelihood Estimation (MLE) purely using source data (without any modification) achieves the *minimax* optimality for covariate shift under the *well-specified* setting. That is, *no* algorithm performs better than MLE in this setting (up to a constant factor), justifying MLE is all you need.
Our result holds for a very rich class of parametric models, and does not require any boundedness condition on the density ratio. We illustrate the wide applicability of our framework by instantiating it to three concrete examples---linear regression, logistic regression, and phase retrieval. This paper further complement the study by proving that, under the *misspecified setting*, MLE is no longer the optimal choice, whereas Maximum Weighted Likelihood Estimator (MWLE) emerges as minimax optimal in certain scenarios.
****
Xinyu Shi,Jianhao Ding,Zecheng Hao,Zhaofei Yu
Towards Energy Efficient Spiking Neural Networks: An Unstructured Pruning Framework
Spiking Neural Networks (SNNs) have emerged as energy-efficient alternatives to Artificial Neural Networks (ANNs) when deployed on neuromorphic chips. While recent studies have demonstrated the impressive performance of deep SNNs on challenging tasks, their energy efficiency advantage has been diminished. Existing methods targeting energy consumption reduction do not fully exploit sparsity, whereas powerful pruning methods can achieve high sparsity but are not directly targeted at energy efficiency, limiting their effectiveness in energy saving. Furthermore, none of these works fully exploit the sparsity of neurons or the potential for unstructured neuron pruning in SNNs. In this paper, we propose a novel pruning framework that combines unstructured weight pruning with unstructured neuron pruning to maximize the utilization of the sparsity of neuromorphic computing, thereby enhancing energy efficiency. To the best of our knowledge, this is the first application of unstructured neuron pruning to deep SNNs. Experimental results demonstrate that our method achieves impressive energy efficiency gains. The sparse network pruned by our method with only 0.63\% remaining connections ca

n achieve a remarkable 91 times increase in energy efficiency compared to the original dense network, requiring only 8.5M SOPs for inference, with merely 2.19\% accuracy loss on the CIFAR-10 dataset. Our work suggests that deep and dense SNNs exhibit high redundancy in energy consumption, highlighting the potential for targeted SNN sparsification to save energy.
****
Mingde Zhao,Safa Alver,Harm van Seijen,Romain Laroche,Doina Precup,Yoshua Bengio
Consciousness-Inspired Spatio-Temporal Abstractions for Better Generalization in Reinforcement Learning
Inspired by human conscious planning, we propose Skipper, a model-based reinforcement learning framework utilizing spatio-temporal abstractions to generalize better in novel situations. It automatically decomposes the given task into smaller, more manageable subtasks, and thus enables sparse decision-making and focused computation on the relevant parts of the environment. The decomposition relies on the extraction of an abstracted proxy problem represented as a directed graph, in which vertices and edges are learned end-to-end from hindsight. Our theoretical analyses provide performance guarantees under appropriate assumptions and establish where our approach is expected to be helpful. Generalization-focused experiments validate Skipper’s significant advantage in zero-shot generalization, compared to some existing state-of-the-art hierarchical planning methods.
****
Nan Chen,Zemin Liu,Bryan Hooi,Bingsheng He,Rizal Fathony,Jun Hu,Jia Chen
Consistency Training with Learnable Data Augmentation for Graph Anomaly Detection with Limited Supervision
Graph Anomaly Detection (GAD) has surfaced as a significant field of research, predominantly due to its substantial influence in production environments. Although existing approaches for node anomaly detection have shown effectiveness, they have yet to fully address two major challenges: operating in settings with limited supervision and managing class imbalance effectively. In response to these challenges, we propose a novel model, ConsisGAD, which is tailored for GAD in scenarios characterized by limited supervision and is anchored in the principles of consistency training. Under limited supervision, ConsisGAD effectively leverages the abundance of unlabeled data for consistency training by incorporating a novel learnable data augmentation mechanism, thereby introducing controlled noise into the dataset. Moreover, ConsisGAD takes advantage of the variance in homophily distribution between normal and anomalous nodes to craft a simplified GNN backbone, enhancing its capability to distinguish effectively between these two classes. Comprehensive experiments on several benchmark datasets validate the superior performance of ConsisGAD in comparison to state-of-the-art baselines. Our code is available at https://github.com/Xtra-Computing/ConsisGAD.
****
Ahmad Bdeir,Kristian Schwethelm,Niels Landwehr
Fully Hyperbolic Convolutional Neural Networks for Computer Vision
Real-world visual data exhibit intrinsic hierarchical structures that can be represented effectively in hyperbolic spaces. Hyperbolic neural networks (HNNs) are a promising approach for learning feature representations in such spaces. However, current HNNs in computer vision rely on Euclidean backbones and only project features to the hyperbolic space in the task heads, limiting their ability to fully leverage the benefits of hyperbolic geometry. To address this, we present HCNN, a fully hyperbolic convolutional neural network (CNN) designed for computer vision tasks. Based on the Lorentz model, we generalize fundamental components of CNNs and propose novel formulations of the convolutional layer, batch normalization, and multinomial logistic regression. Experiments on standard vision tasks demonstrate the promising performance of our HCNN framework in both hybrid and fully hyperbolic settings. Overall, we believe our contributions provide a foundation for developing more powerful HNNs that can better represent complex structures found in image data. Our code is publicly available at https://github.com/kschwethelm/HyperbolicCV.
****
Satwik Bhattamishra,Arkil Patel,Phil Blunsom,Varun Kanade

Understanding In-Context Learning in Transformers and LLMs by Learning to Learn Discrete Functions
In order to understand the in-context learning phenomenon, recent works have adopted a stylized experimental framework and demonstrated that Transformers can match the performance of gradient-based learning algorithms for various classes of real-valued functions. However, the limitations of Transformers in implementing learning algorithms, and their ability to learn other forms of algorithms are not well understood. Additionally, the degree to which these capabilities are confined to attention-based models is unclear. Furthermore, it remains to be seen whether the insights derived from these stylized settings can be extrapolated to pretrained Large Language Models (LLMs). In this work, we take a step towards answering these questions by demonstrating the following: (a) On a test-bed with a variety of Boolean function classes, we find that Transformers can nearly match the optimal learning algorithm for 'simpler' tasks, while their performance deteriorates on more 'complex' tasks. Additionally, we find that certain attention-free models perform (almost) identically to Transformers on a range of tasks. (b) When provided a *teaching sequence*, i.e. a set of examples that uniquely identifies a function in a class, we show that Transformers learn more sample-efficiently. Interestingly, our results show that Transformers can learn to implement *two distinct* algorithms to solve a *single* task, and can adaptively select the more sample-efficient algorithm depending on the sequence of in-context examples. (c) Lastly, we show that extant LLMs, e.g. LLaMA-2, GPT-4, can compete with nearest-neighbor baselines on prediction tasks that are guaranteed to not be in their training set.
****
Yuhta Takida,Masaaki Imaizumi,Takashi Shibuya,Chieh-Hsin Lai,Toshimitsu Uesaka,Naoki Murata,Yuki Mitsufuji
SAN: Inducing Metrizability of GAN with Discriminative Normalized Linear Layer
Generative adversarial networks (GANs) learn a target probability distribution by optimizing a generator and a discriminator with minimax objectives. This paper addresses the question of whether such optimization actually provides the generator with gradients that make its distribution close to the target distribution. We derive *metrizable conditions*, sufficient conditions for the discriminator to serve as the distance between the distributions, by connecting the GAN formulation with the concept of sliced optimal transport. Furthermore, by leveraging these theoretical results, we propose a novel GAN training scheme called the Slicing Adversarial Network (SAN). With only simple modifications, a broad class of existing GANs can be converted to SANs. Experiments on synthetic and image datasets support our theoretical results and the effectiveness of SAN as compared to the usual GANs. We also apply SAN to StyleGAN-XL, which leads to a state-of-the-art FID score amongst GANs for class conditional generation on CIFAR10 and ImageNet 256$\times$256. The code is available at https://github.com/sony/san.
****
Marcel Binz,Eric Schulz
Turning large language models into cognitive models
Large language models are powerful systems that excel at many tasks, ranging from translation to mathematical reasoning. Yet, at the same time, these models often show unhuman-like characteristics. In the present paper, we address this gap and ask whether large language models can be turned into cognitive models. We find that -- after finetuning them on data from psychological experiments -- these models offer accurate representations of human behavior, even outperforming traditional cognitive models in two decision-making domains. In addition, we show that their representations contain the information necessary to model behavior on the level of individual subjects. Finally, we demonstrate that finetuning on multiple tasks enables large language models to predict human behavior in a previously unseen task. Taken together, these results suggest that large, pre-trained models can be adapted to become models of human cognition, which opens up future research directions toward building more general cognitive models.
****
Zhiqian Lan,Yuxuan Jiang,Yao Mu,Chen Chen,Shengbo Eben Li

SEPT: Towards Efficient Scene Representation Learning for Motion Prediction
Motion prediction is crucial for autonomous vehicles to operate safely in complex traffic environments. Extracting effective spatiotemporal relationships among traffic elements is key to accurate forecasting. Inspired by the successful practice of pretrained large language models, this paper presents SEPT, a modeling framework that leverages self-supervised learning to develop powerful spatiotemporal understanding for complex traffic scenes. Specifically, our approach involves three masking-reconstruction modeling tasks on scene inputs including agents' trajectories and road network, pretraining the scene encoder to capture kinematics within trajectory, spatial structure of road network, and interactions among roads and agents. The pretrained encoder is then finetuned on the downstream forecasting task. Extensive experiments demonstrate that SEPT, without elaborate architectural design or manual feature engineering, achieves state-of-the-art performance on the Argoverse 1 and Argoverse 2 motion forecasting benchmarks, outperforming previous methods on all main metrics by a large margin.
****
Hiroki Furuta,Kuang-Huei Lee,Ofir Nachum,Yutaka Matsuo,Aleksandra Faust,Shixiang Shane Gu,Izzeddin Gur
Multimodal Web Navigation with Instruction-Finetuned Foundation Models
The progress of autonomous web navigation has been hindered by the dependence on billions of exploratory interactions via online reinforcement learning, and domain-specific model designs that make it difficult to leverage generalization from rich out-of-domain data.
In this work, we study data-driven offline training for web agents with vision-language foundation models.
We propose an instruction-following multimodal agent, WebGUM, that observes both webpage screenshots and HTML pages and outputs web navigation actions, such as click and type.
WebGUM is trained by jointly finetuning an instruction-finetuned language model and a vision encoder with temporal and local perception on a large corpus of demonstrations.
We empirically demonstrate this recipe improves the agent's ability of grounded multimodal perception, HTML comprehension, and multi-step reasoning, outperforming prior works by a significant margin.
On the MiniWoB, we improve over the previous best offline methods by more than 45.8%, even outperforming online-finetuned SoTA, humans, and GPT-4-based agent.
On the WebShop benchmark, our 3-billion-parameter model achieves superior performance to the existing SoTA, PaLM-540B.
Furthermore, WebGUM exhibits strong positive transfer to the real-world planning tasks on the Mind2Web.
We also collect 347K high-quality demonstrations using our trained models, 38 times larger than prior work, and make them available to promote future research in this direction.
****
Binghui Xie,Yatao Bian,Kaiwen Zhou,Yongqiang Chen,Peilin Zhao,Bo Han,Wei Meng,James Cheng
Enhancing Neural Subset Selection: Integrating Background Information into Set Representations
Learning neural subset selection tasks, such as compound selection in AI-aided drug discovery, have become increasingly pivotal across diverse applications. The existing methodologies in the field primarily concentrate on constructing models that capture the relationship between utility function values and subsets within their respective supersets. However, these approaches tend to overlook the valuable information contained within the superset when utilizing neural networks to model set functions. In this work, we address this oversight by adopting a probabilistic perspective. Our theoretical findings demonstrate that when the target value is conditioned on both the input set and subset, it is essential to incorporate an invariant sufficient statistic of the superset into the subset of interest for effective learning. This ensures that the output value remains invariant to permutations of the subset and its corresponding superset, enabling ident

ification of the specific superset from which the subset originated. Motivated by these insights, we propose a simple yet effective information aggregation module designed to merge the representations of subsets and supersets from a permutation invariance perspective. Comprehensive empirical evaluations across diverse tasks and datasets validate the enhanced efficacy of our approach over conventional methods, underscoring the practicality and potency of our proposed strategies in real-world contexts.
****
Jason Piquenot,Aldo Moscatelli,Maxime Berar,Pierre Héroux,Romain Raveaux,Jean-Yves RAMEL,Sébastien Adam
G^{2}N^2 : Weisfeiler and Lehman go grammatical
This paper introduces a framework for formally establishing a connection between a portion of an algebraic language and a Graph Neural Network (GNN). The framework leverages Context-Free Grammars (CFG) to organize algebraic operations into generative rules that can be translated into a GNN layer model. As CFGs derived directly from a language tend to contain redundancies in their rules and variables, we present a grammar reduction scheme. By applying this strategy, we define a CFG that conforms to the third-order Weisfeiler-Lehman (3-WL) test using the matricial language MATLANG. From this 3-WL CFG, we derive a GNN model, named G^{2}N^2, which is provably 3-WL compliant. Through various experiments, we demonstrate the superior efficiency of G^{2}N^2 compared to other 3-WL GNNs across numerous downstream tasks. Specifically, one experiment highlights the benefits of grammar reduction within our framework.
****
Ilan Naiman,N. Benjamin Erichson,Pu Ren,Michael W. Mahoney,Omri Azencot
Generative Modeling of Regular and Irregular Time Series Data via Koopman VAEs
Generating realistic time series data is important for many engineering and scientific applications.
Existing work tackles this problem using generative adversarial networks (GANs).
However, GANs are unstable during training, and they can suffer from mode collapse.
While variational autoencoders (VAEs) are known to be more robust to the these issues, they are (surprisingly) less considered for time series generation.
In this work, we introduce Koopman VAE (KoVAE), a new generative framework that is based on a novel design for the model prior, and that can be optimized for either regular and irregular training data.
Inspired by Koopman theory, we represent the latent conditional prior dynamics using a linear map.
Our approach enhances generative modeling with two desired features: (i) incorporating domain knowledge can be achieved by leveraging spectral tools that prescribe constraints on the eigenvalues of the linear map; and (ii) studying the qualitative behavior and stability of the system can be performed using tools from dynamical systems theory.
Our results show that KoVAE outperforms state-of-the-art GAN and VAE methods across several challenging synthetic and real-world time series generation benchmarks.
Whether trained on regular or irregular data, KoVAE generates time series that improve both discriminative and predictive metrics.
We also present visual evidence suggesting that KoVAE learns probability density functions that better approximate the empirical ground truth distribution.
****
Sravanthi Gurugubelli,Sundeep Prabhakar Chepuri
SaNN: Simple Yet Powerful Simplicial-aware Neural Networks
Simplicial neural networks (SNNs) are deep models for higher-order graph representation learning. SNNs learn low-dimensional embeddings of simplices in a simplicial complex by aggregating features of their respective upper, lower, boundary, and coboundary adjacent simplices. The aggregation in SNNs is carried out during training. Since the number of simplices of various orders in a simplicial complex is significantly large, the memory and training-time requirement in SNNs is enormous. In this work, we propose a scalable simplicial-aware neural network (S

aNN) model with a constant run-time and memory requirements independent of the size of the simplicial complex and the density of interactions in it. SaNN is based on pre-aggregated simplicial-aware features as inputs to a neural network, so it has a strong simplicial-structural inductive bias. We provide theoretical conditions under which SaNN is provably more powerful than the Weisfeiler-Lehman (WL) graph isomorphism test and as powerful as the simplicial Weisfeiler-Lehman (SWL) test. We also show that SaNN is permutation and orientation equivariant and satisfies simplicial-awareness of the highest order in a simplicial complex. We demonstrate via numerical experiments that despite being computationally economical, the proposed model achieves state-of-the-art performance in predicting trajectories, simplicial closures, and classifying graphs.
****
Yu-Yu Wu,Hung-Jui Wang,Shang-Tse Chen
Annealing Self-Distillation Rectification Improves Adversarial Training
In standard adversarial training, models are optimized to fit invariant one-hot labels for adversarial data when the perturbations are within allowable budgets. However, the overconfident target harms generalization and causes the problem of robust overfitting. To address this issue and enhance adversarial robustness, we analyze the characteristics of robust models and identify that robust models tend to produce smoother and well-calibrated outputs. Based on the observation, we propose a simple yet effective method, Annealing Self-Distillation Rectification (ADR), which generates soft labels as a better guidance mechanism that reflects the underlying distribution of data. By utilizing ADR, we can obtain rectified labels that improve model robustness without the need for pre-trained models or extensive extra computation. Moreover, our method facilitates seamless plug-and-play integration with other adversarial training techniques by replacing the hard labels in their objectives. We demonstrate the efficacy of ADR through extensive experiments and strong performances across datasets.
****
Ruoyu Wang,Yongqi Yang,Zhihao Qian,Ye Zhu,Yu Wu
Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation
Originating from the diffusion phenomenon in physics that describes particle movement, the diffusion generative models inherit the characteristics of stochastic random walk in the data space along the denoising trajectory. However, the intrinsic mutual interference among image regions contradicts the need for practical downstream application scenarios where the preservation of low-level pixel information from given conditioning is desired (e.g., customization tasks like personalized generation and inpainting based on a user-provided single image). In this work, we investigate the diffusion (physics) in diffusion (machine learning) properties and propose our Cyclic One-Way Diffusion (COW) method to control the direction of diffusion phenomenon given a pre-trained frozen diffusion model for versatile customization application scenarios, where the low-level pixel information from the conditioning needs to be preserved. Notably, unlike most current methods that incorporate additional conditions by fine-tuning the base text-to-image diffusion model or learning auxiliary networks, our method provides a novel perspective to understand the task needs and is applicable to a wider range of customization scenarios in a learning-free manner. Extensive experiment results show that our proposed COW can achieve more flexible customization based on strict visual conditions in different application settings. Project page: https://wangruoyu02.github.io/cow.github.io/.
****
Masanori Koyama,Kenji Fukumizu,Kohei Hayashi,Takeru Miyato
Neural Fourier Transform: A General Approach to Equivariant Representation Learning
Symmetry learning has proven to be an effective approach for extracting the hidden structure of data, with the concept of equivariance relation playing the central role.
However, most of the current studies are built on architectural theory and corresponding assumptions on the form of data.

We propose Neural Fourier Transform (NFT), a general framework of learning the l
atent linear action of the group without assuming explicit knowledge of how the
group acts on data.
We present the theoretical foundations of NFT and show that
the existence of a linear equivariant feature, which has been assumed ubiquitous
ly in equivariance learning, is equivalent to the existence of a group invariant
 kernel on the dataspace.
We also provide experimental results to demonstrate the application of NFT in ty
pical scenarios with varying levels of knowledge about the acting group.
****
Seungjae Shin,HeeSun Bae,Byeonghu Na,Yoon-Yeong Kim,Il-chul Moon
Unknown Domain Inconsistency Minimization for Domain Generalization
The objective of domain generalization (DG) is to enhance the transferability of
 the model learned from a source domain to unobserved domains. To prevent overfi
tting to a specific domain, Sharpness-Aware Minimization (SAM) reduces source do
main’s loss sharpness. Although SAM variants have delivered significant improvem
ents in DG, we highlight that there’s still potential for improvement in general
izing to unknown domains through the exploration on data space. This paper intro
duces an objective rooted in both parameter and data perturbed regions for domai
n generalization, coined Unknown Domain Inconsistency Minimization (UDIM). UDIM
reduces the loss landscape inconsistency between source domain and unknown domai
ns. As unknown domains are inaccessible, these domains are empirically crafted b
y perturbing instances from the source domain dataset. In particular, by alignin
g the loss landscape acquired in the source domain to the loss landscape of pert
urbed domains, we expect to achieve generalization grounded on these flat minima
 for the unknown domains. Theoretically, we validate that merging SAM optimizati
on with the UDIM objective establishes an upper bound for the true objective of
the DG task. In an empirical aspect, UDIM consistently outperforms SAM variants
across multiple DG benchmark datasets. Notably, UDIM shows statistically signifi
cant improvements in scenarios with more restrictive domain information, undersc
oring UDIM’s generalization capability in unseen domains.
****
Ge Gao,Qitong Gao,Xi Yang,Song Ju,Miroslav Pajic,Min Chi
On Trajectory Augmentations for Off-Policy Evaluation
In the realm of reinforcement learning (RL), off-policy evaluation (OPE) holds a
 pivotal position, especially in high-stake human-involved scenarios such as e-l
earning and healthcare. Applying OPE to these domains is often challenging with
scarce and underrepresentative offline training trajectories. Data augmentation
has been a successful technique to enrich training data. However, directly emplo
ying existing data augmentation methods to OPE may not be feasible, due to the M
arkovian nature within the offline trajectories and the desire for generalizabil
ity across diverse target policies. In this work, we propose an offline trajecto
ry augmentation approach to specifically facilitate OPE in human-involved scenar
ios. We propose sub-trajectory mining to extract potentially valuable sub-trajec
tories from offline data, and diversify the behaviors within those sub-trajector
ies by varying coverage of the state-action space. Our work was empirically eval
uated in a wide array of environments, encompassing both simulated scenarios and
 real-world domains like robotic control, healthcare, and e-learning, where the
training trajectories include varying levels of coverage of the state-action spa
ce. By enhancing the performance of a variety of OPE methods, our work offers a
promising path forward for tackling OPE challenges in situations where data may
be limited or underrepresentative.
****
Yiliu Wang,Wei Chen,Milan Vojnovic
Combinatorial Bandits for Maximum Value Reward Function under Value-Index Feedba
ck
We investigate the combinatorial multi-armed bandit problem where an action is t
o select k arms from a set of base arms, and its reward is the maximum of the
sample values of these k arms, under a weak feedback structure that only retur
ns the value and index of the arm with the maximum value. This novel feedback st

ructure is much weaker than the semi-bandit feedback previously studied and is only slightly stronger than the full-bandit feedback, and thus it presents a new challenge for the online learning task. We propose an algorithm and derive a regret bound for instances where arm outcomes follow distributions with finite supports. Our algorithm introduces a novel concept of biased arm replacement to address the weak feedback challenge, and it achieves a distribution-dependent regret bound of $O((k/\Delta)\log(T))$ and a distribution-independent regret bound of $\tilde{O}(\sqrt{T})$, where Δ is the reward gap and T is the time horizon.
Notably, our regret bound is comparable to the bounds obtained under the more informative semi-bandit feedback.
We demonstrate the effectiveness of our algorithm through experimental results.
****
Zilin Si,Gu Zhang,Qingwei Ben,Branden Romero,Zhou Xian,Chao Liu,Chuang Gan
DIFFTACTILE: A Physics-based Differentiable Tactile Simulator for Contact-rich Robotic Manipulation
We introduce DIFFTACTILE, a physics-based differentiable tactile simulation system designed to enhance robotic manipulation with dense and physically accurate tactile feedback. In contrast to prior tactile simulators which primarily focus on manipulating rigid bodies and often rely on simplified approximations to model stress and deformations of materials in contact, DIFFTACTILE emphasizes physics-based contact modeling with high fidelity, supporting simulations of diverse contact modes and interactions with objects possessing a wide range of material properties. Our system incorporates several key components, including a Finite Element Method (FEM)-based soft body model for simulating the sensing elastomer, a multi-material simulator for modeling diverse object types (such as elastic, elastoplastic, cables) under manipulation, a penalty-based contact model for handling contact dynamics. The differentiable nature of our system facilitates gradient-based optimization for both 1) refining physical properties in simulation using real-world data, hence narrowing the sim-to-real gap and 2) efficient learning of tactile-assisted grasping and contact-rich manipulation skills. Additionally, we introduce a method to infer the optical response of our tactile sensor to contact using an efficient pixel-based neural module. We anticipate that DIFFTACTILE will serve as a useful platform for studying contact-rich manipulations, leveraging the benefits of dense tactile feedback and differentiable physics. Code and supplementary materials are available at the project website https://difftactile.github.io/.
****
Xuefeng Liu,Takuma Yoneda,Rick Stevens,Matthew Walter,Yuxin Chen
Blending Imitation and Reinforcement Learning for Robust Policy Improvement
While reinforcement learning (RL) has shown promising performance, its sample complexity continues to be a substantial hurdle, restricting its broader application across a variety of domains. Imitation learning (IL) utilizes oracles to improve sample efficiency, yet it is often constrained by the quality of the oracles deployed. To address the demand for robust policy improvement in real-world scenarios, we introduce a novel algorithm, Robust Policy Improvement (RPI), which actively interleaves between IL and RL based on an online estimate of their performance. RPI draws on the strengths of IL, using oracle queries to facilitate exploration—an aspect that is notably challenging in sparse-reward RL—particularly during the early stages of learning. As learning unfolds, RPI gradually transitions to RL, effectively treating the learned policy as an improved oracle. This algorithm is capable of learning from and improving upon a diverse set of black-box oracles. Integral to RPI are Robust Active Policy Selection (RAPS) and Robust Policy Gradient (RPG), both of which reason over whether to perform state-wise imitation from the oracles or learn from its own value function when the learner’s performance surpasses that of the oracles in a specific state. Empirical evaluations and theoretical analysis validate that RPI excels in comparison to existing state-of-the-art methodologies, demonstrating superior performance across various benchmark domains.
****

Pingzhi Li,Zhenyu Zhang,Prateek Yadav,Yi-Lin Sung,Yu Cheng,Mohit Bansal,Tianlong Chen
Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy
Sparsely activated Mixture-of-Experts (SMoE) has shown promise to scale up the learning capacity of neural networks, however, they have issues like: (a) $\textit{High Memory Usage,}$ due to duplication of the network layers into multiple copies as experts; and (b) $\textit{Redundancy in Experts,}$ as common learning-based routing policies suffer from representational collapse. Therefore, vanilla SMoE models are memory inefficient and non-scalable, especially for resource-constrained downstream scenarios. In this paper, we ask: Can we craft a compact SMoE model by consolidating expert information? What is the best recipe to merge multiple experts into fewer but more knowledgeable experts? Our pilot investigation reveals that conventional model merging methods fail to be effective in such expert merging for SMoE. The potential reasons are: (1) redundant information overshadows critical experts; (2) appropriate neuron permutation for each expert is missing to bring all of them in alignment. To address these challenges, we propose a novel merging algorithm for SMoE, $\textit{i.e.}$, $\texttt{M-SMoE}$, which leverages routing statistics to guide expert merging. Specifically, it starts with neuron permutation alignment for experts; then, dominant experts and their "group members" are formed based on routing policies; lastly, every expert group is merged into a single expert by utilizing each expert's activation frequency as their weight for merging, thus diminishing the impact of insignificant experts. Moreover, we draw an interesting observation that our proposed merging promotes a low dimensionality in the merged expert's weight space, naturally paving the way for additional compression. Hence, our final method, $\texttt{MC-SMoE}$ ($\textit{i.e.}$, Merge, then Compress SMoE), further decomposes the merged experts into low-rank and structural sparse alternatives. Extensive experiments across 8 benchmarks validate the effectiveness of our proposals. For instance, our $\texttt{MC-SMoE}$ achieves up to 80% memory and a 20% FLOPs reduction, with virtually no loss in performance. Our code is provided as supplementary material.
****
Marc Rigter,Minqi Jiang,Ingmar Posner
Reward-Free Curricula for Training Robust World Models
There has been a recent surge of interest in developing generally-capable agents that can adapt to new tasks without additional training in the environment. Learning world models from reward-free exploration is a promising approach, and enables policies to be trained using imagined experience for new tasks. However, achieving a general agent requires robustness across different environments. In this work, we address the novel problem of generating curricula in the reward-free setting to train robust world models. We consider robustness in terms of minimax regret over all environment instantiations and show that the minimax regret can be connected to minimising the maximum error in the world model across environment instances. This result informs our algorithm, WAKER: Weighted Acquisition of Knowledge across Environments for Robustness. WAKER selects environments for data collection based on the estimated error of the world model for each environment. Our experiments demonstrate that WAKER outperforms na■■ve domain randomisation, resulting in improved robustness, efficiency, and generalisation.
****
Jia-Wang Bian,Wenjing Bian,Victor Adrian Prisacariu,Philip Torr
PORF: POSE RESIDUAL FIELD FOR ACCURATE NEURAL SURFACE RECONSTRUCTION
Neural surface reconstruction is sensitive to the camera pose noise, even when state-of-the-art pose estimators like COLMAP or ARKit are used. Existing Pose-NeRF joint optimisation methods have struggled to improve pose accuracy in challenging real-world scenarios. To overcome the challenges, we introduce the pose residual field (PoRF), a novel implicit representation that uses an MLP for regressing pose updates. Compared with the conventional per-frame pose parameter optimisation, this new representation is more robust due to parameter sharing that leverages global information over the entire sequence. Furthermore, we propose an ep

ipolar geometry loss to enhance the supervision that leverages the correspondences exported from COLMAP results without the extra computational overhead. Our method yields promising results. On the DTU dataset, we reduce the rotation error of COLMAP poses by 78\%, leading to the reduced reconstruction Chamfer distance from 3.48mm to 0.85mm. On the MobileBrick dataset that contains casually captured unbounded 360-degree videos, our method refines ARKit poses and improves the reconstruction F1 score from 69.18 to 75.67, outperforming that with the provided ground-truth pose (75.14). These achievements demonstrate the efficacy of our approach in refining camera poses and improving the accuracy of neural surface reconstruction in real-world scenarios.
****
Junsong Chen,Jincheng YU,Chongjian GE,Lewei Yao,Enze Xie,Zhongdao Wang,James Kwok,Ping Luo,Huchuan Lu,Zhenguo Li
PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis
The most advanced text-to-image (T2I) models require significant training costs (e.g., millions of GPU hours), seriously hindering the fundamental innovation for the AIGC community while increasing CO2 emissions. This paper introduces PixArt-α, a Transformer-based T2I diffusion model whose image generation quality is competitive with state-of-the-art image generators (e.g., Imagen, SDXL, and even Midjourney), reaching near-commercial application standards. Additionally, it supports high-resolution image synthesis up to 1024px resolution with low training cost, as shown in Figure 1 and 2. To achieve this goal, three core designs are proposed: (1) Training strategy decomposition: We devise three distinct training steps that separately optimize pixel dependency, text-image alignment, and image aesthetic quality; (2) Efficient T2I Transformer: We incorporate cross-attention modules into Diffusion Transformer (DiT) to inject text conditions and streamline the computation-intensive class-condition branch; (3) High-informative data: We emphasize the significance of concept density in text-image pairs and leverage a large Vision-Language model to auto-label dense pseudo-captions to assist text-image alignment learning. As a result, PixArt-α's training speed markedly surpasses existing large-scale T2I models, e.g., PixArt-α only takes 10.8% of Stable Diffusion v1.5's training time (~675 vs. ~6,250 A100 GPU days), saving nearly \\$300,000 (\\$26,000 vs. \\$320,000) and reducing 90% CO2 emissions. Moreover, compared with a larger SOTA model, RAPHAEL, our training cost is merely 1%. Extensive experiments demonstrate that PixArt-α excels in image quality, artistry, and semantic control. We hope PixArt-α will provide new insights to the AIGC community and startups to accelerate building their own high-quality yet low-cost generative models from scratch.
****
Lizhang Chen,Bo Liu,Kaizhao Liang,qiang liu
Lion Secretly Solves a Constrained Optimization: As Lyapunov Predicts
Lion (Evolved Sign Momentum), a new optimizer discovered through program search, has shown promising results in training large AI models. It achieves results comparable to AdamW but with greater memory efficiency. As what we can expect from the result of the random search, Lion blends a number of elements from existing algorithms, including signed momentum, decoupled weight decay, Polayk and Nesterov momentum, but doesn't fit into any existing category of theoretically grounded optimizers. Thus, even though Lion appears to perform well as a general-purpose optimizer for a wide range of tasks, its theoretical basis remains uncertain. This absence of theoretical clarity limits opportunities to further enhance and expand Lion's efficacy. This work aims to demystify Lion. Using both continuous-time and discrete-time analysis, we demonstrate that Lion is a novel and theoretically grounded approach for minimizing a general loss function $f(x)$ while enforcing a bound constraint $||x||_\infty \leq 1/\lambda$. Lion achieves this through the incorporation of decoupled weight decay, where λ represents the weight decay coefficient. Our analysis is facilitated by the development of a new Lyapunov function for the Lion updates. It applies to a wide range of Lion-ϕ algorithms, where the $sign(\cdot)$ operator in Lion is replaced by the subgradient of a convex function ϕ, leading to the solution of the general c

omposite optimization problem $\min_x f(x) + \phi^*(x)$. Our findings provide valuable insights into the dynamics of Lion and pave the way for further enhancements and extensions of Lion-related algorithms.
****
Daniel Severo,Lucas Theis,Johannes Ballé
The Unreasonable Effectiveness of Linear Prediction as a Perceptual Metric
We show how perceptual embeddings of the visual system can be constructed at inference-time with no training data or deep neural network features. Our perceptual embeddings are solutions to a weighted least squares (WLS) problem, defined at the pixel-level, and solved at inference-time, that can capture global and local image characteristics. The distance in embedding space is used to define a perceptual similary metric which we call \emph{LASI: Linear Autoregressive Similarity Index}. Experiments on full-reference image quality assessment datasets show LASI performs competitively with learned deep feature based methods like LPIPS \citep{zhang2018unreasonable} and PIM \citep{bhardwaj2020unsupervised}, at a similar computational cost to hand-crafted methods such as MS-SSIM \citep{wang2003multiscale}. We found that increasing the dimensionality of the embedding space consistently reduces the WLS loss while increasing performance on perceptual tasks, at the cost of increasing the computational complexity. LASI is fully differentiable, scales cubically with the number of embedding dimensions, and can be parallelized at the pixel-level. A Maximum Differentiation (MAD) competition \citep{wang2008maximum} between LASI and LPIPS shows that both methods are capable of finding failure points for the other, suggesting these metrics can be combined.
****
Tong Zhou,Shaolei Ren,Xiaolin Xu
ArchLock: Locking DNN Transferability at the Architecture Level with a Zero-Cost Binary Predictor
Deep neural network (DNN) models, despite their impressive performance, are vulnerable to exploitation by attackers who attempt to transfer them to other tasks for their own benefit. Current defense strategies mainly address this vulnerability at the model parameter level, leaving the potential of architectural-level defense largely unexplored. This paper, for the first time, addresses the issue of model protection by reducing transferability at the architecture level. Specifically, we present a novel neural architecture search (NAS)-enabled algorithm that employs zero-cost proxies and evolutionary search, to explore model architectures with low transferability. Our method, namely ArchLock, aims to achieve high performance on the source task, while degrading the performance on potential target tasks, i.e., locking the transferability of a DNN model. To achieve efficient cross-task search without accurately knowing the training data owned by the attackers, we utilize zero-cost proxies to speed up architecture evaluation and simulate potential target task embeddings to assist cross-task search with a binary performance predictor. Extensive experiments on NAS-Bench-201 and TransNAS-Bench-101 demonstrate that ArchLock reduces transferability by up to 30% and 50%, respectively, with negligible performance degradation on source tasks (<2%). The code is available at https://github.com/Tongzhou0101/ArchLock.
****
Aaron Zweig,Joan Bruna
Symmetric Single Index Learning
Few neural architectures lend themselves to provable learning with gradient based methods. One popular model is the single-index model, in which labels are produced by composing an unknown linear projection with a possibly unknown scalar link function. Learning this model with SGD is relatively well-understood, whereby the so-called information exponent of the link function governs a polynomial sample complexity rate. However, extending this analysis to deeper or more complicated architectures remains challenging.

In this work, we consider single index learning in the setting of symmetric neural networks. Under analytic assumptions on the activation and maximum degree assumptions on the link function, we prove that gradient flow recovers the hidden planted direction, represented as a finitely supported vector in the feature spa

ce of power sum polynomials. We characterize a notion of information exponent a dapted to our setting that controls the efficiency of learning.
****
Zhiqiu Xu,Yanjie Chen,Kirill Vishniakov,Yida Yin,Zhiqiang Shen,Trevor Darrell,Lingjie Liu,Zhuang Liu
Initializing Models with Larger Ones
Weight initialization plays an important role in neural network training. Widely used initialization methods are proposed and evaluated for networks that are trained from scratch. However, the growing number of pretrained models now offers new opportunities for tackling this classical problem of weight initialization. In this work, we introduce weight selection, a method for initializing smaller models by selecting a subset of weights from a pretrained larger model. This enables the transfer of knowledge from pretrained weights to smaller models. Our experiments demonstrate that weight selection can significantly enhance the performance of small models and reduce their training time. Notably, it can also be used together with knowledge distillation. Weight selection offers a new approach to leverage the power of pretrained models in resource-constrained settings, and we hope it can be a useful tool for training small models in the large-model era.
****
Sohyun An,Hayeon Lee,Jaehyeong Jo,Seanie Lee,Sung Ju Hwang
DiffusionNAG: Predictor-guided Neural Architecture Generation with Diffusion Models
Existing NAS methods suffer from either an excessive amount of time for repetitive sampling and training of many task-irrelevant architectures. To tackle such limitations of existing NAS methods, we propose a paradigm shift from NAS to a novel conditional Neural Architecture Generation (NAG) framework based on diffusion models, dubbed DiffusionNAG. Specifically, we consider the neural architectures as directed graphs and propose a graph diffusion model for generating them. Moreover, with the guidance of parameterized predictors, DiffusionNAG can flexibly generate task-optimal architectures with the desired properties for diverse tasks, by sampling from a region that is more likely to satisfy the properties. This conditional NAG scheme is significantly more efficient than previous NAS schemes which sample the architectures and filter them using the property predictors. We validate the effectiveness of DiffusionNAG through extensive experiments in two predictor-based NAS scenarios: Transferable NAS and Bayesian Optimization (BO)-based NAS. DiffusionNAG achieves superior performance with speedups of up to 35$\times$ when compared to the baselines on Transferable NAS benchmarks. Furthermore, when integrated into a BO-based algorithm, DiffusionNAG outperforms existing BO-based NAS approaches, particularly in the large MobileNetV3 search space on the ImageNet 1K dataset. Code is available at https://github.com/CownowAn/DiffusionNAG.
****
Zi Wang,Aaron J Havens,Alexandre Araujo,Yang Zheng,Bin Hu,Yudong Chen,Somesh Jha
On the Scalability and Memory Efficiency of Semidefinite Programs for Lipschitz Constant Estimation of Neural Networks
Lipschitz constant estimation plays an important role for understanding generalization, robustness, and fairness in deep learning. Unlike naive bounds based on the network weight norm product, semidefinite programs (SDPs) have shown great promise in providing less conservative Lipschitz bounds with polynomial-time complexity guarantees. However, due to the memory consumption and running speed, standard SDP algorithms cannot scale to modern neural network structures. In this paper, we transform the SDPs for Lipschitz constant estimation into an eigenvalue problem, which aligns with the modern large optimization paradigms based on first-order methods. This is amenable to autodiff frameworks such as PyTorch and TensorFlow, requiring significantly less memory than standard SDP algorithms. The transformation also allows us to leverage various existing numerical techniques for eigenvalue optimization, opening the way for further memory improvement and computational speedup. The essential technique of our eigenvalue-problem transformation is to introduce redundant quadratic constraints and then utilize both La

grangian and Shor's SDP relaxations. Numerical examples demonstrate that our tec hnique is more scalable than existing approaches. For networks that existing SDP solvers cannot handle, we improve the Lipschitz constant estimation by up to 58 \% compared to the weight matrix norm product bound.
****
Xian Liu,Jian Ren,Aliaksandr Siarohin,Ivan Skorokhodov,Yanyu Li,Dahua Lin,Xihui Liu,Ziwei Liu,Sergey Tulyakov
HyperHuman: Hyper-Realistic Human Generation with Latent Structural Diffusion
Despite significant advances in large-scale text-to-image models, achieving hype r-realistic human image generation remains a desirable yet unsolved task. Existi ng models like Stable Diffusion and DALL·E 2 tend to generate human images with incoherent parts or unnatural poses. To tackle these challenges, our key insight is that human image is inherently structural over multiple granularities, from the coarse-level body skeleton to fine-grained spatial geometry. Therefore, capt uring such correlations between the explicit appearance and latent structure in one model is essential to generate coherent and natural human images. To this en d, we propose a unified framework, HyperHuman, that generates in-the-wild human images of high realism and diverse layouts. Specifically, 1) we first build a la rge-scale human-centric dataset, named HumanVerse, which consists of 340M images with comprehensive annotations like human pose, depth, and surface normal. 2) N ext, we propose a Latent Structural Diffusion Model that simultaneously denoises the depth and surface normal along with the synthesized RGB image. Our model en forces the joint learning of image appearance, spatial relationship, and geometr y in a unified network, where each branch in the model complements to each other with both structural awareness and textural richness. 3) Finally, to further bo ost the visual quality, we propose a Structure-Guided Refiner to compose the pre dicted conditions for more detailed generation of higher resolution. Extensive e xperiments demonstrate that our framework yields the state-of-the-art performanc e, generating hyper-realistic human images under diverse scenarios.
****
Yanzhou Li,Kangjie Chen,Tianlin Li,Jian Zhang,Shangqing Liu,Wenhan Wang,Tianwei Zhang,Yang Liu
BadEdit: Backdooring Large Language Models by Model Editing
Mainstream backdoor attack methods typically demand substantial tuning data for poisoning, limiting their practicality and potentially degrading the overall per formance when applied to Large Language Models (LLMs). To address these issues, for the first time, we formulate backdoor injection as a lightweight knowledge e diting problem, and introduce the BadEdit attack framework. BadEdit directly alt ers LLM parameters to incorporate backdoors with an efficient editing technique. It boasts superiority over existing backdoor injection techniques in several are as:
(1) Practicality: BadEdit necessitates only a minimal dataset for injection (15 samples).
(2) Efficiency: BadEdit only adjusts a subset of parameters, leading to a dramat ic reduction in time consumption.
(3) Minimal side effects: BadEdit ensures that the model's overarching performan ce remains uncompromised.
(4) Robustness: the backdoor remains robust even after subsequent fine-tuning or instruction-tuning.
Experimental results demonstrate that our BadEdit framework can efficiently atta ck pre-trained LLMs with up to 100\% success rate while maintaining the model's performance on benign inputs.
****
Nayoung Lee,Kartik Sreenivasan,Jason D. Lee,Kangwook Lee,Dimitris Papailiopoulos
Teaching Arithmetic to Small Transformers
Large language models like GPT-4 exhibit emergent capabilities across general-pu rpose tasks, such as basic arithmetic, when trained on extensive text data, even though these tasks are not explicitly encoded by the unsupervised, next-token p rediction objective. This study investigates how even small transformers, traine d from random initialization, can efficiently learn arithmetic operations such a

s addition, multiplication, and elementary functions like square root, using the
 next-token prediction objective. We first demonstrate that conventional trainin
g data is not the most effective for arithmetic learning, and simple formatting
changes can significantly improve accuracy. This leads to sharp phase transition
s as a function of training data scale, which, in some cases, can be explained t
hrough connections to low-rank matrix completion. Building on prior work, we the
n train on chain-of-thought style data that includes intermediate step results.
Even in the complete absence of pretraining, this approach significantly and sim
ultaneously improves accuracy, sample complexity, and convergence speed. We also
 study the interplay between arithmetic and text data during training and examin
e the effects of few-shot prompting, pretraining, and parameter scaling. Additio
nally, we discuss the challenges associated with length generalization. Our work
 highlights the importance of high-quality, instructive data that considers the
particular characteristics of the next-word prediction loss for rapidly elicitin
g arithmetic capabilities.
****
Zhen Yang,Ganggui Ding,Wen Wang,Hao Chen,Bohan Zhuang,Chunhua Shen
Object-Aware Inversion and Reassembly for Image Editing
Diffusion-based image editing methods have achieved remarkable advances in text-
driven image editing. The editing task aims to convert an input image with the o
riginal text prompt into the desired image that is well-aligned with the target
text prompt. By comparing the original and target prompts, we can obtain numerou
s editing pairs, each comprising an object and its corresponding editing target.
 To allow editability while maintaining fidelity to the input image, existing ed
iting methods typically involve a fixed number of inversion steps that project t
he whole input image to its noisier latent representation, followed by a denoisi
ng process guided by the target prompt. However, we find that the optimal number
 of inversion steps for achieving ideal editing results varies significantly amo
ng different editing pairs, owing to varying editing difficulties. Therefore, th
e current literature, which relies on a fixed number of inversion steps, produce
s sub-optimal generation quality, especially when handling multiple editing pair
s in a natural image.
To this end, we propose a new image editing paradigm, dubbed Object-aware Invers
ion and Reassembly (OIR), to enable object-level fine-grained editing. Specifica
lly, we design a new search metric, which determines the optimal inversion steps
 for each editing pair, by jointly considering the editability of the target and
 the fidelity of the non-editing region. We use our search metric to find the op
timal inversion step for each editing pair when editing an image. We then edit t
hese editing pairs separately to avoid \concept. Subsequently, we propose an add
itional reassembly step to seamlessly integrate the respective editing results a
nd the non-editing region to obtain the final edited image. To systematically ev
aluate the effectiveness of our method, we collect two datasets called OIRBench
for benchmarking single- and multi-object editing, respectively. Experiments dem
onstrate that our method achieves superior performance in editing object shapes,
 colors, materials, categories, \textit{etc.}, especially in multi-object editin
g scenarios.
The project page can be found in https://aim-uofa.github.io/OIR-Diffusion/.
****
Namjun Kim,Chanho Min,Sejun Park
Minimum width for universal approximation using ReLU networks on compact domain
It has been shown that deep neural networks of a large enough width are universa
l approximators but they are not if the width is too small.
There were several attempts to characterize the minimum width $w_{\min}$ enablin
g the universal approximation property; however, only a few of them found the ex
act values.
In this work, we show that the minimum width for L^p approximation of L^p fu
nctions from $[0,1]^{d_x}$ to $\mathbb R^{d_y}$ is exactly $\max\\{d_x,d_y,2\\}$
 if an activation function is ReLU-Like (e.g., ReLU, GELU, Softplus).
Compared to the known result for ReLU networks, $w_{\min}=\max\\{d_x+1,d_y\\}$ w
hen the domain is ${\mathbb R^{d_x}}$, our result first shows that approximation

 on a compact domain requires smaller width than on ${\mathbb R^{d_x}}$.
We next prove a lower bound on $w_{\min}$ for uniform approximation using genera
l activation functions including ReLU: $w_{\min}\ge d_y+1$ if $d_x<d_y\le2d_x$.
Together with our first result, this shows a dichotomy between L^p and uniform
 approximations for general activation functions and input/output dimensions.
****
Tao Dai,Beiliang Wu,Peiyuan Liu,Naiqi Li,Jigang Bao,Yong Jiang,Shu-Tao Xia
Periodicity Decoupling Framework for Long-term Series Forecasting
Convolutional neural network (CNN)-based and Transformer-based methods have rece
ntly made significant strides in time series forecasting, which excel at modelin
g local temporal variations or capturing long-term dependencies. However, real-w
orld time series usually contain intricate temporal patterns, thus making it cha
llenging for existing methods that mainly focus on temporal variations modeling
from the 1D time series directly. Based on the intrinsic periodicity of time ser
ies, we propose a novel Periodicity Decoupling Framework (PDF) to capture 2D tem
poral variations of decoupled series for long-term series forecasting. Our PDF m
ainly consists of three components: multi-periodic decoupling block (MDB), dual
variations modeling block (DVMB), and variations aggregation block (VAB). Unlike
 the previous methods that model 1D temporal variations, our PDF mainly models 2
D temporal variations, decoupled from 1D time series by MDB. After that, DVMB at
tempts to further capture short-term and long-term variations, followed by VAB t
o make final predictions. Extensive experimental results across seven real-world
 long-term time series datasets demonstrate the superiority of our method over o
ther state-of-the-art methods, in terms of both forecasting performance and comp
utational efficiency. Code is available at https://github.com/Hank0626/PDF.
****
Siyuan Li,Weiyang Jin,Zedong Wang,Fang Wu,Zicheng Liu,Cheng Tan,Stan Z. Li
SemiReward: A General Reward Model for Semi-supervised Learning
Semi-supervised learning (SSL) has witnessed great progress with various improve
ments in the self-training framework with pseudo labeling. The main challenge is
 how to distinguish high-quality pseudo labels against the confirmation bias. Ho
wever, existing pseudo-label selection strategies are limited to pre-defined sch
emes or complex hand-crafted policies specially designed for classification, fai
ling to achieve high-quality labels, fast convergence, and task versatility simu
ltaneously. To these ends, we propose a Semi-supervised Reward framework (SemiRe
ward) that predicts reward scores to evaluate and filter out high-quality pseudo
 labels, which is pluggable to mainstream SSL methods in wide task types and sce
narios. To mitigate confirmation bias, SemiReward is trained online in two stage
s with a generator model and subsampling strategy. With classification and regre
ssion tasks on 13 standard SSL benchmarks across three modalities, extensive exp
eriments verify that SemiReward achieves significant performance gains and faste
r convergence speeds upon Pseudo Label, FlexMatch, and Free/SoftMatch. Code and
models are available at https://github.com/Westlake-AI/SemiReward.
****
Qinyu Zhao,Ming Xu,Kartik Gupta,Akshay Asthana,Liang Zheng,Stephen Gould
Towards Optimal Feature-Shaping Methods for Out-of-Distribution Detection
Feature shaping refers to a family of methods that exhibit state-of-the-art perf
ormance for out-of-distribution (OOD) detection. These approaches manipulate the
 feature representation, typically from the penultimate layer of a pre-trained d
eep learning model, so as to better differentiate between in-distribution (ID) a
nd OOD samples. However, existing feature-shaping methods usually employ rules m
anually designed for specific model architectures and OOD datasets, which conseq
uently limit their generalization ability. To address this gap, we first formula
te an abstract optimization framework for studying feature-shaping methods. We t
hen propose a concrete reduction of the framework with a simple piecewise consta
nt shaping function and show that existing feature-shaping methods approximate t
he optimal solution to the concrete optimization problem. Further, assuming that
 OOD data is inaccessible, we propose a formulation that yields a closed-form so
lution for the piecewise constant shaping function, utilizing solely the ID data
. Through extensive experiments, we show that the feature-shaping function optim

ized by our method improves the generalization ability of OOD detection across a large variety of datasets and model architectures. Our code is available at https://github.com/Qinyu-Allen-Zhao/OptFSOOD.
****
Oren Katzir,Or Patashnik,Daniel Cohen-Or,Dani Lischinski
Noise-free Score Distillation
Score Distillation Sampling (SDS) has emerged as the de facto approach for text-to-content generation in non-image domains. In this paper, we reexamine the SDS process and introduce a straightforward interpretation that demystifies the necessity for large Classifier-Free Guidance (CFG) scales, rooted in the distillation of an undesired noise term. Building upon our interpretation, we propose a novel Noise-Free Score Distillation (NFSD) process, which requires minimal modifications to the original SDS framework. Through this streamlined design, we achieve more effective distillation of pre-trained text-to-image diffusion models while using a nominal CFG scale. This strategic choice allows us to prevent the over-smoothing of results, ensuring that the generated data is both realistic and complies with the desired prompt. To demonstrate the efficacy of NFSD, we provide qualitative examples that compare NFSD and SDS, as well as several other methods.
****
Austin Tripp,Krzysztof Maziarz,Sarah Lewis,Marwin Segler,José Miguel Hernández-Lobato
Retro-fallback: retrosynthetic planning in an uncertain world
Retrosynthesis is the task of proposing a series of chemical reactions to create a desired molecule from simpler, buyable molecules. While previous works have proposed algorithms to find optimal solutions for a range of metrics (e.g. shortest, lowest-cost), these works generally overlook the fact that we have imperfect knowledge of the space of possible reactions, meaning plans created by the algorithm may not work in a laboratory. In this paper we propose a novel formulation of retrosynthesis in terms of stochastic processes to account for this uncertainty. We then propose a novel greedy algorithm called retro-fallback which maximizes the probability that at least one synthesis plan can be executed in the lab. Using in-silico benchmarks we demonstrate that retro-fallback generally produces better sets of synthesis plans than the popular MCTS and retro* algorithms.
****
Haitz Sáez de Ocáriz Borde,Anastasis Kratsios
Neural Snowflakes: Universal Latent Graph Inference via Trainable Latent Geometries
The inductive bias of a graph neural network (GNN) is largely encoded in its specified graph. Latent graph inference relies on latent geometric representations to dynamically rewire or infer a GNN's graph to maximize the GNN's predictive downstream performance, but it lacks solid theoretical foundations in terms of embedding-based representation guarantees. This paper addresses this issue by introducing a trainable deep learning architecture, coined \textit{neural snowflake}, that can adaptively implement fractal-like metrics on $\mathbb{R}^d$. We prove that any given finite weights graph can be isometrically embedded by a standard MLP encoder. Furthermore, when the latent graph can be represented in the feature space of a sufficiently regular kernel, we show that the combined neural snowflake and MLP encoder do not succumb to the curse of dimensionality by using only a low-degree polynomial number of parameters in the number of nodes. This implementation enables a low-dimensional isometric embedding of the latent graph. We conduct synthetic experiments to demonstrate the superior metric learning capabilities of neural snowflakes when compared to more familiar spaces like Euclidean space. Additionally, we carry out latent graph inference experiments on graph benchmarks. Consistently, the neural snowflake model achieves predictive performance that either matches or surpasses that of the state-of-the-art latent graph inference models. Importantly, this performance improvement is achieved without requiring random search for optimal latent geometry. Instead, the neural snowflake model achieves this enhancement in a differentiable manner.
****
Dustin Podell,Zion English,Kyle Lacey,Andreas Blattmann,Tim Dockhorn,Jonas Mülle

r,Joe Penna,Robin Rombach
SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis
We present Stable Diffusion XL (SDXL), a latent diffusion model for text-to-imag
e synthesis. Compared to previous versions of Stable Diffusion, SDXL leverages a
 three times larger UNet backbone, achieved by significantly increasing the numb
er of attention blocks and including a second text encoder. Further, we design m
ultiple novel conditioning schemes and train SDXL on multiple aspect ratios. To
ensure highest quality results, we also introduce a refinement model which is us
ed to improve the visual fidelity of samples generated by SDXL using a post-hoc
image-to-image technique. We demonstrate that SDXL improves dramatically over pr
evious versions of Stable Diffusion and achieves results competitive with those
of black-box state-of-the-art image generators such as Midjourney.
****
Yonghao Song,Bingchuan Liu,Xiang Li,Nanlin Shi,Yijun Wang,Xiaorong Gao
Decoding Natural Images from EEG for Object Recognition
Electroencephalography (EEG) signals, known for convenient non-invasive acquisit
ion but low signal-to-noise ratio, have recently gained substantial attention du
e to the potential to decode natural images. This paper presents a self-supervis
ed framework to demonstrate the feasibility of learning image representations fr
om EEG signals, particularly for object recognition. The framework utilizes imag
e and EEG encoders to extract features from paired image stimuli and EEG respons
es. Contrastive learning aligns these two modalities by constraining their simil
arity. Our approach achieves state-of-the-art results on a comprehensive EEG-ima
ge dataset, with a top-1 accuracy of 15.6% and a top-5 accuracy of 42.8% in 200-
way zero-shot tasks. Moreover, we perform extensive experiments to explore the b
iological plausibility by resolving the temporal, spatial, spectral, and semanti
c aspects of EEG signals. Besides, we introduce attention modules to capture spa
tial correlations, providing implicit evidence of the brain activity perceived f
rom EEG data. These findings yield valuable insights for neural decoding and bra
in-computer interfaces in real-world scenarios. Code available at https://github
.com/eeyhsong/NICE-EEG.
****
Anthony Bardou,Patrick Thiran,Thomas Begin
Relaxing the Additivity Constraints in Decentralized No-Regret High-Dimensional
Bayesian Optimization
Bayesian Optimization (BO) is typically used to optimize an unknown function f
 that is noisy and costly to evaluate, by exploiting an acquisition function tha
t must be maximized at each optimization step. Even if provably asymptotically o
ptimal BO algorithms are efficient at optimizing low-dimensional functions, scal
ing them to high-dimensional spaces remains an open problem, often tackled by as
suming an additive structure for f. By doing so, BO algorithms typically intro
duce additional restrictive assumptions on the additive structure that reduce th
eir applicability domain. This paper contains two main contributions: (i) we rel
ax the restrictive assumptions on the additive structure of f without weakenin
g the maximization guarantees of the acquisition function, and (ii) we address t
he over-exploration problem for decentralized BO algorithms. To these ends, we p
ropose DuMBO, an asymptotically optimal decentralized BO algorithm that achieves
 very competitive performance against state-of-the-art BO algorithms, especially
 when the additive structure of f comprises high-dimensional factors.
****
Thomas Coste,Usman Anwar,Robert Kirk,David Krueger
Reward Model Ensembles Help Mitigate Overoptimization
Reinforcement learning from human feedback (RLHF) is a standard approach for fin
e-tuning large language models to follow instructions. As part of this process,
learned reward models are used to approximately model human preferences. However
, as imperfect representations of the “true” reward, these learned reward models
 are susceptible to overoptimization. Gao et al. (2023) studied this phenomenon
in a synthetic human feedback setup with a significantly larger “gold” reward mo
del acting as the true reward (instead of humans) and showed that overoptimizati
on remains a persistent problem regardless of the size of the proxy reward model

 and training data used. Using a similar setup, we conduct a systematic study to
 evaluate the efficacy of using ensemble-based conservative optimization objecti
ves, specifically worst-case optimization (WCO) and uncertainty-weighted optimiz
ation (UWO), for mitigating reward model overoptimization when using two optimiz
ation methods: (a) best-of-n sampling (BoN) (b) proximal policy optimization (PP
O). We additionally extend the setup of Gao et al. (2023) to include 25% label n
oise to better mirror real-world conditions. Both with and without label noise w
e find that conservative optimization practically eliminates overoptimization an
d improves performance by up to 70% for BoN sampling. For PPO, ensemble-based co
nservative optimization always reduces overoptimization and outperforms single r
eward model optimization. Moreover, combining it with a small KL penalty success
fully prevents overoptimization at no performance cost. Overall, our results dem
onstrate that ensemble-based conservative optimization can effectively counter o
veroptimization.
****
William Rudman,Carsten Eickhoff
Stable Anisotropic Regularization
Given the success of Large Language Models (LLMs), there has been considerable i
nterest in studying the properties of model activations. The literature overwhel
mingly agrees that LLM representations are dominated by a few ``outlier dimensio
ns'' with exceedingly high variance and magnitude. Several studies in Natural La
nguage Processing (NLP) have sought to mitigate the impact of such outlier dimen
sions and force LLMs to be isotropic (i.e., have uniform variance across all dim
ensions in embedding space). Isotropy is thought to be a desirable property for
LLMs that improves model performance and more closely aligns textual representat
ions with human intuition. However, many claims regarding isotropy in NLP have b
een based on the average cosine similarity of embeddings, which has recently bee
n shown to be a flawed measure of isotropy. In this paper, we propose I-STAR: Is
oScore*-based STable Anisotropic Regularization, a novel regularization
 method that can be used to increase or decrease levels of isotropy in embedding
 space during training. I-STAR uses IsoScore*, the first accurate measu
re of isotropy that is both differentiable and stable on mini-batch computations
. In contrast to several previous works, we find that \textit{decreasing} isotro
py in contextualized embeddings improves performance on the majority of tasks an
d models considered in this paper.
****
chuan guo,Yuxuan Mu,Xinxin Zuo,Peng Dai,Youliang Yan,Juwei Lu,Li Cheng
Generative Human Motion Stylization in Latent Space
Human motion stylization aims to revise the style of an input motion while keepi
ng its content unaltered. Unlike existing works that operate directly in pose sp
ace, we leverage the \textit{latent space} of pretrained autoencoders as a more
expressive and robust representation for motion extraction and infusion. Buildin
g upon this, we present a novel \textit{generative} model that produces diverse
stylization results of a single motion (latent) code. During training, a motion
code is decomposed into two coding components: a deterministic content code, and
 a probabilistic style code adhering to a prior distribution; then a generator m
assages the random combination of content and style codes to reconstruct the cor
responding motion codes. Our approach is versatile, allowing the learning of pro
babilistic style space from either style labeled or unlabeled motions, providing
 notable flexibility in stylization as well. In inference, users can opt to styl
ize a motion using style cues from a reference motion or a label. Even in the ab
sence of explicit style input, our model facilitates novel re-stylization by sam
pling from the unconditional style prior distribution. Experimental results show
 that our proposed stylization models, despite their lightweight design, outperf
orm the state-of-the-arts in style reeanactment, content preservation, and gener
alization across various applications and settings.
****
Yoni Shafir,Guy Tevet,Roy Kapon,Amit Haim Bermano
Human Motion Diffusion as a Generative Prior
Recent work has demonstrated the significant potential of denoising diffusion mo

dels
for generating human motion, including text-to-motion capabilities.
However, these methods are restricted by the paucity of annotated motion data,
a focus on single-person motions, and a lack of detailed control.
In this paper, we introduce three forms of composition based on diffusion priors
:
sequential, parallel, and model composition.
Using sequential composition, we tackle the challenge of long sequence
generation. We introduce DoubleTake, an inference-time method with which
we generate long animations consisting of sequences of prompted intervals
and their transitions, using a prior trained only for short clips.
Using parallel composition, we show promising steps toward two-person generation
.
Beginning with two fixed priors as well as a few two-person training examples, we learn a slim
communication block, ComMDM, to coordinate interaction between the two resulting motions.
Lastly, using model composition, we first train individual priors
to complete motions that realize a prescribed motion for a given joint.
We then introduce DiffusionBlending, an interpolation mechanism to effectively blend several
such models to enable flexible and efficient fine-grained joint and trajectory-level control and editing.
We evaluate the composition methods using an off-the-shelf motion diffusion model,
and further compare the results to dedicated models trained for these specific tasks.
****
Chen Geng,Hong-Xing Yu,Sida Peng,Xiaowei Zhou,Jiajun Wu
Neural Polynomial Gabor Fields for Macro Motion Analysis
We study macro motion analysis, where macro motion refers to the collection of all visually observable motions in a dynamic scene. Traditional filtering-based methods on motion analysis typically focus only on local and tiny motions, yet fail to represent large motions or 3D scenes. Recent dynamic neural representations can faithfully represent motions using correspondences, but they cannot be directly used for motion analysis. In this work, we propose Phase-based neural polynomial Gabor fields (Phase-PGF), which learns to represent scene dynamics with low-dimensional time-varying phases. We theoretically show that Phase-PGF has several properties suitable for macro motion analysis. In our experiments, we collect diverse 2D and 3D dynamic scenes and show that Phase-PGF enables dynamic scene analysis and editing tasks including motion loop detection, motion factorization, motion smoothing, and motion magnification. Project page: https://chen-geng.com/phasepgf
****
Sihyun Yu,Weili Nie,De-An Huang,Boyi Li,Jinwoo Shin,Anima Anandkumar
Efficient Video Diffusion Models via Content-Frame Motion-Latent Decomposition
Video diffusion models have recently made great progress in generation quality, but are still limited by the high memory and computational requirements. This is because current video diffusion models often attempt to process high-dimensional videos directly. To tackle this issue, we propose content-motion latent diffusion model (CMD), a novel efficient extension of pretrained image diffusion models for video generation. Specifically, we propose an autoencoder that succinctly encodes a video as a combination of a content frame (like an image) and a low-dimensional motion latent representation. The former represents the common content, and the latter represents the underlying motion in the video, respectively. We generate the content frame by fine-tuning a pretrained image diffusion model, and we generate the motion latent representation by training a new lightweight diffusion model. A key innovation here is the design of a compact latent space that can directly utilizes a pretrained image diffusion model, which has not been done in previous latent video diffusion models. This leads to considerably better

quality generation and reduced computational costs. For instance, CMD can sampl
e a video 7.7$\times$ faster than prior approaches by generating a video of 512$
\times$1024 resolution and length 16 in 3.1 seconds. Moreover, CMD achieves an F
VD score of 238.3 on WebVid-10M, 18.5% better than the previous state-of-the-art
 of 292.4.
****
Erik J Bekkers,Sharvaree Vadgama,Rob Hesselink,Putri A Van der Linden,David W. R
omero
Fast, Expressive $\mathrm{SE}(n)$ Equivariant Networks through Weight-Sharing in
 Position-Orientation Space
Based on the theory of homogeneous spaces we derive *geometrically optimal edge
attributes* to be used within the flexible message-passing framework. We formali
ze the notion of weight sharing in convolutional networks as the sharing of mess
age functions over point-pairs that should be treated equally. We define equival
ence classes of point-pairs that are identical up to a transformation in the gro
up and derive attributes that uniquely identify these classes. Weight sharing is
 then obtained by conditioning message functions on these attributes. As an appl
ication of the theory, we develop an efficient equivariant group convolutional n
etwork for processing 3D point clouds. The theory of homogeneous spaces tells us
 how to do group convolutions with feature maps over the homogeneous space of po
sitions $\mathbb{R}^3$, position and orientations $\mathbb{R}^3 {\times} S^2$, a
nd the group $SE(3)$ itself. Among these, $\mathbb{R}^3 {\times} S^2$ is an opti
mal choice due to the ability to represent directional information, which $\math
bb{R}^3$ methods cannot, and it significantly enhances computational efficiency
compared to indexing features on the full $SE(3)$ group. We support this claim w
ith state-of-the-art results —in accuracy and speed— on five different benchmark
s in 2D and 3D, including interatomic potential energy prediction, trajectory fo
recasting in N-body systems, and generating molecules via equivariant diffusion
models.

*Code available at [https://github.com/ebekkers/ponita](https://github.com/ebekk
ers/ponita)*
****
Matthew Finlayson,John Hewitt,Alexander Koller,Swabha Swayamdipta,Ashish Sabharw
al
Closing the Curious Case of Neural Text Degeneration
Despite their ubiquity in language generation, it remains unknown why truncation
 sampling heuristics like nucleus sampling are so effective. We provide a theore
tical explanation for the effectiveness of the truncation sampling by proving th
at truncation methods that discard tokens below some probability threshold (the
most common type of truncation) can guarantee that all sampled tokens have nonze
ro true probability. However, thresholds are a coarse heuristic, and necessarily
 discard some tokens with nonzero true probability as well. In pursuit of a more
 precise sampling strategy, we show that we can leverage a known source of model
 errors, the softmax bottleneck, to prove that certain tokens have nonzero true
probability, without relying on a threshold. Based on our findings, we develop a
n experimental truncation strategy and the present pilot studies demonstrating t
he promise of this type of algorithm. Our evaluations show that our method outpe
rforms its threshold-based counterparts under automatic and human evaluation met
rics for low-entropy (i.e., close to greedy) open-ended text generation. Our the
oretical findings and pilot experiments provide both insight into why truncation
 sampling works, and make progress toward more expressive sampling algorithms th
at better surface the generative capabilities of large language models.
****
Nilesh Gupta,Fnu Devvrit,Ankit Singh Rawat,Srinadh Bhojanapalli,Prateek Jain,Ind
erjit S Dhillon
Dual-Encoders for Extreme Multi-label Classification
Dual-encoder (DE) models are widely used in retrieval tasks, most commonly studi
ed on open QA benchmarks that are often characterized by multi-class and limited
 training data. In contrast, their performance in multi-label and data-rich retr

ieval settings like extreme multi-label classification (XMC), remains under-expl
ored. Current empirical evidence indicates that DE models fall significantly sho
rt on XMC benchmarks, where SOTA methods linearly scale the number of learnable
parameters with the total number of classes (documents in the corpus) by employi
ng per-class classification head. To this end, we first study and highlight that
 existing multi-label contrastive training losses are not appropriate for traini
ng DE models on XMC tasks. We propose decoupled softmax loss - a simple modifica
tion to the InfoNCE loss - that overcomes the limitations of existing contrastiv
e losses. We further extend our loss design to a soft top-k operator-based loss
which is tailored to optimize top-k prediction performance. When trained with ou
r proposed loss functions, standard DE models alone can match or outperform SOTA
 methods by up to 2\% at Precision@1 even on the largest XMC datasets while bein
g 20× smaller in terms of the number of trainable parameters. This leads to more
 parameter-efficient and universally applicable solutions for retrieval tasks. O
ur code and models are publicly available [here](https://github.com/nilesh2797/d
exml).
****
An-Chieh Cheng,Xueting Li,Sifei Liu,Xiaolong Wang
TUVF: Learning Generalizable Texture UV Radiance Fields
Textures are a vital aspect of creating visually appealing and realistic 3D mode
ls. In this paper, we study the problem of generating high-fidelity texture give
n shapes of 3D assets, which has been relatively less explored compared with gen
eric 3D shape modeling. Our goal is to facilitate a controllable texture generat
ion process, such that one texture code can correspond to a particular appearanc
e style independent of any input shapes from a category. We introduce Texture UV
 Radiance Fields (TUVF) that generate textures in a learnable UV sphere space ra
ther than directly on the 3D shape. This allows the texture to be disentangled f
rom the underlying shape and transferable to other shapes that share the same UV
 space, i.e., from the same category. We integrate the UV sphere space with the
radiance field, which provides a more efficient and accurate representation of t
extures than traditional texture maps. We perform our experiments on synthetic a
nd real-world object datasets where we achieve not only realistic synthesis but
also substantial improvements over state-of-the-arts on texture controlling and
editing.
****
Ismail Yunus Akhalwaya,Shashanka Ubaru,Kenneth L. Clarkson,Mark S. Squillante,Vi
shnu Jejjala,Yang-Hui He,Kugendran Naidoo,Vasileios Kalantzis,Lior Horesh
Topological data analysis on noisy quantum computers
Topological data analysis (TDA) is a powerful technique for extracting complex a
nd valuable shape-related summaries of high-dimensional data. However, the compu
tational demands of classical algorithms for computing TDA are exorbitant, and q
uickly become impractical for high-order characteristics. Quantum computers offe
r the potential of achieving significant speedup for certain computational probl
ems. Indeed, TDA has been purported to be one such problem, yet, quantum computi
ng algorithms proposed for the problem, such as the original Quantum TDA (QTDA)
formulation by Lloyd, Garnerone and Zanardi, require fault-tolerance qualificati
ons that are currently unavailable. In this study, we present NISQ-TDA, a fully
implemented end-to-end quantum machine learning algorithm needing only a short c
ircuit-depth, that is applicable to high-dimensional classical data, and with pr
ovable asymptotic speedup for certain classes of problems. The algorithm neither
 suffers from the data-loading problem nor does it need to store the input data
on the quantum computer explicitly. The algorithm was successfully executed on q
uantum computing devices, as well as on noisy quantum simulators, applied to sma
ll datasets. Preliminary empirical results suggest that the algorithm is robust
to noise.
****
Mehrdad Saberi,Vinu Sankar Sadasivan,Keivan Rezaei,Aounon Kumar,Atoosa Chegini,W
enxiao Wang,Soheil Feizi
Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks
In light of recent advancements in generative AI models, it has become essential

to distinguish genuine content from AI-generated one to prevent the malicious usage of fake materials as authentic ones and vice versa. Various techniques have been introduced for identifying AI-generated images, with watermarking emerging as a promising approach. In this paper, we analyze the robustness of various AI-image detectors including watermarking and classifier-based deepfake detectors. For watermarking methods that introduce subtle image perturbations (i.e., low perturbation budget methods), we reveal a fundamental trade-off between the evasion error rate (i.e., the fraction of watermarked images detected as non-watermarked ones) and the spoofing error rate (i.e., the fraction of non-watermarked images detected as watermarked ones) upon an application of a diffusion purification attack. In this regime, we also empirically show that diffusion purification effectively removes watermarks with minimal changes to images. For high perturbation watermarking methods where notable changes are applied to images, the diffusion purification attack is not effective. In this case, we develop a model substitution adversarial attack that can successfully remove watermarks. Moreover, we show that watermarking methods are vulnerable to spoofing attacks where the attacker aims to have real images (potentially obscene) identified as watermarked ones, damaging the reputation of the developers. In particular, by just having black-box access to the watermarking method, we show that one can generate a watermarked noise image which can be added to the real images to have them falsely flagged as watermarked ones. Finally, we extend our theory to characterize a fundamental trade-off between the robustness and reliability of classifier-based deep fake detectors and demonstrate it through experiments. Code is available at https://github.com/mehrdadsaberi/watermark_robustness.
****
Boya Shi,Zhengqin Xu,Shuai Jia,Chao Ma
Prompt Learning with Quaternion Networks
Multimodal pre-trained models have shown impressive potential in enhancing performance on downstream tasks. However, existing fusion strategies for modalities primarily rely on explicit interaction structures that fail to capture the diverse aspects and patterns inherent in input data. This yields limited performance in zero-shot contexts, especially when fine-grained classifications and abstract interpretations are required. To address this, we propose an effective approach, namely Prompt Learning with Quaternion Networks (QNet), for semantic alignment across diverse modalities. QNet employs a quaternion hidden space where the mutually orthogonal imaginary axes capture rich intermodal semantic spatial correlations from various perspectives. Hierarchical features across multilayers are utilized to encode intricate interdependencies within various modalities with reduced parameters. Our experiments on 11 datasets demonstrate that QNet outperforms state-of-the-art prompt learning techniques in base-to-novel generalization, cross-dataset transfer, and domain transfer scenarios with fewer learnable parameters. The source code is available at https://github.com/VISION-SJTU/QNet.
****
Hritik Bansal,John Dang,Aditya Grover
Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models
Aligning large language models (LLMs) with human values and intents critically involves the use of human or AI feedback. While dense feedback annotations are expensive to acquire and integrate, sparse feedback presents a structural design choice between ratings (e.g., score Response A on a scale of 1-7) and rankings (e.g., is Response A better than Response B?). In this work, we analyze the effect of this design choice for the alignment and evaluation of LLMs. We uncover an inconsistency problem wherein the preferences inferred from ratings and rankings significantly disagree 60% for both human and AI annotators. Our subsequent analysis identifies various facets of annotator biases that explain this phenomena such as human annotators would rate denser responses higher while preferring accuracy during pairwise judgments, for a particular comparison instance. To our surprise, we observe that the choice of feedback protocol has a significant effect on the evaluation of aligned LLMs. In particular, we find that LLMs that leverage rankings data for alignment (say model X) are preferred over those that levera

ge ratings data (say model Y), with a rank-based evaluation protocol (is X/Y's response better than reference response?) but not with a rating-based evaluation protocol (score Rank X/Y's response on a scale of 1-7). Our findings thus shed light on critical gaps in methods for evaluating the real-world utility of language models and their strong dependence on the feedback protocol used for alignment. Our code and data are available at \url{https://github.com/Hritikbansal/sparse_feedback}.
****
Yujia Qin,Shihao Liang,Yining Ye,Kunlun Zhu,Lan Yan,Yaxi Lu,Yankai Lin,Xin Cong,Xiangru Tang,Bill Qian,Sihan Zhao,Lauren Hong,Runchu Tian,Ruobing Xie,Jie Zhou,Mark Gerstein,dahai li,Zhiyuan Liu,Maosong Sun
ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs
Despite the advancements of open-source large language models (LLMs), e.g., LLaMA, they remain significantly limited in tool-use capabilities, i.e., using external tools (APIs) to fulfill human instructions. The reason is that current instruction tuning largely focuses on basic language tasks but ignores the tool-use domain. This is in contrast to the excellent tool-use capabilities of state-of-the-art (SOTA) closed-source LLMs, e.g., ChatGPT. To bridge this gap, we introduce ToolLLM, a general tool-use framework encompassing data construction, model training, and evaluation. We first present ToolBench, an instruction-tuning dataset for tool use, which is constructed automatically using ChatGPT. Specifically, the construction can be divided into three stages: (i) API collection: we collect 16,464 real-world RESTful APIs spanning 49 categories from RapidAPI Hub; (ii) instruction generation: we prompt ChatGPT to generate diverse instructions involving these APIs, covering both single-tool and multi-tool scenarios; (iii) solution path annotation: we use ChatGPT to search for a valid solution path (chain of API calls) for each instruction. To enhance the reasoning capabilities of LLMs, we develop a novel depth-first search-based decision tree algorithm. It enables LLMs to evaluate multiple reasoning traces and expand the search space. Moreover, to evaluate the tool-use capabilities of LLMs, we develop an automatic evaluator: ToolEval. Based on ToolBench, we fine-tune LLaMA to obtain an LLM ToolLLaMA, and equip it with a neural API retriever to recommend appropriate APIs for each instruction. Experiments show that ToolLLaMA demonstrates a remarkable ability to execute complex instructions and generalize to unseen APIs, and exhibits comparable performance to ChatGPT. Our ToolLLaMA also demonstrates strong zero-shot generalization ability in an out-of-distribution tool-use dataset: APIBench.
****
Haitao Yang,Xiangru Huang,Bo Sun,Chandrajit L. Bajaj,Qixing Huang
GenCorres: Consistent Shape Matching via Coupled Implicit-Explicit Shape Generative Models
This paper introduces GenCorres, a novel unsupervised joint shape matching (JSM) approach. Our key idea is to learn a mesh generator to fit an unorganized deformable shape collection while constraining deformations between adjacent synthetic shapes to preserve geometric structures such as local rigidity and local conformality. GenCorres presents three appealing advantages over existing JSM techniques. First, GenCorres performs JSM among a synthetic shape collection whose size is much bigger than the input shapes and fully leverages the datadriven power of JSM. Second, GenCorres unifies consistent shape matching and pairwise matching (i.e., by enforcing deformation priors between adjacent synthetic shapes). Third, the generator provides a concise encoding of consistent shape correspondences. However, learning a mesh generator from an unorganized shape collection is challenging, requiring a good initialization. GenCorres addresses this issue by learning an implicit generator from the input shapes, which provides intermediate shapes between two arbitrary shapes. We introduce a novel approach for computing correspondences between adjacent implicit surfaces, which we use to regularize the implicit generator. Synthetic shapes of the implicit generator then guide initial fittings (i.e., via template-based deformation) for learning the mesh generator. Experimental results show that GenCorres considerably outperforms state-of-the-art JSM techniques. The synthetic shapes of GenCorres also achieve salient performance gains against state-of-the-art deformable shape generators.

****
Faeze Brahman,Chandra Bhagavatula,Valentina Pyatkin,Jena D. Hwang,Xiang Lorraine
 Li,Hirona Jacqueline Arai,Soumya Sanyal,Keisuke Sakaguchi,Xiang Ren,Yejin Choi
PlaSma: Procedural Knowledge Models for Language-based Planning and Re-Planning
Procedural planning, which entails decomposing a high-level goal into a sequence
 of temporally ordered steps, is an important yet intricate task for machines. I
t involves integrating common-sense knowledge to reason about complex and often
contextualized situations, e.g. ``scheduling a doctor's appointment without a ph
one''. While current approaches show encouraging results using large language mo
dels (LLMs), they are hindered by drawbacks such as costly API calls and reprodu
cibility issues. In this paper, we advocate planning using smaller language mode
ls. We present PlaSma, a novel two-pronged approach to endow small language mode
ls with procedural knowledge and (constrained) language-based planning capabilit
ies. More concretely, we develop *symbolic procedural knowledge distillation* to
 enhance the commonsense knowledge in small language models and an *inference-ti
me algorithm* to facilitate more structured and accurate reasoning. In addition,
 we introduce a new related task, *Replanning*, that requires a revision of a pl
an to cope with a constrained situation. In both the planning and replanning set
tings, we show that orders-of-magnitude smaller models (770M-11B parameters) can
 compete and often surpass their larger teacher models' capabilities. Finally, w
e showcase successful application of PlaSma in an embodied environment, VirtualH
ome.
****
Runyu Zhang,Yang Hu,Na Li
Soft Robust MDPs and Risk-Sensitive MDPs: Equivalence, Policy Gradient, and Samp
le Complexity
Robust Markov Decision Processes (MDPs) and risk-sensitive MDPs are both powerfu
l tools for making decisions in the presence of uncertainties. Previous efforts
have aimed to establish their connections, revealing equivalences in specific fo
rmulations. This paper introduces a new formulation for risk-sensitive MDPs, whi
ch assesses risk in a slightly different manner compared to the classical Markov
 risk measure [Ruszczy■nski 2010], and establishes its equivalence with a class
 of soft robust MDP (RMDP) problems, including the standard RMDP as a special ca
se. Leveraging this equivalence, we further derive the policy gradient theorem f
or both problems, proving gradient domination and global convergence of the exac
t policy gradient method under the tabular setting with direct parameterization.
 This forms a sharp contrast to the Markov risk measure, known to be potentially
 non-gradient-dominant [Huang et al. 2021]. We also propose a sample-based offli
ne learning algorithm, namely the robust fitted-Z iteration (RFZI), for a specif
ic soft RMDP problem with a KL-divergence regularization term (or equivalently t
he risk-sensitive MDP with an entropy risk measure). We showcase its streamlined
design and less stringent assumptions due to the equivalence and analyze its sam
ple complexity.
****
Christopher Mohri,Daniel Andor,Eunsol Choi,Michael Collins,Anqi Mao,Yutao Zhong
Learning to Reject with a Fixed Predictor: Application to Decontextualization
We study the problem of classification with a reject option for a fixed predicto
r, crucial to natural language processing. We introduce a new problem formulatio
n for this scenario, and an algorithm minimizing a new surrogate loss function.
We provide a complete theoretical analysis of the surrogate loss function with a
 strong H-consistency guarantee. For evaluation, we choose the \textit{deconte
xtualization} task, and provide a manually-labelled dataset of $2\mathord,000$ e
xamples. Our algorithm significantly outperforms the baselines considered, with
a ~ 25% improvement in coverage when halving the error rate, which is only
~ 3% away from the theoretical limit.
****
Lei Li,Yekun Chai,Shuohuan Wang,Yu Sun,Hao Tian,Ningyu Zhang,Hua Wu
Tool-Augmented Reward Modeling
Reward modeling (*a.k.a.*, preference modeling) is instrumental for aligning lar
ge language models with human preferences, particularly within the context of re

inforcement learning from human feedback (RLHF). While conventional reward model
s (RMs) have exhibited remarkable scalability, they oft struggle with fundamenta
l functionality such as arithmetic computation, code execution, and factual look
up. In this paper, we propose a tool-augmented preference modeling approach, nam
ed Themis, to address these limitations by empowering RMs with access to externa
l environments, including calculators and search engines. This approach not only
 fosters synergy between tool utilization and reward grading but also enhances i
nterpretive capacity and scoring reliability. Our study delves into the integrat
ion of external tools into RMs, enabling them to interact with diverse external
sources and construct task-specific tool engagement and reasoning traces in an a
utoregressive manner. We validate our approach across a wide range of domains, i
ncorporating seven distinct external tools. Our experimental results demonstrate
 a noteworthy overall improvement of 17.7% across eight tasks in preference rank
ing. Furthermore, our approach outperforms Gopher 280B by 7.3% on TruthfulQA tas
k in zero-shot evaluation. In human evaluations, RLHF trained with Themis attain
s an average win rate of 32% when compared to baselines across four distinct tas
ks. Additionally, we provide a comprehensive collection of tool-related RM datas
ets, incorporating data from seven distinct tool APIs, totaling 15,000 instances
. We have made the code, data, and model checkpoints publicly available to facil
itate and inspire further research advancements (https://github.com/ernie-resear
ch/Tool-Augmented-Reward-Model).
****
Mitchell Wortsman,Peter J Liu,Lechao Xiao,Katie E Everett,Alexander A Alemi,Ben
Adlam,John D Co-Reyes,Izzeddin Gur,Abhishek Kumar,Roman Novak,Jeffrey Pennington
,Jascha Sohl-Dickstein,Kelvin Xu,Jaehoon Lee,Justin Gilmer,Simon Kornblith
Small-scale proxies for large-scale Transformer training instabilities
Teams that have trained large Transformer-based models have reported training in
stabilities at large scale that did not appear when training with the same hyper
parameters at smaller scales. Although the causes of such instabilities are of s
cientific interest, the amount of resources required to reproduce them has made
investigation difficult. In this work, we seek ways to reproduce and study train
ing instability at smaller scales. First, we focus on two sources of training in
stability described in previous work: the growth of logits in attention layers (
Dehghani et al., 2023) and divergence of the output logits from the log probabil
ities (Chowdhery et al., 2022). By measuring the relationship between learning r
ate and loss across scales, we show that these instabilities also appear in smal
l models when training at high learning rates, and that mitigations previously e
mployed at large scales are equally effective in this regime. This prompts us to
 investigate the extent to which other known optimizer and model interventions i
nfluence the sensitivity of the final loss to changes in the learning rate. To t
his end, we study methods such as warm-up, weight decay, and the MuParam (Yang e
t al., 2022), and combine techniques to train small models that achieve similar
losses across orders of magnitude of learning rate variation. Finally, to conclu
de our exploration we study two cases where instabilities can be predicted befor
e they emerge by examining the scaling behavior of model characteristics such as
 activation and gradient norms.
****
Petr Mokrov,Alexander Korotin,Alexander Kolesov,Nikita Gushchin,Evgeny Burnaev
Energy-guided Entropic Neural Optimal Transport
Energy-based models (EBMs) are known in the Machine Learning community for decad
es. Since the seminal works devoted to EBMs dating back to the noughties, there
have been a lot of efficient methods which solve the generative modelling proble
m by means of energy potentials (unnormalized likelihood functions). In contrast
, the realm of Optimal Transport (OT) and, in particular, neural OT solvers is m
uch less explored and limited by few recent works (excluding WGAN-based approach
es which utilize OT as a loss function and do not model OT maps themselves). In
our work, we bridge the gap between EBMs and Entropy-regularized OT. We present
a novel methodology which allows utilizing the recent developments and technical
 improvements of the former in order to enrich the latter. From the theoretical
perspective, we prove generalization bounds for our technique. In practice, we v

alidate its applicability in toy 2D and image domains. To showcase the scalability, we empower our method with a pre-trained StyleGAN and apply it to high-res AFHQ 512×512 unpaired I2I translation. For simplicity, we choose simple short- and long-run EBMs as a backbone of our Energy-guided Entropic OT approach, leaving the application of more sophisticated EBMs for future research. Our code is available at: https://github.com/PetrMokrov/Energy-guided-Entropic-OT
****
LIN Yong,Lu Tan,Yifan HAO,Ho Nam Wong,Hanze Dong,WEIZHONG ZHANG,Yujiu Yang,Tong Zhang
Spurious Feature Diversification Improves Out-of-distribution Generalization
Generalization to out-of-distribution (OOD) data is a critical challenge in machine learning. Ensemble-based methods, like weight space ensembles that interpolate model parameters, have been shown to achieve superior OOD performance. However, the underlying mechanism for their effectiveness remains unclear.

In this study, we closely examine WiSE-FT, a popular weight space ensemble method that interpolates between a pre-trained and a fine-tuned model. We observe an unexpected ``FalseFalseTrue" phenomenon, in which WiSE-FT successfully corrects many cases where each individual model makes incorrect predictions, which contributes significantly to its OOD effectiveness. To gain further insights, we conduct theoretical analysis in a multi-class setting with a large number of spurious features. Our analysis predicts the above phenomenon and it further shows that ensemble-based models reduce prediction errors in the OOD settings by utilizing a more diverse set of spurious features. Contrary to the conventional wisdom that focuses on learning invariant features for better OOD performance, our findings suggest that incorporating a large number of diverse spurious features weakens their individual contributions, leading to improved overall OOD generalization performance. Additionally, our findings provide the first explanation for the mysterious phenomenon of weight space ensembles outperforming output space ensembles in OOD. Empirically we demonstrate the effectiveness of utilizing diverse spurious features on a MultiColorMNIST dataset, and our experimental results are consistent with the theoretical analysis.

Building upon the new theoretical insights into the efficacy of ensemble methods, we further identify an issue of WiSE-FT caused by the overconfidence of fine-tuned models in OOD situations. This overconfidence magnifies the fine-tuned model's incorrect prediction, leading to deteriorated OOD ensemble performance. To remedy this problem, we propose a novel method called BAlaNced averaGing (BANG) to mitigate the overconfidence problem, which significantly enhances the OOD performance of WiSE-FT.
****
Tamir David Hay,Lior Wolf
Dynamic Layer Tying for Parameter-Efficient Transformers
In the pursuit of reducing the number of trainable parameters in deep transformer networks, we employ Reinforcement Learning to dynamically select layers during training and tie them together. Every few iterations, the RL agent is asked whether to train each layer i independently or to copy the weights of a previous layer $j<i$. This facilitates weight sharing, reduces the number of trainable parameters, and also serves as an effective regularization technique. Experimental evaluations validate that our model modestly outperforms the baseline transformer model with regard to perplexity and drastically reduces the number of trainable parameters. In particular, the memory consumption during training is up to one order of magnitude less than the conventional training method.
****
Renrui Zhang,Jiaming Han,Chris Liu,Aojun Zhou,Pan Lu,Hongsheng Li,Peng Gao,Yu Qiao
LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention
With the rising tide of large language models (LLMs), there has been a growing interest in developing general-purpose instruction-following models, e.g., ChatGP

T. To this end, we present LLaMA-Adapter, a lightweight adaption method for effi
cient instruction tuning of LLaMA. Using 52K self-instruct demonstrations, LLaMA
-Adapter only introduces 1.2M learnable parameters upon the frozen LLaMA 7B mode
l, and costs less than one hour for fine-tuning. Specifically, a zero-initialize
d attention mechanism is proposed. It adopts a learnable zero gating to adaptive
ly inject the instructional cues into LLaMA within self-attention layers, contri
buting to a stable training process and superior final performance. In this way,
 LLaMA-Adapter can generate high-quality responses to diverse language instructi
ons, comparable to Alpaca with fully fine-tuned 7B parameters. Besides language
commands, by incorporating an image encoder, our approach can be simply extended
 to a multi-modal LLM for image-conditioned instruction following, which achieve
s superior multi-modal reasoning capacity on several popular benchmarks (MME, MM
Bench, LVLM-eHub). Furthermore, we also verify the proposed zero-initialized att
ention mechanism for fine-tuning other pre-trained models (ViT, RoBERTa, CLIP) o
n traditional vision and language tasks, demonstrating the effectiveness and gen
eralizability of our approach.
****
Valentyn Melnychuk,Dennis Frauen,Stefan Feuerriegel
Bounds on Representation-Induced Confounding Bias for Treatment Effect Estimatio
n
State-of-the-art methods for conditional average treatment effect (CATE) estimat
ion make widespread use of representation learning. Here, the idea is to reduce
the variance of the low-sample CATE estimation by a (potentially constrained) lo
w-dimensional representation. However, low-dimensional representations can lose
information about the observed confounders and thus lead to bias, because of whi
ch the validity of representation learning for CATE estimation is typically viol
ated. In this paper, we propose a new, representation-agnostic refutation framew
ork for estimating bounds on the representation-induced confounding bias that co
mes from dimensionality reduction (or other constraints on the representations)
in CATE estimation. First, we establish theoretically under which conditions CAT
E is non-identifiable given low-dimensional (constrained) representations. Secon
d, as our remedy, we propose a neural refutation framework which performs partia
l identification of CATE or, equivalently, aims at estimating lower and upper bo
unds of the representation-induced confounding bias. We demonstrate the effectiv
eness of our bounds in a series of experiments. In sum, our refutation framework
 is of direct relevance in practice where the validity of CATE estimation is of
importance.
****
Zeyu Tang,Jialu Wang,Yang Liu,Peter Spirtes,Kun Zhang
Procedural Fairness Through Decoupling Objectionable Data Generating Components
We reveal and address the frequently overlooked yet important issue of _disguise
d procedural unfairness_, namely, the potentially inadvertent alterations on the
 behavior of neutral (i.e., not problematic) aspects of data generating process,
 and/or the lack of procedural assurance of the greatest benefit of the least ad
vantaged individuals. Inspired by John Rawls's advocacy for _pure procedural jus
tice_ (Rawls, 1971; 2001), we view automated decision-making as a microcosm of s
ocial institutions, and consider how the data generating process itself can sati
sfy the requirements of procedural fairness. We propose a framework that decoupl
es the objectionable data generating components from the neutral ones by utilizi
ng reference points and the associated value instantiation rule. Our findings hi
ghlight the necessity of preventing _disguised procedural unfairness_, drawing a
ttention not only to the objectionable data generating components that we aim to
 mitigate, but also more importantly, to the neutral components that we intend t
o keep unaffected.
****
Xingyu Zhou,Sayak Ray Chowdhury
On Differentially Private Federated Linear Contextual Bandits
We consider cross-silo federated linear contextual bandit (LCB) problem under di
fferential privacy, where multiple silos interact with their respective local us
ers and communicate via a central server to realize collaboration without sacrif

icing each user's privacy. We identify three issues in the state-of-the-art~\cit ep{dubey2020differentially}: (i) failure of claimed privacy protection, (ii) inc orrect regret bound due to noise miscalculation and (iii) ungrounded communicati on cost.
To resolve these issues, we take a two-step approach. First, we design an algori thmic framework consisting of a generic federated LCB algorithm and flexible pri vacy protocols. Then, leveraging the proposed framework, we study federated LCBs under two different privacy constraints. We first establish privacy and regret guarantees under silo-level local differential privacy, which fix the issues pre sent in state-of-the-art algorithm.
To further improve the regret performance, we next consider shuffle model of dif ferential privacy, under which we show that our algorithm can achieve nearly ``o ptimal'' regret without a trusted server.
We accomplish this via two different schemes -- one relies on a new result on p rivacy amplification via shuffling for DP mechanisms and another one leverages t he integration of a shuffle protocol for vector sum into the tree-based mechanis m, both of which might be of independent interest. Finally, we support our theor etical results with
numerical evaluations over contextual bandit instances generated from both synth etic and real-life data.
****
Suhan Shetty,Teng Xue,Sylvain Calinon
Generalized Policy Iteration using Tensor Approximation for Hybrid Control
Control of dynamic systems involving hybrid actions is a challenging task in rob otics. To address this, we present a novel algorithm called Generalized Policy Iteration using Tensor Train (TTPI) that belongs to the class of Approximate Dyn amic Programming (ADP). We use a low-rank tensor approximation technique called Tensor Train (TT) to approximate the state-value and advantage function which en ables us to efficiently handle hybrid systems. We demonstrate the superiority of our approach over previous baselines for some benchmark problems with hybrid ac tion spaces. Additionally, the robustness and generalization of the policy for h ybrid systems are showcased through a real-world robotics experiment involving a non-prehensile manipulation task which is considered to be a highly challenging control problem.
****
Guillaume Bono,Leonid Antsfeld,Boris Chidlovskii,Philippe Weinzaepfel,Christian Wolf
End-to-End (Instance)-Image Goal Navigation through Correspondence as an Emergen t Phenomenon
Most recent work in goal oriented visual navigation resorts to large-scale machi ne learning in simulated environments. The main challenge lies in learning compa ct representations generalizable to unseen environments and in learning high-cap acity perception modules capable of reasoning on high-dimensional input. The lat ter is particularly difficult when the goal is not given as a category ("ObjectN av") but as an exemplar image ("ImageNav"), as the perception module needs to le arn a comparison strategy requiring to solve an underlying visual correspondence problem. This has been shown to be difficult from reward alone or with standard auxiliary tasks. We address this problem through a sequence of two pretext task s, which serve as a prior for what we argue is one of the main bottleneck in per ception, extremely wide-baseline relative pose estimation and visibility predict ion in complex scenes. The first pretext task, cross-view completion is a proxy for the underlying visual correspondence problem, while the second task addresse s goal detection and finding directly. We propose a new dual encoder with a larg e-capacity binocular ViT model and show that correspondence solutions naturally emerge from the training signals. Experiments show significant improvements and SOTA performance on the two benchmarks, ImageNav and the Instance-ImageNav vari ant, where camera intrinsics and height differ between observation and goal.
****
Omer Nahum,Gali Noti,David C. Parkes,Nir Rosenfeld
Decongestion by Representation: Learning to Improve Economic Welfare in Marketpl

aces
Congestion is a common failure mode of markets, where consumers compete ineffici
ently on the same subset of goods (e.g., chasing the same small set of properti
es on a vacation rental platform). The typical economic story is that prices de
congest by balancing supply and demand. But in modern online marketplaces, price
s are typically set in a decentralized way by sellers, and the information about
 items is inevitably partial. The power of a platform is limited to controlling
 representations---the subset of information about items presented by default
to users. This motivates the present study of *decongestion by representation*,
where a platform seeks to learn representations that reduce congestion and thus
 improve social welfare. The technical challenge is twofold: relying only on rev
ealed preferences from the choices of consumers, rather than true preferences; a
nd the combinatorial problem associated with representations that determine the
  features to reveal in the default view. We tackle both challenges by proposin
g a *differentiable proxy of welfare* that can be trained end-to-end on consumer
 choice data. We develop sufficient conditions for when decongestion promotes we
lfare, and present the results of extensive experiments on both synthetic and re
al data that demonstrate the utility of our approach.
****
Francois Charton
Learning the greatest common divisor: explaining transformer predictions
The predictions of small transformers, trained to calculate the greatest common
divisor (GCD) of two positive integers, can be fully characterized by looking at
 model inputs and outputs.
As training proceeds, the model learns a list $\mathcal D$ of integers, products
 of divisors of the base used to represent integers and small primes, and predic
ts the largest element of $\mathcal D$ that divides both inputs.
Training distributions impact performance. Models trained from uniform operands
only learn a handful of GCD (up to 38 GCD ≤100). Log-uniform operands boo
st performance to 73 GCD ≤ 100, and a log-uniform distribution of outcome
s (i.e. GCD) to 91. However, training from uniform (balanced) GCD breaks expla
inability.
****
Yuchen Hu,CHEN CHEN,Chao-Han Huck Yang,Ruizhe Li,Chao Zhang,Pin-Yu Chen,Ensiong
Chng
Large Language Models are Efficient Learners of Noise-Robust Speech Recognition
Recent advances in large language models (LLMs) have promoted generative error c
orrection (GER) for automatic speech recognition (ASR), which leverages the rich
 linguistic knowledge and powerful reasoning ability of LLMs to improve recognit
ion results. The latest work proposes a GER benchmark with "HyPoradise" dataset
to learn the mapping from ASR N-best hypotheses to ground-truth transcription by
 efficient LLM finetuning, which shows great effectiveness but lacks specificity
 on noise-robust ASR. In this work, we extend the benchmark to noisy conditions
and investigate if we can teach LLMs to perform denoising for GER just like what
 robust ASR do, where one solution is introducing noise information as a conditi
oner into LLM. However, directly incorporating noise embeddings from audio encod
er could harm the LLM tuning due to cross-modality gap. To this end, we propose
to extract a language-space noise embedding from the N-best list to represent th
e noise conditions of source speech, which can promote the denoising process in
GER. Furthermore, in order to enhance its representation ability of audio noise,
 we design a knowledge distillation (KD) approach via mutual information estimat
ion to distill the real noise information in audio embeddings to our language em
bedding. Experiments on various latest LLMs demonstrate our approach achieves a
new breakthrough with up to 53.9% correction improvement in terms of word error
rate while with limited training data. Analysis shows that our language-space no
ise embedding can well represent the noise conditions of source speech, under wh
ich off-the-shelf LLMs show strong ability of language-space denoising.
****
Yongtao Wu,Fanghui Liu,Carl-Johann Simon-Gabriel,Grigorios Chrysos,Volkan Cevher
Robust NAS under adversarial training: benchmark, theory, and beyond

Recent developments in neural architecture search (NAS) emphasize the significan
ce of considering robust architectures against malicious data. However, there is
 a notable absence of benchmark evaluations and theoretical guarantees for searc
hing these robust architectures, especially when adversarial training is conside
red. In this work, we aim to address these two challenges, making twofold contri
butions. First, we release a comprehensive data set that encompasses both clean
accuracy and robust accuracy for a vast array of adversarially trained networks
from the NAS-Bench-201 search space on image datasets. Then, leveraging the neur
al tangent kernel (NTK) tool from deep learning theory, we establish a generaliz
ation theory for searching architecture in terms of clean accuracy and robust ac
curacy under multi-objective adversarial training. We firmly believe that our be
nchmark and theoretical insights will significantly benefit the NAS community th
rough reliable reproducibility, efficient assessment, and theoretical foundation
, particularly in the pursuit of robust architectures.
****
Canyu Chen,Kai Shu
Can LLM-Generated Misinformation Be Detected?
The advent of Large Language Models (LLMs) has made a transformative impact. How
ever, the potential that LLMs such as ChatGPT can be exploited to generate misin
formation has posed a serious concern to online safety and public trust. A funda
mental research question is: will LLM-generated misinformation cause more harm t
han human-written misinformation? We propose to tackle this question from the pe
rspective of detection difficulty. We first build a taxonomy of LLM-generated mi
sinformation. Then we categorize and validate the potential real-world methods f
or generating misinformation with LLMs. Then, through extensive empirical invest
igation, we discover that LLM-generated misinformation can be harder to detect f
or humans and detectors compared to human-written misinformation with the same s
emantics, which suggests it can have more deceptive styles and potentially cause
 more harm. We also discuss the implications of our discovery on combating misin
formation in the age of LLMs and the countermeasures.
****
Ian Gemp,Luke Marris,Georgios Piliouras
Approximating Nash Equilibria in Normal-Form Games via Stochastic Optimization
We propose the first loss function for approximate Nash equilibria of normal-for
m games that is amenable to unbiased Monte Carlo estimation. This construction a
llows us to deploy standard non-convex stochastic optimization techniques for ap
proximating Nash equilibria, resulting in novel algorithms with provable guaran
tees. We complement our theoretical analysis with experiments demonstrating that
 stochastic gradient descent can outperform previous state-of-the-art approaches
.
****
Changyou Chen,Han Ding,Bunyamin Sisman,Yi Xu,Ouye Xie,Benjamin Z. Yao,Son Dinh T
ran,Belinda Zeng
Diffusion Models for Multi-Task Generative Modeling
Generative modeling via diffusion-based models has been achieving state-of-the-a
rt results on various generation tasks. Most existing diffusion models, however,
 are limited to a single-generation modeling. Can we generalize diffusion models
 with the ability of multi-task generative training for more generalizable model
ing? In this paper, we propose a principled way to define a diffusion model for
this purpose by constructing a unified multi-task diffusion model in a common {\
em diffusion space}. We define the forward diffusion process to be driven by an
information aggregation from multiple types of task-data, {\it e.g.}, images for
 a generation task and labels for a classification task. In the reverse process,
 we enforce information sharing by parameterizing a shared backbone denoising ne
twork with additional task-specific decoder heads. Such a structure can simultan
eously learn to generate different types of multi-task data with a multi-task lo
ss, which is derived from a multi-task variational lower bound that generalizes
the standard diffusion model. We propose several multi-task generation settings
to verify our framework, including image transition, masked-image training, join
t image-label and joint image-representation generative modeling. Extensive expe

rimental results on ImageNet indicate the effectiveness of our framework for var
ious multi-task generative modeling, which we believe is an important research d
irection worthy of more future explorations.
****
Marcus J. Min,Yangruibo Ding,Luca Buratti,Saurabh Pujar,Gail Kaiser,Suman Jana,B
aishakhi Ray
Beyond Accuracy: Evaluating Self-Consistency of Code Large Language Models with
IdentityChain
Code Large Language Models (Code LLMs) are being increasingly employed in real-l
ife applications, so evaluating them is critical. While the conventional accurac
y evaluates the performance of Code LLMs on a set of individual tasks, their sel
f-consistency across different tasks is overlooked. Intuitively, a trustworthy m
odel should be self-consistent when generating natural language specifications f
or its own code and generating code for its own specifications. Failure to prese
rve self-consistency reveals a lack of understanding of the shared semantics und
erlying natural language and programming language, and therefore undermines the
trustworthiness of a model. In this paper, we first formally define the self-con
sistency of Code LLMs and then design a framework, IdentityChain, which effectiv
ely and efficiently evaluates the self-consistency and conventional accuracy of
a model at the same time. We study eleven Code LLMs and show that they fail to p
reserve self-consistency, which is indeed a distinct aspect from conventional ac
curacy. Furthermore, we show that IdentityChain can be used as a model debugging
 tool to expose weaknesses of Code LLMs by demonstrating three major weaknesses
that we identify in current models using IdentityChain. Our code is available at
 https://github.com/marcusm117/IdentityChain.
****
Felix Petersen,Aashwin Ananda Mishra,Hilde Kuehne,Christian Borgelt,Oliver Deuss
en,Mikhail Yurochkin
Uncertainty Quantification via Stable Distribution Propagation
We propose a new approach for propagating stable probability distributions throu
gh neural networks. Our method is based on local linearization, which we show to
 be an optimal approximation in terms of total variation distance for the ReLU n
on-linearity. This allows propagating Gaussian and Cauchy input uncertainties th
rough neural networks to quantify their output uncertainties. To demonstrate the
 utility of propagating distributions, we apply the proposed method to predictin
g calibrated confidence intervals and selective prediction on out-of-distributio
n data. The results demonstrate a broad applicability of propagating distributio
ns and show the advantages of our method over other approaches such as moment ma
tching.
****
Yuning You,Ruida Zhou,Jiwoong Park,Haotian Xu,Chao Tian,Zhangyang Wang,Yang Shen
Latent 3D Graph Diffusion
Generating 3D graphs of symmetry-group equivariance is of intriguing potential i
n broad applications from machine vision to molecular discovery. Emerging approa
ches adopt diffusion generative models (DGMs) with proper re-engineering to capt
ure 3D graph distributions. In this paper, we raise an orthogonal and fundamenta
l question of in what (latent) space we should diffuse 3D graphs. ❶ We motivate
the study with theoretical analysis showing that the performance bound of 3D gra
ph diffusion can be improved in a latent space versus the original space, provid
ed that the latent space is of (i) low dimensionality yet (ii) high quality (i.e
., low reconstruction error) and DGMs have (iii) symmetry preservation as an ind
uctive bias. ❷ Guided by the theoretical guidelines, we propose to perform 3D gr
aph diffusion in a low-dimensional latent space, which is learned through cascad
ed 2D–3D graph autoencoders for low-error reconstruction and symmetry-group inva
riance. The overall pipeline is dubbed latent 3D graph diffusion. ❸ Motivated by
 applications in molecular discovery, we further extend latent 3D graph diffusio
n to conditional generation given SE(3)-invariant attributes or equivariant 3D o
bjects. ❹ We also demonstrate empirically that out-of-distribution conditional g
eneration can be further improved by regularizing the latent space via graph sel
f-supervised learning. We validate through comprehensive experiments that our me

thod generates 3D molecules of higher validity / drug-likeliness and comparable or better conformations / energetics, while being an order of magnitude faster in training. Codes are released at https://github.com/Shen-Lab/LDM-3DG.
****
Rafael Alberto Rivera Soto,Kailin Koch,Aleem Khan,Barry Y. Chen,Marcus Bishop,Nicholas Andrews
Few-Shot Detection of Machine-Generated Text using Style Representations
The advent of instruction-tuned language models that convincingly mimic human writing poses a significant risk of abuse. For example, such models could be used for plagiarism, disinformation, spam, or phishing. However, such abuse may be counteracted with the ability to detect whether a piece of text was composed by a language model rather than a human. Some previous approaches to this problem have relied on supervised methods trained on corpora of confirmed human and machine-written documents. Unfortunately, model under-specification poses an unavoidable challenge for such detectors, making them brittle in the face of data shifts, such as the release of further language models producing still more fluent text than the models used to train the detectors. Other previous approaches require access to the models that generated the text to be detected at inference or detection time, which is often impractical. In light of these challenge, we pursue a fundamentally different approach not relying on samples from language models of concern at training time. Instead, we propose to leverage representations of writing style estimated from human-authored text. Indeed, we find that features effective at distinguishing among human authors are also effective at distinguishing human from machine authors, including state of the art large language models like Llama 2, ChatGPT, and GPT-4. Furthermore, given handfuls of examples composed by each of several specific language models of interest, our approach affords the ability to predict which model specifically generated a given document.
****
Li Meng,Morten Goodwin,Anis Yazidi,Paal E. Engelstad
State Representation Learning Using an Unbalanced Atlas
The manifold hypothesis posits that high-dimensional data often lies on a lower-dimensional manifold and that utilizing this manifold as the target space yields more efficient representations. While numerous traditional manifold-based techniques exist for dimensionality reduction, their application in self-supervised learning has witnessed slow progress. The recent MSimCLR method combines manifold encoding with SimCLR but requires extremely low target encoding dimensions to outperform SimCLR, limiting its applicability. This paper introduces a novel learning paradigm using an unbalanced atlas (UA), capable of surpassing state-of-the-art self-supervised learning approaches. We investigated and engineered the DeepInfomax with an unbalanced atlas (DIM-UA) method by adapting the Spatiotemporal DeepInfomax (ST-DIM) framework to align with our proposed UA paradigm. The efficacy of DIM-UA is demonstrated through training and evaluation on the Atari Annotated RAM Interface (AtariARI) benchmark, a modified version of the Atari 2600 framework that produces annotated image samples for representation learning. The UA paradigm improves existing algorithms significantly as the number of target encoding dimensions grows. For instance, the mean F1 score averaged over categories of DIM-UA is~75% compared to ~70% of ST-DIM when using 16384 hidden units.
****
Liyang Zhu,Meng Ding,Vaneet Aggarwal,Jinhui Xu,Di Wang
Improved Analysis of Sparse Linear Regression in Local Differential Privacy Model
In this paper, we revisit
the problem of sparse linear regression in the local differential privacy (LDP) model. Existing research in the non-interactive and sequentially local models has focused on obtaining the lower bounds for the case where the underlying parameter is 1-sparse, and extending such bounds to the more general k-sparse case has proven to be challenging. Moreover, it is unclear whether efficient non-interactive LDP (NLDP) algorithms exist. To address these issues,
we first consider the problem in the ϵ non-interactive LDP model and provide a lower bound of $\Omega(\frac{\sqrt{dk\log d}}{\sqrt{n}\epsilon})$ on th

e ℓ_2-norm estimation error for sub-Gaussian data, where n is the sample
size and d is the dimension of the space.
We propose an innovative NLDP algorithm, the very first of its kind for the prob
lem. As a remarkable outcome, this algorithm also yields a novel and highly effi
cient estimator as a valuable by-product. Our algorithm achieves an upper bound
of $\tilde{O}({\frac{d\sqrt{k}}{\sqrt{n}\epsilon}})$ for the estimation error wh
en the data is sub-Gaussian, which can be further improved by a factor of $O(\s
qrt{d})$ if the server has additional public but unlabeled data.
For the sequentially interactive LDP model, we show a similar lower bound of $\O
mega({\frac{\sqrt{dk}}{\sqrt{n}\epsilon}})$. As for the upper bound, we rectify
a previous method and show that it is possible to achieve a bound of $\tilde{O}(
\frac{k\sqrt{d}}{\sqrt{n}\epsilon})$. Our findings reveal fundamental difference
s between the non-private case, central DP model, and local DP model in the spar
se linear regression problem.
****
Renat Sergazinov,Elizabeth Chun,Valeriya Rogovchenko,Nathaniel J Fernandes,Nicho
las Kasman,Irina Gaynanova
GlucoBench: Curated List of Continuous Glucose Monitoring Datasets with Predicti
on Benchmarks
The rising rates of diabetes necessitate innovative methods for its management.
Continuous glucose monitors (CGM) are small medical devices that measure blood g
lucose levels at regular intervals providing insights into daily patterns of glu
cose variation. Forecasting of glucose trajectories based on CGM data holds the
 potential to substantially improve diabetes management, by both refining artifi
cial pancreas systems and enabling individuals to make adjustments based on pre
dictions to maintain optimal glycemic range. Despite numerous methods proposed f
or CGM-based glucose trajectory prediction, these methods are typically evaluate
d on small, private datasets, impeding reproducibility, further research, and pr
actical adoption. The absence of standardized prediction tasks and systematic co
mparisons between methods has led to uncoordinated research efforts, obstructing
 the identification of optimal tools for tackling specific challenges. As a resu
lt, only a limited number of prediction methods have been implemented in clinica
l practice.

To address these challenges, we present a comprehensive resource that provides (
1) a consolidated repository of curated publicly available CGM datasets to foste
r reproducibility and accessibility; (2) a standardized task list to unify resea
rch objectives and facilitate coordinated efforts; (3) a set of benchmark models
 with established baseline performance, enabling the research community to objec
tively gauge new methods' efficacy; and (4) a detailed analysis of performance-i
nfluencing factors for model development. We anticipate these resources to prope
l collaborative research endeavors in the critical domain of CGM-based glucose p
redictions. Our code is available online at github.com/IrinaStatsLab/GlucoBench.
****
Xingxuan Li,Ruochen Zhao,Yew Ken Chia,Bosheng Ding,Shafiq Joty,Soujanya Poria,Li
dong Bing
Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapti
ng over Heterogeneous Sources
We present chain-of-knowledge (CoK), a novel framework that augments large langu
age models (LLMs) by dynamically incorporating grounding information from hetero
geneous sources. It results in more factual rationales and reduced hallucination
 in generation.
Specifically, CoK consists of three stages: reasoning preparation, dynamic knowl
edge adapting, and answer consolidation.
Given a knowledge-intensive question, CoK first prepares several preliminary rat
ionales and answers while identifying the relevant knowledge domains.
If there is no majority consensus among the answers from samples, CoK corrects t
he rationales step by step by adapting knowledge from the identified domains.
These corrected rationales can plausibly serve as a better foundation for the fi
nal answer consolidation.

Unlike prior studies that primarily use unstructured data, CoK also leverages structured knowledge sources such as Wikidata and tables that provide more reliable factual information.
To access both unstructured and structured knowledge sources in the dynamic knowledge adapting stage, we propose an adaptive query generator that allows the generation of queries for various types of query languages, including SPARQL, SQL, and natural sentences. Moreover, to minimize error propagation between rationales, CoK corrects the rationales progressively using preceding corrected rationales to generate and correct subsequent rationales.
Extensive experiments show that CoK consistently improves the performance of LLMs on knowledge-intensive tasks across different domains.
****
Mengxi Ya,Yiming Li,Tao Dai,Bin Wang,Yong Jiang,Shu-Tao Xia
Towards Faithful XAI Evaluation via Generalization-Limited Backdoor Watermark
Saliency-based representation visualization (SRV) ($e.g.$, Grad-CAM) is one of the most classical and widely adopted explainable artificial intelligence (XAI) methods for its simplicity and efficiency. It can be used to interpret deep neural networks by locating saliency areas contributing the most to their predictions. However, it is difficult to automatically measure and evaluate the performance of SRV methods due to the lack of ground-truth salience areas of samples. In this paper, we revisit the backdoor-based SRV evaluation, which is currently the only feasible method to alleviate the previous problem. We first reveal its \emph{implementation limitations} and \emph{unreliable nature} due to the trigger generalization of existing backdoor watermarks. Given these findings, we propose a generalization-limited backdoor watermark (GLBW), based on which we design a more faithful XAI evaluation. Specifically, we formulate the training of watermarked DNNs as a min-max problem, where we find the `worst' potential trigger (with the highest attack effectiveness and differences from the ground-truth trigger) via inner maximization and minimize its effects and the loss over benign and poisoned samples via outer minimization in each iteration. In particular, we design an adaptive optimization method to find desired potential triggers in each inner maximization. Extensive experiments on benchmark datasets are conducted, verifying the effectiveness of our generalization-limited watermark. Our codes are available at \url{https://github.com/yamengxi/GLBW}.
****
Changyao Tian,Chenxin Tao,Jifeng Dai,Hao Li,Ziheng Li,Lewei Lu,Xiaogang Wang,Hongsheng Li,Gao Huang,Xizhou Zhu
ADDP: Learning General Representations for Image Recognition and Generation with Alternating Denoising Diffusion Process
Image recognition and generation have long been developed independently of each other. With the recent trend towards general-purpose representation learning, the development of general representations for both recognition and generation tasks is also promoted. However, preliminary attempts mainly focus on generation performance, but are still inferior on recognition tasks. These methods are modeled in the vector-quantized (VQ) space, whereas leading recognition methods use pixels as inputs. Our key insights are twofold: *(1) pixels as inputs are crucial for recognition tasks; (2) VQ tokens as reconstruction targets are beneficial for generation tasks.* These observations motivate us to propose an **Alternating Denoising Diffusion Process (ADDP)** that integrates these two spaces within a single representation learning framework. In each denoising step, our method first decodes pixels from previous VQ tokens, then generates new VQ tokens from the decoded pixels. The diffusion process gradually masks out a portion of VQ tokens to construct the training samples. The learned representations can be used to generate diverse high-fidelity images and also demonstrate excellent transfer performance on recognition tasks. Extensive experiments show that our method achieves competitive performance on unconditional generation, ImageNet classification, COCO detection, and ADE20k segmentation. Importantly, our method represents *the first successful development* of general representations applicable to both generation and dense recognition tasks. Code shall be released.
****

Chuanqing Wang,Di Wu,Chaoming Fang,Jie Yang,Mohamad Sawan
Exploring Effective Stimulus Encoding via Vision System Modeling for Visual Prostheses
Visual prostheses are potential devices to restore vision for blind people, which highly depends on the quality of stimulation patterns of the implanted electrode array. However, existing processing frameworks prioritize the generation of stimulation while disregarding the potential impact of restoration effects and fail to assess the quality of the generated stimulation properly. In this paper, we propose for the first time an end-to-end visual prosthesis framework (StimuSEE) that generates stimulation patterns with proper quality verification using V1 neuron spike patterns as supervision. StimuSEE consists of a retinal network to predict the stimulation pattern, a phosphene model, and a primary vision system network (PVS-net) to simulate the signal processing from the retina to the visual cortex and predict the firing rate of V1 neurons. Experimental results show that the predicted stimulation shares similar patterns to the original scenes, whose different stimulus amplitudes contribute to a similar firing rate with normal cells. Numerically, the predicted firing rate and the recorded response of normal neurons achieve a Pearson correlation coefficient of 0.78.
****
Yunhe Zhang,Yan Sun,Jinyu Cai,Jicong Fan
Deep Orthogonal Hypersphere Compression for Anomaly Detection
Many well-known and effective anomaly detection methods assume that a reasonable decision boundary has a hypersphere shape, which however is difficult to obtain in practice and is not sufficiently compact, especially when the data are in high-dimensional spaces. In this paper, we first propose a novel deep anomaly detection model that improves the original hypersphere learning through an orthogonal projection layer, which ensures that the training data distribution is consistent with the hypersphere hypothesis, thereby increasing the true positive rate and decreasing the false negative rate. Moreover, we propose a bi-hypersphere compression method to obtain a hyperspherical shell that yields a more compact decision region than a hyperball, which is demonstrated theoretically and numerically. The proposed methods are not confined to common datasets such as image and tabular data, but are also extended to a more challenging but promising scenario, graph-level anomaly detection, which learns graph representation with maximum mutual information between the substructure and global structure features while exploring orthogonal single- or bi-hypersphere anomaly decision boundaries. The numerical and visualization results on benchmark datasets demonstrate the superiority of our methods in comparison to many baselines and state-of-the-art methods.
****
Yuan Gao,WEIZHONG ZHANG,Wenhan Luo,Lin Ma,Jin-Gang Yu,Gui-Song Xia,Jiayi Ma
Aux-NAS: Exploiting Auxiliary Labels with Negligibly Extra Inference Cost
We aim at exploiting additional auxiliary labels from an independent (auxiliary) task to boost the primary task performance which we focus on, while preserving a single task inference cost of the primary task. While most existing auxiliary learning methods are optimization-based relying on loss weights/gradients manipulation, our method is architecture-based with a flexible asymmetric structure for the primary and auxiliary tasks, which produces different networks for training and inference. Specifically, starting from two single task networks/branches (each representing a task), we propose a novel method with evolving networks where only primary-to-auxiliary links exist as the cross-task connections after convergence. These connections can be removed during the primary task inference, resulting in a single task inference cost. We achieve this by formulating a Neural Architecture Search (NAS) problem, where we initialize bi-directional connections in the search space and guide the NAS optimization converging to an architecture with only the single-side primary-to-auxiliary connections. Moreover, our method can be incorporated with existing optimization-based auxiliary learning approaches. Extensive experiments with 6 tasks on NYU v2, CityScapes, and Taskonomy datasets using VGG-16, ResNet-50, and ViTBase backbones validate the promising performance. The codes will be released.

****
Ilmin Kang,HyounYoung Bae,Kangil Kim
Label-Focused Inductive Bias over Latent Object Features in Visual Classificatio
n
Most neural networks for classification primarily learn features differentiated by input-domain related information such as visual similarity of objects in an image. While this focus is natural behavior, it can inadvertently introduce an inductive bias that conflicts with unseen relations in an implicit output-domain determined by human labeling based on their own world knowledge. Such conflicts can limit generalization of models by potential dominance of the input-domain focused bias in inference.
To overcome this limitation without external resources, we introduce Output-Domain focused Biasing (ODB) training strategy that constructs inductive biases on features differentiated by only output labels. It has four steps: 1) it learns intermediate latent object features in an unsupervised manner; 2) it decouples their visual dependencies by assigning new independent embedding parameters; 3) it captures structured features optimized for the original classification task; and 4) it integrates the structured features with the original visual features for the final prediction.
We implement the ODB on a vision transformer architecture, and achieved significant improvements on image classification benchmarks. This paper offers a straightforward and effective method to obtain and utilize output-domain focused inductive bias for classification mapping two different domains.
****
Jonathan Brokman,Roy Betser,Rotem Turjeman,Tom Berkov,Ido Cohen,Guy Gilboa
Enhancing Neural Training via a Correlated Dynamics Model
As neural networks grow in scale, their training becomes both computationally demanding and rich in dynamics. Amidst the flourishing interest in these training dynamics, we present a novel observation: Parameters during training exhibit intrinsic correlations over time. Capitalizing on this, we introduce \emph{correlation mode decomposition} (CMD). This algorithm clusters the parameter space into groups, termed modes, that display synchronized behavior across epochs. This enables CMD to efficiently represent the training dynamics of complex networks, like ResNets and Transformers, using only a few modes. Moreover, test set generalization is enhanced.

We introduce an efficient CMD variant, designed to run concurrently with training. Our experiments indicate that CMD surpasses the state-of-the-art method for compactly modeled dynamics on image classification. Our modeling can improve training efficiency and lower communication overhead, as shown by our preliminary experiments in the context of federated learning.
****
Zhen Xiang,Fengqing Jiang,Zidi Xiong,Bhaskar Ramasubramanian,Radha Poovendran,Bo Li
BadChain: Backdoor Chain-of-Thought Prompting for Large Language Models
Large language models (LLMs) are shown to benefit from chain-of-thought (COT) prompting, particularly when tackling tasks that require systematic reasoning processes. On the other hand, COT prompting also poses new vulnerabilities in the form of backdoor attacks, wherein the model will output unintended malicious content under specific backdoor-triggered conditions during inference. Traditional methods for launching backdoor attacks involve either contaminating the training dataset with backdoored instances or directly manipulating the model parameters during deployment. However, these approaches are not practical for commercial LLMs that typically operate via API access. In this paper, we propose BadChain, the first backdoor attack against LLMs employing COT prompting, which does not require access to the training dataset or model parameters and imposes low computational overhead. BadChain leverages the inherent reasoning capabilities of LLMs by inserting a backdoor reasoning step into the sequence of reasoning steps of the model output, thereby altering the final response when a backdoor trigger is embedded in the query prompt. In particular, a subset of demonstrations will be ma

nipulated to incorporate a backdoor reasoning step in COT prompting. Consequently, given any query prompt containing the backdoor trigger, the LLM will be misled to output unintended content. Empirically, we show the effectiveness of BadChain for two COT strategies across four LLMs (Llama2, GPT-3.5, PaLM2, and GPT-4) and six complex benchmark tasks encompassing arithmetic, commonsense, and symbolic reasoning. We show that the baseline backdoor attacks designed for simpler tasks such as semantic classification will fail on these complicated tasks. In addition, our findings reveal that LLMs endowed with stronger reasoning capabilities exhibit higher susceptibility to BadChain, exemplified by a high average attack success rate of 97.0\% across the six benchmark tasks on GPT-4. We also demonstrate the interpretability of BadChain by showing that the relationship between the trigger and the backdoor reasoning step can be well-explained based on the output of the backdoored model. Finally, we propose two defenses based on shuffling and demonstrate their overall ineffectiveness against BadChain. Therefore, BadChain remains a severe threat to LLMs, underscoring the urgency for the development of robust and effective future defenses.
****
Aojun Zhou,Ke Wang,Zimu Lu,Weikang Shi,Sichun Luo,Zipeng Qin,Shaoqing Lu,Anya Jia,Linqi Song,Mingjie Zhan,Hongsheng Li
Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification
Recent progress in large language models (LLMs) like GPT-4 and PaLM-2 has brought significant advancements in addressing math reasoning problems. In particular, OpenAI's latest version of GPT-4, known as GPT-4 Code Interpreter, shows remarkable performance on challenging math datasets. In this paper, we explore the effect of code on enhancing LLMs' reasoning capability by introducing different constraints on the Code Usage Frequency of GPT-4 Code Interpreter. We found that its success can be largely attributed to its powerful skills in generating and executing code, evaluating the output of code execution, and rectifying its solution when receiving unreasonable outputs. Based on this insight, we propose a novel and effective prompting method, explicit $\underline{\text{c}}$ode-based $\underline{\text{s}}$elf-$\underline{\text{v}}$erification (CSV), to further boost the mathematical reasoning potential of GPT-4 Code Interpreter. This method employs a zero-shot prompt on GPT-4 Code Interpreter to encourage it to use code to self-verify its answers. In instances where the verification state registers as "False", the model shall automatically amend its solution, analogous to our approach of rectifying errors during a mathematics examination. Furthermore, we recognize that the states of the verification result indicate the confidence of a solution, which can improve the effectiveness of majority voting. With GPT-4 Code Interpreter and CSV, we achieve an impressive zero-shot accuracy on MATH dataset.
****
Tianyu Fan,Lirong Wu,Yufei Huang,Haitao Lin,Cheng Tan,Zhangyang Gao,Stan Z. Li
Decoupling Weighing and Selecting for Integrating Multiple Graph Pre-training Tasks
Recent years have witnessed the great success of graph pre-training for graph representation learning. With hundreds of graph pre-training tasks proposed, integrating knowledge acquired from multiple pre-training tasks has become a popular research topic. In this paper, we identify two important collaborative processes for this topic: (1) select: how to select an optimal task combination from a given task pool based on their compatibility, and (2) weigh: how to weigh the selected tasks based on their importance. While there currently has been a lot of work focused on weighing, comparatively little effort has been devoted to selecting. This paper proposes a novel instance-level framework for integrating multiple graph pre-training tasks,
Weigh And Select (WAS), where the two collaborative processes, weighing and selecting, are combined by decoupled siamese networks. Specifically, it first adaptively learns an optimal combination of tasks for each instance from a given task pool, based on which a customized instance-level task weighing strategy is learned. Extensive experiments on 16 graph datasets across node-level and graph-level downstream tasks have demonstrated that by combining a few simple but classical

tasks, WAS can achieve comparable performance to other leading counterparts. The code is available at https://github.com/TianyuFan0504/WAS.
****
Mingxuan Liu,Subhankar Roy,Wenjing Li,Zhun Zhong,Nicu Sebe,Elisa Ricci
Democratizing Fine-grained Visual Recognition with Large Language Models
Identifying subordinate-level categories from images is a longstanding task in computer vision and is referred to as fine-grained visual recognition (FGVR). It has tremendous significance in real-world applications since an average layperson does not excel at differentiating species of birds or mushrooms due to subtle differences among the species. A major bottleneck in developing FGVR systems is caused by the need of high-quality paired expert annotations. To circumvent the need of expert knowledge we propose Fine-grained Semantic Category Reasoning (FineR) that internally leverages the world knowledge of large language models (LLMs) as a proxy in order to reason about fine-grained category names. In detail, to bridge the modality gap between images and LLM, we extract part-level visual attributes from images as text and feed that information to a LLM. Based on the visual attributes and its internal world knowledge the LLM reasons about the subordinate-level category names. Our training-free FineR outperforms several state-of-the-art FGVR and language and vision assistant models and shows promise in working in the wild and in new domains where gathering expert annotation is arduous.
****
Seohong Park,Oleh Rybkin,Sergey Levine
METRA: Scalable Unsupervised RL with Metric-Aware Abstraction
Unsupervised pre-training strategies have proven to be highly effective in natural language processing and computer vision. Likewise, unsupervised reinforcement learning (RL) holds the promise of discovering a variety of potentially useful behaviors that can accelerate the learning of a wide array of downstream tasks. Previous unsupervised RL approaches have mainly focused on pure exploration and mutual information skill learning. However, despite the previous attempts, making unsupervised RL truly scalable still remains a major open challenge: pure exploration approaches might struggle in complex environments with large state spaces, where covering every possible transition is infeasible, and mutual information skill learning approaches might completely fail to explore the environment due to the lack of incentives. To make unsupervised RL scalable to complex, high-dimensional environments, we propose a novel unsupervised RL objective, which we call Metric-Aware Abstraction (METRA). Our main idea is, instead of directly covering the entire state space, to only cover a compact latent space $\mathcal{Z}$ that is metrically connected to the state space $\mathcal{S}$ by temporal distances. By learning to move in every direction in the latent space, METRA obtains a tractable set of diverse behaviors that approximately cover the state space, being scalable to high-dimensional environments. Through our experiments in five locomotion and manipulation environments, we demonstrate that METRA can discover a variety of useful behaviors even in complex, pixel-based environments, being the first unsupervised RL method that discovers diverse locomotion behaviors in pixel-based Quadruped and Humanoid. Our code and videos are available at https://seohong.me/projects/metra/
****
Jingyun Xiao,Ran Liu,Eva L Dyer
GAFormer: Enhancing Timeseries Transformers Through Group-Aware Embeddings
Analyzing multivariate time series is important in many domains. However, it has been difficult to learn robust and generalizable representations within multivariate datasets due to complex inter-channel relationships and dynamic shifts. In this paper, we introduce a novel approach for learning spatiotemporal structure and using it to improve the application of transformers to timeseries datasets. Our framework learns a set of group tokens, and builds an instance-specific group embedding (GE) layer that assigns input tokens to a small number of group tokens to incorporate structure into learning. We then introduce a novel architecture, Group-Aware transFormer (GAFormer), which incorporates both spatial and temporal group embeddings to achieve state-of-the-art performance on a number of ti

me-series classification and regression tasks. In evaluations on a number of div
erse timeseries datasets, we show that GE on its own can provide a nice enhancem
ent to a number of backbones, and that by coupling spatial and temporal group em
beddings, the GAFormer can outperform the existing baselines. Finally, we show h
ow our approach discerns latent structures in data even without information abou
t the spatial ordering of channels, and yields a more interpretable decompositio
n of spatial and temporal structure underlying complex multivariate datasets.
****
Shuai Yang,Yukang Chen,Luozhou WANG,Shu Liu,Ying-Cong Chen
Denoising Diffusion Step-aware Models
Denoising Diffusion Probabilistic Models (DDPMs) have garnered popularity for da
ta generation across various domains. However, a significant bottleneck is the n
ecessity for whole-network computation during every step of the generative proce
ss, leading to high computational overheads. This paper presents a novel framewo
rk, Denoising Diffusion Step-aware Models (DDSM), to address this challenge. Unl
ike conventional approaches, DDSM employs a spectrum of neural networks whose si
zes are adapted according to the importance of each generative step, as determin
ed through evolutionary search. This step-wise network variation effectively cir
cumvents redundant computational efforts, particularly in less critical steps, t
hereby enhancing the efficiency of the diffusion model. Furthermore, the step-aw
are design can be seamlessly integrated with other efficiency-geared diffusion m
odels such as DDIMs and latent diffusion, thus broadening the scope of computati
onal savings. Empirical evaluations demonstrate that DDSM achieves computational
 savings of 49% for CIFAR-10, 61% for CelebA-HQ, 59% for LSUN-bedroom, 71% for A
FHQ, and 76% for ImageNet, all without compromising the generation quality. Our
code and models are available at https://github.com/EnVision-Research/DDSM.
****
Weiyun Wang,Min Shi,Qingyun Li,Wenhai Wang,Zhenhang Huang,Linjie Xing,Zhe Chen,H
ao Li,Xizhou Zhu,Zhiguo Cao,Yushi Chen,Tong Lu,Jifeng Dai,Yu Qiao
The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of
 the Open World
We present the All-Seeing (AS) project: a large-scale dataset and model for reco
gnizing and understanding everything in the open world.
Using a scalable data engine that incorporates human feedback and efficient mode
ls in the loop, we create a new dataset (AS-1B) with over 1.2 billion regions an
notated with semantic tags, question-answering pairs, and detailed captions. It
covers a wide range of 3.5 million common and rare concepts in the real world an
d has 132.2 billion tokens that describe the concepts and their attributes. Leve
raging this new dataset, we develop the All-Seeing model (ASM), a unified framew
ork for panoptic visual recognition and understanding. The model is trained with
 open-ended language prompts and locations, which allows it to generalize to var
ious vision and language tasks with remarkable zero-shot performance, including
both region- and image-level retrieval, region recognition, captioning, and ques
tion-answering. We hope that this project can serve as a foundation for vision-l
anguage artificial general intelligence research. Code is available at https://g
ithub.com/OpenGVLab/all-seeing.
****
Kevin Black,Mitsuhiko Nakamoto,Pranav Atreya,Homer Rich Walke,Chelsea Finn,Avira
l Kumar,Sergey Levine
Zero-Shot Robotic Manipulation with Pre-Trained Image-Editing Diffusion Models
If generalist robots are to operate in truly unstructured environments, they nee
d to be able to recognize and reason about novel objects and scenarios. Such obj
ects and scenarios might not be present in the robot's own training data. We pro
pose SuSIE, a method that leverages an image editing diffusion model to act as a
 high-level planner by proposing intermediate subgoals that a low-level controll
er attains. Specifically, we fine-tune InstructPix2Pix on robot data such that i
t outputs a hypothetical future observation given the robot's current observatio
n and a language command. We then use the same robot data to train a low-level g
oal-conditioned policy to reach a given image observation. We find that when the
se components are combined, the resulting system exhibits robust generalization

capabilities. The high-level planner utilizes its Internet-scale pre-training and visual understanding to guide the low-level goal-conditioned policy, achieving significantly better generalization than conventional language-conditioned policies. We demonstrate that this approach solves real robot control tasks involving novel objects, distractors, and even environments, both in the real world and in simulation. The project website can be found at http://subgoal-image-editing.github.io.
****
Finn Rietz,Erik Schaffernicht,Stefan Heinrich,Johannes A. Stork
Prioritized Soft Q-Decomposition for Lexicographic Reinforcement Learning
Reinforcement learning (RL) for complex tasks remains a challenge, primarily due to the difficulties of engineering scalar reward functions and the inherent inefficiency of training models from scratch. Instead, it would be better to specify complex tasks in terms of elementary subtasks and to reuse subtask solutions whenever possible. In this work, we address continuous space lexicographic multi-objective RL problems, consisting of prioritized subtasks, which are notoriously difficult to solve. We show that these can be scalarized with a subtask transformation and then solved incrementally using value decomposition. Exploiting this insight, we propose prioritized soft Q-decomposition (PSQD), a novel algorithm for learning and adapting subtask solutions under lexicographic priorities in continuous state-action spaces. PSQD offers the ability to reuse previously learned subtask solutions in a zero-shot composition, followed by an adaptation step. Its ability to use retained subtask training data for offline learning eliminates the need for new environment interaction during adaptation. We demonstrate the efficacy of our approach by presenting successful learning, reuse, and adaptation results for both low- and high-dimensional simulated robot control tasks, as well as offline learning results. In contrast to baseline approaches, PSQD does not trade off between conflicting subtasks or priority constraints and satisfies subtask priorities during learning. PSQD provides an intuitive framework for tackling complex RL problems, offering insights into the inner workings of the subtask composition.
****
Zibin Dong,Yifu Yuan,Jianye HAO,Fei Ni,Yao Mu,YAN ZHENG,Yujing Hu,Tangjie Lv,Changjie Fan,Zhipeng Hu
AlignDiff: Aligning Diverse Human Preferences via Behavior-Customisable Diffusion Model
Aligning agent behaviors with diverse human preferences remains a challenging problem in reinforcement learning (RL), owing to the inherent abstractness and mutability of human preferences. To address these issues, we propose AlignDiff, a novel framework that leverages RLHF to quantify human preferences, covering abstractness, and utilizes them to guide diffusion planning for zero-shot behavior customizing, covering mutability. AlignDiff can accurately match user-customized behaviors and efficiently switch from one to another. To build the framework, we first establish the multi-perspective human feedback datasets, which contain comparisons for the attributes of diverse behaviors, and then train an attribute strength model to predict quantified relative strengths. After relabeling behavioral datasets with relative strengths, we proceed to train an attribute-conditioned diffusion model, which serves as a planner with the attribute strength model as a director for preference aligning at the inference phase. We evaluate AlignDiff on various locomotion tasks and demonstrate its superior performance on preference matching, switching, and covering compared to other baselines. Its capability of completing unseen downstream tasks under human instructions also showcases the promising potential for human-AI collaboration. More visualization videos are released on https://aligndiff.github.io/.
****
Pengfei He,Han Xu,Jie Ren,Yingqian Cui,Shenglai Zeng,Hui Liu,Charu C. Aggarwal,Jiliang Tang
Sharpness-Aware Data Poisoning Attack
Recent research has highlighted the vulnerability of Deep Neural Networks (DNNs) against data poisoning attacks. These attacks aim to inject poisoning samples i

nto the models' training dataset such that the trained models have inference failures. While previous studies have executed different types of attacks, one major challenge that greatly limits their effectiveness is the
uncertainty of the re-training process after the injection of poisoning samples. It includes the uncertainty of training initialization, algorithm and model architecture. To address this challenge, we propose a new strategy called **Sharpness-Aware Data Poisoning Attack (SAPA)**. In particular, it leverages the concept of DNNs' loss landscape sharpness to optimize the poisoning effect on the (approximately) worst re-trained model. Extensive experiments demonstrate that SAPA offers a general and principled strategy that significantly enhances various types of poisoning attacks against various types of re-training uncertainty.
****
Haoyu Han,Xiaorui Liu,Li Ma,MohamadAli Torkamani,Hui Liu,Jiliang Tang,Makoto Yamada
Structural Fairness-aware Active Learning for Graph Neural Networks
Graph Neural Networks (GNNs) have seen significant achievements in semi-supervised node classification. Yet, their efficacy often hinges on access to high-quality labeled node samples, which may not always be available in real-world scenarios. While active learning is commonly employed across various domains to pinpoint and label high-quality samples based on data features, graph data present unique challenges due to their intrinsic structures that render nodes non-i.i.d. Furthermore, biases emerge from the positioning of labeled nodes; for instance, nodes closer to the labeled counterparts often yield better performance. To better leverage graph structure and mitigate structural bias in active learning, we present a unified optimization framework (SCARCE), which is also easily incorporated with node features. Extensive experiments demonstrate that the proposed method not only improves the GNNs performance but also paves the way for more fair results.
****
Qihang Zhou,Guansong Pang,Yu Tian,Shibo He,Jiming Chen
AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection
Zero-shot anomaly detection (ZSAD) requires detection models trained using auxiliary
data to detect anomalies without any training sample in a target dataset. It
is a crucial task when training data is not accessible due to various concerns, e.g.,
data privacy, yet it is challenging since the models need to generalize to anomalies
across different domains where the appearance of foreground objects, abnormal
regions, and background features, such as defects/tumors on different products/
organs, can vary significantly. Recently large pre-trained vision-language
models (VLMs), such as CLIP, have demonstrated strong zero-shot recognition
ability in various vision tasks, including anomaly detection. However, their ZSAD
performance is weak since the VLMs focus more on modeling the class semantics
of the foreground objects rather than the abnormality/normality in the images. In
this paper we introduce a novel approach, namely AnomalyCLIP, to adapt CLIP
for accurate ZSAD across different domains. The key insight of AnomalyCLIP
is to learn object-agnostic text prompts that capture generic normality and abnormality
in an image regardless of its foreground objects. This allows our model to
focus on the abnormal image regions rather than the object semantics, enabling
generalized normality and abnormality recognition on diverse types of objects.
Large-scale experiments on 17 real-world anomaly detection datasets show that
AnomalyCLIP achieves superior zero-shot performance of detecting and segmenting
anomalies in datasets of highly diverse class semantics from various defect
inspection and medical imaging domains. Code will be made available at https://github.com/zqhang/AnomalyCLIP.
****

Wenhao Li
Efficient Planning with Latent Diffusion
Temporal abstraction and efficient planning pose significant challenges in offli ne reinforcement learning, mainly when dealing with domains that involve tempora lly extended tasks and delayed sparse rewards. Existing methods typically plan i n the raw action space and can be inefficient and inflexible. Latent action spac es offer a more flexible approach, capturing only possible actions within the be havior policy support and decoupling the temporal structure between planning and modeling. However, current latent-action-based methods are limited to discrete spaces and require expensive planning steps. This paper presents a unified frame work for continuous latent action space representation learning and planning by leveraging latent, score-based diffusion models. We establish the theoretical eq uivalence between planning in the latent action space and energy-guided sampling with a pretrained diffusion model and introduce a novel sequence-level exact sa mpling method. Our proposed method, $\texttt{LatentDiffuser}$, demonstrates comp etitive performance on low-dimensional locomotion control tasks and surpasses ex isting methods in higher-dimensional tasks.
****
Kai-Po Chang,Chi-Pin Huang,Wei-Yuan Cheng,Fu-En Yang,Chien-Yi Wang,Yung-Hsuan La i,Yu-Chiang Frank Wang
RAPPER: Reinforced Rationale-Prompted Paradigm for Natural Language Explanation in Visual Question Answering
Natural Language Explanation (NLE) in vision and language tasks aims to provide human-understandable explanations for the associated decision-making process. In practice, one might encounter explanations which lack informativeness or contra dict visual-grounded facts, known as implausibility and hallucination problems, respectively. To tackle these challenging issues, we consider the task of visual question answering (VQA) and introduce Rapper, a two-stage Reinforced Rationale -Prompted Paradigm. By knowledge distillation, the former stage of Rapper infuse s rationale-prompting via large language models (LLMs), encouraging the rational es supported by language-based facts. As for the latter stage, a unique Reinforc ement Learning from NLE Feedback (RLNF) is introduced for injecting visual facts into NLE generation. Finally, quantitative and qualitative experiments on two V L-NLE benchmarks show that Rapper surpasses state-of-the-art VQA-NLE methods whi le providing plausible and faithful NLE.
****
Brian Hu Zhang,Gabriele Farina,Tuomas Sandholm
Mediator Interpretation and Faster Learning Algorithms for Linear Correlated Equ ilibria in General Sequential Games
A recent paper by Farina and Pipis (2023) established the existence of uncoupled no-linear-swap regret dynamics with polynomial-time iterations in extensive-for m games. The equilibrium points reached by these dynamics, known as linear corre lated equilibria, are currently the tightest known relaxation of correlated equi librium that can be learned in polynomial time in any finite extensive-form game . However, their properties remain vastly unexplored, and their computation is o nerous. In this paper, we provide several contributions shedding light on the fu ndamental nature of linear-swap regret. First, we show a connection between line ar deviations and a generalization of communication deviations in which the play er can make queries to a ``mediator'' who replies with action recommendations, a nd, critically, the player is not constrained to match the timing of the game as would be the case for communication deviations. We coin this latter set the unt imed communication (UTC) deviations. We show that the UTC deviations coincide pr ecisely with the linear deviations, and therefore that any player minimizing UTC regret also minimizes linear-swap regret. We then leverage this connection to d evelop state-of-the-art no-regret algorithms for computing linear correlated equ ilibria, both in theory and in practice. In theory, our algorithms achieve polyn omially better per-iteration runtimes; in practice, our algorithms represent the state of the art by several orders of magnitude.
****
Patrick Schnell,Nils Thuerey

Stabilizing Backpropagation Through Time to Learn Complex Physics
Of all the vector fields surrounding the minima of recurrent learning setups, th
e gradient field with its exploding and vanishing updates appears a poor choice
for optimization, offering little beyond efficient computability. We seek to imp
rove this suboptimal practice in the context of physics simulations, where backp
ropagating feedback through many unrolled time steps is considered crucial to ac
quiring temporally coherent behavior. The alternative vector field we propose fo
llows from two principles: physics simulators, unlike neural networks, have a ba
lanced gradient flow and certain modifications to the backpropagation pass leave
 the positions of the original minima unchanged. As any modification of backprop
agation decouples forward and backward pass, the rotation-free character of the
gradient field is lost. Therefore, we discuss the negative implications of using
 such a rotational vector field for optimization and how to counteract them. Our
 final procedure is easily implementable via a sequence of gradient stopping and
 component-wise comparison operations, which do not negatively affect scalabilit
y. Our experiments on three control problems show that especially as we increase
 the complexity of each task, the unbalanced updates from the gradient can no lo
nger provide the precise control signals necessary while our method still solves
 the tasks. Our code can be found at https://github.com/tum-pbs/StableBPTT.
****
Feiyang YE,Yueming Lyu,Xuehao Wang,Yu Zhang,Ivor Tsang
Adaptive Stochastic Gradient Algorithm for Black-box Multi-Objective Learning
Multi-objective optimization (MOO) has become an influential framework for vario
us machine learning problems, including reinforcement learning and multi-task le
arning. In this paper, we study the black-box multi-objective optimization probl
em, where we aim to optimize multiple potentially conflicting objectives with fu
nction queries only. To address this challenging problem and find a Pareto optim
al solution or the Pareto stationary solution,
we propose a novel adaptive stochastic gradient algorithm for black-box MOO, cal
led ASMG.
Specifically, we use the stochastic gradient approximation method to obtain the
gradient for the distribution parameters of the Gaussian smoothed MOO with funct
ion queries only. Subsequently, an adaptive weight is employed to aggregate all
stochastic gradients to optimize all objective functions effectively.
Theoretically, we explicitly provide the connection between the original MOO pro
blem and the corresponding Gaussian smoothed MOO problem and prove the convergen
ce rate for the proposed ASMG algorithm in both convex and non-convex scenarios.
Empirically, the proposed ASMG method achieves competitive performance on multip
le numerical benchmark problems. Additionally, the state-of-the-art performance
on the black-box multi-task learning problem demonstrates the effectiveness of t
he proposed ASMG method.
****
DIPANJYOTI PAUL,Arpita Chowdhury,Xinqi Xiong,Feng-Ju Chang,David Edward Carlyn,S
amuel Stevens,Kaiya L Provost,Anuj Karpatne,Bryan Carstens,Daniel Rubenstein,Cha
rles Stewart,Tanya Berger-Wolf,Yu Su,Wei-Lun Chao
A Simple Interpretable Transformer for Fine-Grained Image Classification and Ana
lysis
We present a novel usage of Transformers to make image classification interpreta
ble. Unlike mainstream classifiers that wait until the last fully-connected laye
r to incorporate class information to make predictions, we investigate a proacti
ve approach, asking each class to search for itself in an image. We realize this
 idea via a Transformer encoder-decoder inspired by DEtection TRansformer (DETR)
. We learn "class-specific'' queries (one for each class) as input to the decode
r, enabling each class to localize its patterns in an image via cross-attention.
 We name our approach INterpretable TRansformer (INTR), which is fairly easy to
implement and exhibits several compelling properties. We show that INTR intrinsi
cally encourages each class to attend distinctively; the cross-attention weights
 thus provide a faithful interpretation of the prediction. Interestingly, via "m
ulti-head'' cross-attention, INTR could identify different "attributes'' of a cl
ass, making it particularly suitable for fine-grained classification and analysi

s, which we demonstrate on eight datasets.
****
Mara Finkelstein,Markus Freitag
MBR and QE Finetuning: Training-time Distillation of the Best and Most Expensive Decoding Methods
Recent research in decoding methods for Natural Language Generation (NLG) tasks has shown that MAP decoding is not optimal, because model probabilities do not always align with human preferences. Stronger decoding methods, including Quality Estimation (QE) reranking and Minimum Bayes' Risk (MBR) decoding, have since been proposed to mitigate the model-perplexity-vs-quality mismatch. While these decoding methods achieve state-of-the-art performance, they are prohibitively expensive to compute. In this work, we propose MBR finetuning and QE finetuning, which distill the quality gains from these decoding methods at training time, while using an efficient decoding algorithm at inference time. Using the canonical NLG task of Neural Machine Translation (NMT), we show that even with self-training, these finetuning methods significantly outperform the base model. Moreover, when using an external LLM as a teacher model, these finetuning methods outperform finetuning on human-generated references. These findings suggest new ways to leverage monolingual data to achieve improvements in model quality that are on par with, or even exceed, improvements from human-curated data, while maintaining maximum efficiency during decoding.
****
Yixin Cheng,Grigorios Chrysos,Markos Georgopoulos,Volkan Cevher
Multilinear Operator Networks
Despite the remarkable capabilities of deep neural networks in image recognition, the dependence on activation functions remains a largely unexplored area and has yet to be eliminated. On the other hand, Polynomial Networks is a class of models that does not require activation functions, but have yet to perform on par with modern architectures. In this work, we aim close this gap and propose MONet, which relies *solely* on multilinear operators. The core layer of MONet, called Mu-Layer, captures multiplicative interactions of the elements of the input token. MONet captures high-degree interactions of the input elements and we demonstrate the efficacy of our approach on a series of image recognition and scientific computing benchmarks. The proposed model outperforms prior polynomial networks and performs on par with modern architectures. We believe that MONet can inspire further research on models that use entirely multilinear operations.
****
Duy Kien Nguyen,Yanghao Li,Vaibhav Aggarwal,Martin R. Oswald,Alexander Kirillov,Cees G. M. Snoek,Xinlei Chen
R-MAE: Regions Meet Masked Autoencoders
In this work, we explore regions as a potential visual analogue of words for self-supervised image representation learning. Inspired by Masked Autoencoding (MAE), a generative pre-training baseline, we propose masked region autoencoding to learn from groups of pixels or regions. Specifically, we design an architecture which efficiently addresses the one-to-many mapping between images and regions, while being highly effective especially with high-quality regions. When integrated with MAE, our approach (R-MAE) demonstrates consistent improvements across various pre-training datasets and downstream detection and segmentation benchmarks, with negligible computational overheads. Beyond the quantitative evaluation, our analysis indicates the models pre-trained with masked region autoencoding unlock the potential for interactive segmentation. The code is provided at https://github.com/facebookresearch/r-mae.
****
Zhongqi Yue,Jiankun Wang,Qianru Sun,Lei Ji,Eric I-Chao Chang,Hanwang Zhang
Exploring Diffusion Time-steps for Unsupervised Representation Learning
Representation learning is all about discovering the hidden modular attributes that generate the data faithfully. We explore the potential of Denoising Diffusion Probabilistic Model (DM) in unsupervised learning of the modular attributes. We build a theoretical framework that connects the diffusion time-steps and the hidden attributes, which serves as an effective inductive bias for unsupervised l

earning. Specifically, the forward diffusion process incrementally adds Gaussian noise to samples at each time-step, which essentially collapses different samples into similar ones by losing attributes, e.g., fine-grained attributes such as texture are lost with less noise added (i.e., early time-steps), while coarse-grained ones such as shape are lost by adding more noise (i.e., late time-steps). To disentangle the modular attributes, at each time-step t, we learn a t-specific feature to compensate for the newly lost attribute, and the set of all {1,...,t}-specific features, corresponding to the cumulative set of lost attributes, are trained to make up for the reconstruction error of a pre-trained DM at time-step t. On CelebA, FFHQ, and Bedroom datasets, the learned feature significantly improves attribute classification and enables faithful counterfactual generation, e.g., interpolating only one specified attribute between two images, validating the disentanglement quality. Codes are in https://github.com/yue-zhongqi/diti.
****
Zhijian Xu,Ailing Zeng,Qiang Xu
FITS: Modeling Time Series with $10k$ Parameters
In this paper, we introduce FITS, a lightweight yet powerful model for time series analysis. Unlike existing models that directly process raw time-domain data, FITS operates on the principle that time series can be manipulated through interpolation in the complex frequency domain, achieving performance comparable to state-of-the-art models for time series forecasting and anomaly detection tasks. Notably, FITS accomplishes this with a svelte profile of just about $10k$ parameters, making it ideally suited for edge devices and paving the way for a wide range of applications. The code is available for review at: \url{https://anonymous.4open.science/r/FITS}.
****
Daiki Chijiwa
Transferring Learning Trajectories of Neural Networks
Training deep neural networks (DNNs) is computationally expensive, which is problematic especially when performing duplicated or similar training runs in model ensemble or fine-tuning pre-trained models, for example. Once we have trained one DNN on some dataset, we have its learning trajectory (i.e., a sequence of intermediate parameters during training) which may potentially contain useful information for learning the dataset. However, there has been no attempt to utilize such information of a given learning trajectory for another training. In this paper, we formulate the problem of "transferring" a given learning trajectory from one initial parameter to another one (named *learning transfer problem*) and derive the first algorithm to approximately solve it by matching gradients successively along the trajectory via permutation symmetry. We empirically show that the transferred parameters achieve non-trivial accuracy before any direct training, and can be trained significantly faster than training from scratch.
****
QIUHAO Zeng,Changjian Shui,Long-Kai Huang,Peng Liu,Xi Chen,Charles Ling,Boyu Wang
Latent Trajectory Learning for Limited Timestamps under Distribution Shift over Time
Distribution shifts over time are common in real-world machine-learning applications. This scenario is formulated as Evolving Domain Generalization (EDG), where models aim to generalize well to unseen target domains in a time-varying system by learning and leveraging the underlying evolving pattern of the distribution shifts across domains. However, existing methods encounter challenges due to the limited number of timestamps (every domain corresponds to a timestamp) in EDG datasets, leading to difficulties in capturing evolving dynamics and risking overfitting to the sparse timestamps, which hampers their generalization and adaptability to new tasks. To address this limitation, we propose a novel approach SDE-EDG that collects the Infinitely Fined-Grid Evolving Trajectory (IFGET) of the data distribution with continuous-interpolated samples to bridge temporal gaps (intervals between two successive timestamps). Furthermore, by leveraging the inherent capacity of Stochastic Differential Equations (SDEs) to capture continuous trajectories, we propose their use to align SDE-modeled trajectories with IFGET

across domains, thus enabling the capture of evolving distribution trends. We ev
aluate our approach on several benchmark datasets and demonstrate that it can ac
hieve superior performance compared to existing state-of-the-art methods.
****

Changmao Li,Jeffrey Flanigan
Future Language Modeling from Temporal Document History
Predicting the future is of great interest across many aspects of human activity
. Businesses are interested in future trends, traders are interested in future
stock prices, and companies are highly interested in future technological breakt
hroughs. While there are many automated systems for predicting future numerical
 data, such as weather, stock prices, and demand for products, there is relative
ly little work in automatically predicting textual data. Humans are interested
in textual data predictions because it is a natural format for our consumption,
and experts routinely make predictions in a textual format (Christensen et al.,
2004; Tetlock & Gardner, 2015; Frick, 2015). However, there has been relatively
little formalization of this general problem in the machine learning or natural
language processing communities. To address this gap, we introduce the task of
future language modeling: probabilistic modeling of texts in the future based on
 a temporal history of texts. To our knowledge, our work is the first work to f
ormalize the task of predicting the future in this way. We show that it is inde
ed possible to build future language models that improve upon strong non-tempora
l language model baselines, opening the door to working on this important, and w
idely applicable problem.
****

Chengzhi Mao,Carl Vondrick,Hao Wang,Junfeng Yang
Raidar: geneRative AI Detection viA Rewriting
We find that large language models (LLMs) are more likely to modify human-writte
n text than AI-generated text when tasked with rewriting. This tendency arises b
ecause LLMs often perceive AI-generated text as high-quality, leading to fewer m
odifications. We introduce a method to detect AI-generated content by prompting
LLMs to rewrite text and calculating the editing distance of the output. We dubb
ed our geneRative AI Detection viA Rewriting method Raidar. Raidar significantl
y improves the F1 detection scores of existing AI content detection models -- bo
th academic and commercial -- across various domains, including News, creative w
riting, student essays, code, Yelp reviews, and arXiv papers, with gains of up t
o 29 points. Operating solely on word symbols without high-dimensional features,
 our method is compatible with black box LLMs, and is inherently robust on new c
ontent. Our results illustrate the unique imprint of machine-generated text thro
ugh the lens of the machines themselves.
****

Linlu Qiu,Liwei Jiang,Ximing Lu,Melanie Sclar,Valentina Pyatkin,Chandra Bhagavat
ula,Bailin Wang,Yoon Kim,Yejin Choi,Nouha Dziri,Xiang Ren
Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Mo
dels with Hypothesis Refinement
The ability to derive underlying principles from a handful of observations and t
hen generalize to novel situations---known as inductive reasoning---is central t
o human intelligence. Prior work suggests that language models (LMs) often fall
short on inductive reasoning, despite achieving impressive success on research b
enchmarks. In this work, we conduct a systematic study of the inductive reasonin
g capabilities of LMs through $\textit{iterative hypothesis refinement}$, a tech
nique that more closely mirrors the human inductive process than standard input-
output prompting. Iterative hypothesis refinement employs a three-step process:
proposing, selecting, and refining hypotheses in the form of textual rules. By e
xamining the intermediate rules, we observe that LMs are phenomenal $\textit{hyp
othesis proposers}$ (i.e., generating candidate rules), and when coupled with a
(task-specific) symbolic interpreter that is able to systematically filter the p
roposed set of rules, this hybrid approach achieves strong results across induct
ive reasoning benchmarks that require inducing causal relations, language-like i
nstructions, and symbolic concepts. However, they also behave as puzzling $\text
it{inductive reasoners}$, showing notable performance gaps between rule inductio

n (i.e., identifying plausible rules) and rule application (i.e., applying propo sed rules to instances), suggesting that LMs are proposing hypotheses without be ing able to actually apply the rules. Through empirical and human analyses, we f urther reveal several discrepancies between the inductive reasoning processes of LMs and humans, shedding light on both the potentials and limitations of using LMs in inductive reasoning tasks.
****
Mengmeng Xu,Mattia Soldan,Jialin Gao,Shuming Liu,Juan-Manuel Perez-Rua,Bernard G hanem
Boundary Denoising for Video Activity Localization
Video activity localization aims at understanding the semantic content in long, untrimmed videos and retrieving actions of interest. The retrieved action with i ts start and end locations can be used for highlight generation, temporal action detection, etc. Unfortunately, learning the exact boundary location of activiti es is highly challenging because temporal activities are continuous in time, and there are often no clear-cut transitions between actions. Moreover, the definit ion of the start and end of events is subjective, which may confuse the model. T o alleviate the boundary ambiguity, we propose to study the video activity local ization problem from a denoising perspective. Specifically, we propose an encode r-decoder model named DenosieLoc. During training, a set of temporal spans is ra ndomly generated from the ground truth with a controlled noise scale. Then, we a ttempt to reverse this process by boundary denoising, allowing the localizer to predict activities with precise boundaries and resulting in faster convergence s peed. Experiments show that DenosieLoc advances
several video activity understanding tasks. For example, we observe a gain of +1 2.36% average mAP on the QV-Highlights dataset.
Moreover, DenosieLoc achieves state-of-the-art performance on the MAD dataset bu t with much fewer predictions than others.
****
Kai Cheng,Xiaoxiao Long,Wei Yin,Jin Wang,Zhiqiang Wu,Yuexin Ma,Kaixuan Wang,Xiao zhi Chen,Xuejin Chen
UC-NERF: Neural Radiance Field for Under-Calibrated Multi-View Cameras in Autono mous Driving
Multi-camera setups find widespread use across various applications, such as aut onomous driving, as they greatly expand sensing capabilities.
Despite the fast development of Neural radiance field (NeRF) techniques and thei r wide applications in both indoor and outdoor scenes, applying NeRF to multi-ca mera systems remains very challenging. This is primarily due to the inherent und er-calibration issues in multi-camera setup, including inconsistent imaging effe cts stemming from separately calibrated image signal processing units in diverse cameras, and system errors arising from mechanical vibrations during driving th at affect relative camera poses.
In this paper, we present UC-NeRF, a novel method tailored for novel view synthe sis in under-calibrated multi-view camera systems.
Firstly, we propose a layer-based color correction to rectify the color inconsis tency in different image regions. Second, we propose virtual warping to generate more viewpoint-diverse but color-consistent virtual views for color correction and 3D recovery. Finally, a spatiotemporally constrained pose refinement is desi gned for more robust and accurate pose calibration in multi-camera systems.
Our method not only achieves state-of-the-art performance of novel view synthesi s in multi-camera setups, but also effectively facilitates depth estimation in l arge-scale outdoor scenes with the synthesized novel views.
****
Cheng Han,Qifan Wang,Yiming Cui,Wenguan Wang,Lifu Huang,Siyuan Qi,Dongfang Liu
Facing the Elephant in the Room: Visual Prompt Tuning or Full finetuning?
As the scale of vision models continues to grow, the emergence of Visual Prompt Tuning (VPT) as a parameter-efficient transfer learning technique has gained att ention due to its superior performance compared to traditional full-finetuning.
However, the conditions favoring VPT (the "when") and the underlying rationale (the "why") remain unclear. In this paper, we conduct a comprehensive analysis ac

ross 19 distinct datasets and tasks. To understand the "when" aspect, we identify the scenarios where VPT proves favorable by two dimensions: task objectives and data distributions. We find that VPT is preferrable when there is 1) a substantial disparity between the original and the downstream task objectives ($e.g.$, transitioning from classification to counting), or 2) a notable similarity in data distributions between the two tasks ($e.g.$, both involve natural images). In exploring the "why" dimension, our results indicate VPT's success cannot be attributed solely to overfitting and optimization considerations. The unique way VPT preserves original features and adds parameters appears to be a pivotal factor. Our study provides insights into VPT's mechanisms, and offers guidance for its optimal utilization.
****
Shicheng Liu,Minghui Zhu
Meta Inverse Constrained Reinforcement Learning: Convergence Guarantee and Generalization Analysis
This paper considers the problem of learning the reward function and constraints of an expert from few demonstrations. This problem can be considered as a meta-learning problem where we first learn meta-priors over reward functions and constraints from other distinct but related tasks and then adapt the learned meta-priors to new tasks from only few expert demonstrations. We formulate a bi-level optimization problem where the upper level aims to learn a meta-prior over reward functions and the lower level is to learn a meta-prior over constraints. We propose a novel algorithm to solve this problem and formally guarantee that the algorithm reaches the set of ϵ-stationary points at the iteration complexity $O(\frac{1}{\epsilon^2})$. We also quantify the generalization error to an arbitrary new task. Experiments are used to validate that the learned meta-priors can adapt to new tasks with good performance from only few demonstrations.
****
Marco Federici,Patrick Forré,Ryota Tomioka,Bastiaan S. Veeling
Latent Representation and Simulation of Markov Processes via Time-Lagged Information Bottleneck
Markov processes are widely used mathematical models for describing dynamic systems in various fields. However, accurately simulating large-scale systems at long time scales is computationally expensive due to the short time steps required for accurate integration. In this paper, we introduce an inference process that maps complex systems into a simplified representational space and models large jumps in time. To achieve this, we propose Time-lagged Information Bottleneck (T-IB), a principled objective rooted in information theory, which aims to capture relevant temporal features while discarding high-frequency information to simplify the simulation task and minimize the inference error. Our experiments demonstrate that T-IB learns information-optimal representations for accurately modeling the statistical properties and dynamics of the original process at a selected time lag, outperforming existing time-lagged dimensionality reduction methods.
****
Jianhao Yan,Jin Xu,Chiyu Song,Chenming Wu,Yafu Li,Yue Zhang
Understanding In-Context Learning from Repetitions
This paper explores the elusive mechanism underpinning in-context learning in Large Language Models (LLMs). Our work provides a novel perspective by examining in-context learning via the lens of surface repetitions. We quantitatively investigate the role of surface features in text generation, and empirically establish the existence of token co-occurrence reinforcement, a principle that strengthens the relationship between two tokens based on their contextual co-occurrences. By investigating the dual impacts of these features, our research illuminates the internal workings of in-context learning and expounds on the reasons for its failures. This paper provides an essential contribution to the understanding of in-context learning and its potential limitations, providing a fresh perspective on this exciting capability.
****
Sihan Chen,Xingjian He,Handong Li,Xiaojie Jin,Jiashi Feng,Jing Liu
COSA: Concatenated Sample Pretrained Vision-Language Foundation Model

Due to the limited scale and quality of video-text training corpus, most vision-language foundation models employ image-text datasets for pretraining and primarily focus on modeling visually semantic representations while disregarding temporal semantic representations and correlations. To address this issue, we propose COSA, a COncatenated SAmple pretrained vision-language foundation model. COSA can jointly model visual contents and event-level temporal cues using only image-text corpora. We achieve this by sequentially concatenating multiple image-text pairs as inputs for pretraining. This transformation effectively converts existing image-text corpora into a pseudo video-paragraph corpus, enabling richer scene transformations and explicit event-description correspondence. Extensive experiments demonstrate that COSA consistently improves performance across a broad range of semantic vision-language downstream tasks, including paragraph-to-video retrieval, text-to-video/image retrieval, video/image captioning and video QA. Notably, COSA achieves state-of-the-art results on various competitive benchmarks. Code and model are released at https://github.com/TXH-mercury/COSA.
****

Nishant Jain,Karthikeyan Shanmugam,Pradeep Shenoy
Learning model uncertainty as variance-minimizing instance weights
Predictive uncertainty--a model’s self-awareness regarding its accuracy on an input--is key for both building robust models via training interventions and for test-time applications such as selective classification. We propose a novel instance-conditional reweighting approach that captures predictive uncertainty using an auxiliary network, and unifies these train- and test-time applications. The auxiliary network is trained using a meta-objective in a bilevel optimization framework. A key contribution of our proposal is the meta-objective of minimizing dropout variance, an approximation of Bayesian predictive uncertainty, We show in controlled experiments that we effectively capture diverse specific notions of uncertainty through this meta-objective, while previous approaches only capture certain aspects. These results translate to significant gains in real-world settings–selective classification, label noise, domain adaptation, calibration–and across datasets–Imagenet, Cifar100, diabetic retinopathy, Camelyon, WILDs, Imagenet-C,-A,-R, Clothing-1.6M, etc. For Diabetic Retinopathy, we see upto 3.4\%/3.3\% accuracy & AUC gains over SOTA in selective classification. We also improve upon large-scale pretrained models such as PLEX.
****

Linyi Yang,Shuibai Zhang,Zhuohao Yu,Guangsheng Bao,Yidong Wang,Jindong Wang,Ruochen Xu,Wei Ye,Xing Xie,Weizhu Chen,Yue Zhang
Supervised Knowledge Makes Large Language Models Better In-context Learners
Large Language Models (LLMs) exhibit emerging in-context learning abilities through prompt engineering. The recent progress in large-scale generative models has further expanded their use in real-world language applications. However, the critical challenge of improving the generalizability and factuality of LLMs in natural language understanding and question answering remains under-explored. While previous in-context learning research has focused on enhancing models to adhere to users' specific instructions and quality expectations, and to avoid undesired outputs, little to no work has explored the use of task-specific fine-tuned Language Models (SLMs) to improve LLMs' in-context learning during the inference stage. Our primary contribution is the establishment of a simple yet effective framework that enhances the reliability of LLMs as it: 1) generalizes out-of-distribution data, 2) elucidates how LLMs benefit from discriminative models, and 3) minimizes hallucinations in generative tasks. Using our proposed plug-in method, enhanced versions of Llama 2 and ChatGPT surpass their original versions regarding generalizability and factuality. We offer a comprehensive suite of resources, including 16 curated datasets, prompts, model checkpoints, and LLM outputs across 9 distinct tasks. Our empirical analysis sheds light on the advantages of incorporating discriminative models into LLMs and highlights the potential of our methodology in fostering more reliable LLMs.
****

Zhipeng Zhou,Liu Liu,Peilin Zhao,Wei Gong
Pareto Deep Long-Tailed Recognition: A Conflict-Averse Solution

Deep long-tailed recognition (DTLR) has attracted much attention due to its clos
e touch with realistic scenarios. Recent advances have focused on re-balancing a
cross various aspects, e.g., sampling strategy, loss re-weighting, logit adjustm
ent, and input/parameter perturbation, to name a few. However, few studies have
considered dynamic re-balancing to address intrinsic optimization conflicts. In
this paper, we first empirically argue that the optimizations of mainstream DLTR
 methods are still dominated by some categories (e.g., major) due to a fixed re-
balancing strategy. Thus, they fail to deal with gradient conflicts among catego
ries, which naturally deduces the motivation for reaching Pareto optimal solutio
ns. Unfortunately, a naive integration of multi-objective optimization (MOO) wit
h DLTR methods is not applicable due to the gap between multi-task learning (MTL
) and DLTR, and can in turn lead to class-specific feature degradation. Thus, we
 provide effective alternatives by decoupling MOO-based MTL from the temporal ra
ther than structure perspective, and enhancing it via optimizing variability col
lapse loss motivated by the derived MOO-based DLTR generalization bound. Moreove
r, we resort to anticipating worst-case optimization with theoretical insights t
o further ensure convergence. We build a Pareto deep long-tailed recognition met
hod termed PLOT upon the proposed MOO framework. Extensive evaluations demonstra
te that our method not only generally improves mainstream pipelines, but also ac
hieves an augmented version to realize state-of-the-art performance across multi
ple benchmarks.
****
Yu Meng,Jitin Krishnan,Sinong Wang,Qifan Wang,Yuning Mao,Han Fang,Marjan Ghazvin
inejad,Jiawei Han,Luke Zettlemoyer
Representation Deficiency in Masked Language Modeling
Masked Language Modeling (MLM) has been one of the most prominent approaches for
 pretraining bidirectional text encoders due to its simplicity and effectiveness
. One notable concern about MLM is that the special $\texttt{[MASK]}$ symbol cau
ses a discrepancy between pretraining data and downstream data as it is present
only in pretraining but not in fine-tuning. In this work, we offer a new perspec
tive on the consequence of such a discrepancy: We demonstrate empirically and th
eoretically that MLM pretraining allocates some model dimensions exclusively for
 representing $\texttt{[MASK]}$ tokens, resulting in a representation deficiency
 for real tokens and limiting the pretrained model's expressiveness when it is a
dapted to downstream data without $\texttt{[MASK]}$ tokens. Motivated by the ide
ntified issue, we propose MAE-LM, which pretrains the Masked Autoencoder archite
cture with MLM where $\texttt{[MASK]}$ tokens are excluded from the encoder. Emp
irically, we show that MAE-LM improves the utilization of model dimensions for r
eal token representations, and MAE-LM consistently outperforms MLM-pretrained mo
dels on the GLUE and SQuAD benchmarks.
****
Marah I Abdin,Suriya Gunasekar,Varun Chandrasekaran,Jerry Li,Mert Yuksekgonul,Ra
hee Ghosh Peshawaria,Ranjita Naik,Besmira Nushi
KITAB: Evaluating LLMs on Constraint Satisfaction for Information Retrieval
We study the ability of state-of-the art models to answer constraint satisfactio
n queries for information retrieval (e.g., “a list of ice cream shops in San Die
go”). In the past, such queries were considered as tasks that could only be solv
ed via web-search or knowledge bases. More recently, large language models (LLMs
) have demonstrated initial emergent abilities in this task. However, many curre
nt retrieval benchmarks are either saturated or do not measure constraint satisf
action. Motivated by rising concerns around factual incorrectness and hallucinat
ions of LLMs, we present KITAB, a new dataset for measuring constraint satisfact
ion abilities of language models. KITAB consists of book-related data across mor
e than 600 authors and 13,000 queries, and also offers an associated dynamic dat
a collection and constraint verification approach for acquiring similar test dat
a for other authors. Our extended experiments on GPT4 and GPT3.5 characterize an
d decouple common failure modes across dimensions such as information popularity
, constraint types, and context availability. Results show that in the absence o
f context, models exhibit severe limitations as measured by irrelevant informati
on, factual errors, and incompleteness, many of which exacerbate as information

popularity decreases. While context availability mitigates irrelevant information, it is not helpful for satisfying constraints, identifying fundamental barriers to constraint satisfaction. We open source our contributions to foster further research on improving constraint satisfaction abilities of future models.
****

Yilang Zhang,Georgios B. Giannakis

Meta-Learning Priors Using Unrolled Proximal Networks

Relying on prior knowledge accumulated from related tasks, meta-learning offers a powerful approach to learning a novel task from a limited number of training data. Recent approaches use a family of prior probability density functions or recurrent neural network models, whose parameters can be optimized by utilizing labeled data from the observed tasks. While these approaches have appealing empirical performance, expressiveness of their prior is relatively low, which limits generalization and interpretation of meta-learning. Aiming at expressive yet meaningful priors, this contribution puts forth a novel prior representation model that leverages the notion of algorithm unrolling. The key idea is to unroll the proximal gradient descent steps, where learnable piecewise linear functions are developed to approximate the desired proximal operators within *tight* theoretical error bounds established for both smooth and non-smooth proximal functions. The resultant multi-block neural network not only broadens the scope of learnable priors, but also enhances interpretability from an optimization viewpoint. Numerical tests conducted on few-shot learning datasets demonstrate markedly improved performance with flexible, visualizable, and understandable priors.
****

Ahmet Iscen,Mathilde Caron,Alireza Fathi,Cordelia Schmid

Retrieval-Enhanced Contrastive Vision-Text Models

Contrastive image-text models such as CLIP form the building blocks of many state-of-the-art systems. While they excel at recognizing common generic concepts, they still struggle on fine-grained entities which are rare, or even absent from the pre-training dataset. Hence, a key ingredient to their success has been the use of large-scale curated pre-training data aiming at expanding the set of concepts that they can memorize during the pre-training stage. In this work, we explore an alternative to encoding fine-grained knowledge directly into the model's parameters: we instead train the model to retrieve this knowledge from an external memory. Specifically, we propose to equip existing vision-text models with the ability to refine their embedding with cross-modal retrieved information from a memory at inference time, which greatly improves their zero-shot predictions. Remarkably, we show that this can be done with a light-weight, single-layer, fusion transformer on top of a frozen CLIP. Our experiments validate that our retrieval-enhanced contrastive (RECO) training improves CLIP performance substantially on several challenging fine-grained tasks: for example +10.9 on Stanford Cars, +10.2 on CUB-2011 and +7.3 on the recent OVEN benchmark, where we even outperform the fine-tuned models on unseen classes.
****

Hila Chefer,Oran Lang,Mor Geva,Volodymyr Polosukhin,Assaf Shocher,michal Irani,Inbar Mosseri,Lior Wolf

The Hidden Language of Diffusion Models

Text-to-image diffusion models have demonstrated an unparalleled ability to generate high-quality, diverse images from a textual prompt. However, the internal representations learned by these models remain an enigma. In this work, we present Conceptor, a novel method to interpret the internal representation of a textual concept by a diffusion model. This interpretation is obtained by decomposing the concept into a small set of human-interpretable textual elements. Applied over the state-of-the-art Stable Diffusion model, Conceptor reveals non-trivial structures in the representations of concepts. For example, we find surprising visual connections between concepts, that transcend their textual semantics. We additionally discover concepts that rely on mixtures of exemplars, biases, renowned artistic styles, or a simultaneous fusion of multiple meanings of the concept. Through a large battery of experiments, we demonstrate Conceptor's ability to provide meaningful, robust, and faithful decompositions for a wide variety of abst

ract, concrete, and complex textual concepts, while allowing to naturally connect each decomposition element to its corresponding visual impact on the generated images.
****
Maximilian Baader,Mark Niklas Mueller,Yuhao Mao,Martin Vechev
Expressivity of ReLU-Networks under Convex Relaxations
Convex relaxations are a key component of training and certifying provably safe neural networks. However, despite substantial progress, a wide and poorly understood accuracy gap to standard networks remains, raising the question of whether this is due to fundamental limitations of convex relaxations. Initial work investigating this question focused on the simple and widely used IBP relaxation. It revealed that some univariate, convex, continuous piecewise linear (CPWL) functions cannot be encoded by any ReLU network such that its IBP-analysis is precise. To explore whether this limitation is shared by more advanced convex relaxations, we conduct the first in-depth study on the expressive power of ReLU networks across all commonly used convex relaxations. We show that: (i) more advanced relaxations allow a larger class of univariate functions to be expressed as precisely analyzable ReLU networks, (ii) more precise relaxations can allow exponentially larger solution spaces of ReLU networks encoding the same functions, and (iii) even using the most precise single-neuron relaxations, it is impossible to construct precisely analyzable ReLU networks that express multivariate, convex, monotone CPWL functions.
****
Gianluca Bencomo,Jake Snell,Thomas L. Griffiths
Implicit Maximum a Posteriori Filtering via Adaptive Optimization
Bayesian filtering approximates the true underlying behavior of a time-varying system by inverting an explicit generative model to convert noisy measurements into state estimates. This process typically requires matrix storage, inversion, and multiplication or Monte Carlo estimation, none of which are practical in high-dimensional state spaces such as the weight spaces of artificial neural networks. Here, we consider the standard Bayesian filtering problem as optimization over a time-varying objective. Instead of maintaining matrices for the filtering equations or simulating particles, we specify an optimizer that defines the Bayesian filter implicitly. In the linear-Gaussian setting, we show that every Kalman filter has an equivalent formulation using K steps of gradient descent. In the nonlinear setting, our experiments demonstrate that our framework results in filters that are effective, robust, and scalable to high-dimensional systems, comparing well against the standard toolbox of Bayesian filtering solutions. We suggest that it is easier to fine-tune an optimizer than it is to specify the correct filtering equations, making our framework an attractive option for high-dimensional filtering problems.
****
Jian Xie,Kai Zhang,Jiangjie Chen,Renze Lou,Yu Su
Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts
By providing external information to large language models (LLMs), tool augmentation (including retrieval augmentation) has emerged as a promising solution for addressing the limitations of LLMs' static parametric memory.
However, how receptive are LLMs to such external evidence, especially when the evidence conflicts with their parametric memory?
We present the first comprehensive and controlled investigation into the behavior of LLMs when encountering knowledge conflicts.
We propose a systematic framework to elicit high-quality parametric memory from LLMs and construct the corresponding counter-memory, which enables us to conduct a series of controlled experiments.
Our investigation reveals seemingly contradicting behaviors of LLMs.
On the one hand, different from prior wisdom, we find that LLMs can be highly receptive to external evidence even when that conflicts with their parametric memory, given that the external evidence is coherent and convincing.
On the other hand, LLMs also demonstrate a strong confirmation bias when the ext

ernal evidence contains some information that is consistent with their parametri c memory, despite being presented with conflicting evidence at the same time. These results pose important implications that are worth careful consideration f or the further development and deployment of tool- and retrieval-augmented LLMs. Resources are available at https://github.com/OSU-NLP-Group/LLM-Knowledge-Confli ct.
****
Qiang He,Tianyi Zhou,Meng Fang,Setareh Maghsudi
Adaptive Regularization of Representation Rank as an Implicit Constraint of Bell man Equation
Representation rank is an important concept for understanding the role of Neural Networks (NNs) in Deep Reinforcement learning (DRL), which measures the express ive capacity of value networks. Existing studies focus on unboundedly maximizing this rank; nevertheless, that approach would introduce overly complex models in the learning, thus undermining performance. Hence, fine-tuning representation r ank presents a challenging and crucial optimization problem. To address this iss ue, we find a guiding principle for adaptive control of the representation rank. We employ the Bellman equation as a theoretical foundation and derive an upper bound on the cosine similarity of consecutive state-action pairs representations of value networks. We then leverage this upper bound to propose a novel regular izer, namely BEllman Equation-based automatic rank Regularizer (BEER). This regu larizer adaptively regularizes the representation rank, thus improving the DRL a gent's performance. We first validate the effectiveness of automatic control of rank on illustrative experiments. Then, we scale up BEER to complex continuous c ontrol tasks by combining it with the deterministic policy gradient method. Amon g 12 challenging DeepMind control tasks, BEER outperforms the baselines by a lar ge margin. Besides, BEER demonstrates significant advantages in Q-value approxim ation. Our code is available at https://github.com/sweetice/BEER-ICLR2024.
****
Yu Wang,Tong Zhao,Yuying Zhao,Yunchao Liu,Xueqi Cheng,Neil Shah,Tyler Derr
A Topological Perspective on Demystifying GNN-Based Link Prediction Performance
Graph Neural Networks (GNNs) have shown great promise in learning node embedding s for link prediction (LP). While numerous studies improve the overall GNNs' LP performance, none have explored their varying performance across different nodes and the underlying reasons. To this end, we demystify which nodes perform bette r from the perspective of their local topology. Despite the widespread belief th at low-degree nodes exhibit worse LP performance, we surprisingly observe an inc onsistent performance trend. This prompts us to propose a node-level metric, Top ological Concentration (TC), based on the intersection of the local subgraph of each node with the ones of its neighbors. We empirically demonstrate that TC cor relates with LP performance more than other node-level topological metrics, bett er identifying low-performing nodes than using degree. With TC, we discover a no vel topological distribution shift issue in which nodes' newly joined neighbors tend to become less interactive with their existing neighbors, compromising the generalizability of node embeddings for LP at testing time. To make the computat ion of TC scalable, We further propose Approximated Topological Concentration (A TC) and justify its efficacy in approximating TC with reduced computation comple xity. Given the positive correlation between node TC and its LP performance, we explore the potential of boosting LP performance via enhancing TC by re-weightin g edges in the message-passing and discuss its effectiveness with limitations. O ur code is publicly available at https://github.com/YuWVandy/Topo_LP_GNN.
****
Zijian Feng,Hanzhang Zhou,ZIXIAO ZHU,Junlang Qian,Kezhi Mao
Unveiling and Manipulating Prompt Influence in Large Language Models
Prompts play a crucial role in guiding the responses of Large Language Models (L LMs). However, the intricate role of individual tokens in prompts, known as inpu t saliency, in shaping the responses remains largely underexplored. Existing sal iency methods either misalign with LLM generation objectives or rely heavily on linearity assumptions, leading to potential inaccuracies. To address this, we pr opose Token Distribution Dynamics (TDD), an elegantly simple yet remarkably effe

ctive approach to unveil and manipulate the role of prompts in generating LLM ou tputs. TDD leverages the robust interpreting capabilities of the language model head (LM head) to assess input saliency. It projects input tokens into the embed ding space and then estimates their significance based on distribution dynamics over the vocabulary. We introduce three TDD variants: forward, backward, and bid irectional, each offering unique insights into token relevance. Extensive experi ments reveal that the TDD surpasses state-of-the-art baselines with a big margin in elucidating the causal relationships between prompts and LLM outputs. Beyond mere interpretation, we apply TDD to two prompt manipulation tasks for controll ed text generation: zero-shot toxic language suppression and sentiment steering. Empirical results underscore TDD's proficiency in identifying both toxic and se ntimental cues in prompts, subsequently mitigating toxicity or modulating sentim ent in the generated content.
****
Jiahuan Yan,Bo Zheng,Hongxia Xu,Yiheng Zhu,Danny Chen,Jimeng Sun,Jian Wu,Jintai Chen
Making Pre-trained Language Models Great on Tabular Prediction
The transferability of deep neural networks (DNNs) has made significant progress in image and language processing. However, due to the heterogeneity among table s, such DNN bonus is still far from being well exploited on tabular data predict ion (e.g., regression or classification tasks). Condensing knowledge from divers e domains, language models (LMs) possess the capability to comprehend feature na mes from various tables, potentially serving as versatile learners in transferri ng knowledge across distinct tables and diverse prediction tasks, but their disc rete text representation space is inherently incompatible with numerical feature values in tables. In this paper, we present TP-BERTa, a specifically pre-traine d LM for tabular data prediction. Concretely, a novel relative magnitude tokeniz ation converts scalar numerical feature values to finely discrete, high-dimensio nal tokens, and an intra-feature attention approach integrates feature values wi th the corresponding feature names. Comprehensive experiments demonstrate that o ur pre-trained TP-BERTa leads the performance among tabular DNNs and is competit ive with Gradient Boosted Decision Tree models in typical tabular data regime.
****
Szu-Wei Fu,Kuo-Hsuan Hung,Yu Tsao,Yu-Chiang Frank Wang
Self-Supervised Speech Quality Estimation and Enhancement Using Only Clean Speec h
Speech quality estimation has recently undergone a paradigm shift from human-hea ring expert designs to machine-learning models. However, current models rely mai nly on supervised learning, which is time-consuming and expensive for label coll ection. To solve this problem, we propose VQScore, a self-supervised metric for evaluating speech based on the quantization error of a vector-quantized-variatio nal autoencoder (VQ-VAE). The training of VQ-VAE relies on clean speech; hence, large quantization errors can be expected when the speech is distorted. To furth er improve correlation with real quality scores, domain knowledge of speech proc essing is incorporated into the model design. We found that the vector quantizat ion mechanism could also be used for self-supervised speech enhancement (SE) mod el training. To improve the robustness of the encoder for SE, a novel self-disti llation mechanism combined with adversarial training is introduced. In summary, the proposed speech quality estimation method and enhancement models require onl y clean speech for training without any label requirements. Experimental results show that the proposed VQScore and enhancement model are competitive with super vised baselines. The code and pre-trained models will be released
****
Ching Fang,Kim Stachenfeld
Predictive auxiliary objectives in deep RL mimic learning in the brain
The ability to predict upcoming events has been hypothesized to comprise a key a spect of natural and machine cognition. This is supported by trends in deep rein forcement learning (RL), where self-supervised auxiliary objectives such as pred iction are widely used to support representation learning and improve task perfo rmance. Here, we study the effects predictive auxiliary objectives have on repre

sentation learning across different modules of an RL system and how these mimic
representational changes observed in the brain. We find that predictive objectiv
es improve and stabilize learning particularly in resource-limited architectures
, and we identify settings where longer predictive horizons better support repre
sentational transfer. Furthermore, we find that representational changes in this
 RL system bear a striking resemblance to changes in neural activity observed in
 the brain across various experiments. Specifically, we draw a connection betwee
n the auxiliary predictive model of the RL system and hippocampus, an area thoug
ht to learn a predictive model to support memory-guided behavior. We also connec
t the encoder network and the value learning network of the RL system to visual
cortex and striatum in the brain, respectively. This work demonstrates how repre
sentation learning in deep RL systems can provide an interpretable framework for
 modeling multi-region interactions in the brain. The deep RL perspective taken
here also suggests an additional role of the hippocampus in the brain-- that of
an auxiliary learning system that benefits representation learning in other regi
ons.
****
Minh Pham,Kelly O. Marshall,Niv Cohen,Govind Mittal,Chinmay Hegde
Circumventing Concept Erasure Methods For Text-To-Image Generative Models
Text-to-image generative models can produce photo-realistic images for an extrem
ely broad range of concepts, and their usage has proliferated widely among the g
eneral public. On the flip side, these models have numerous drawbacks, including
 their potential to generate images featuring sexually explicit content, mirror
artistic styles without permission, or even hallucinate (or deepfake) the likene
sses of celebrities. Consequently, various methods have been proposed in order t
o "erase" sensitive concepts from text-to-image models. In this work, we examine
 seven recently proposed concept erasure methods, and show that targeted concept
s are not fully excised from any of these methods. Specifically, we leverage the
 existence of special learned word embeddings that can retrieve "erased" concept
s from the sanitized models with no alterations to their weights. Our results hi
ghlight the brittleness of post hoc concept erasure methods, and call into quest
ion their use in the algorithmic toolkit for AI safety.
****
Julius Kunze,Daniel Severo,Giulio Zani,Jan-Willem van de Meent,James Townsend
Entropy Coding of Unordered Data Structures
We present shuffle coding, a general method for optimal compression of sequences
 of unordered objects using bits-back coding. Data structures that can be compre
ssed using shuffle coding include multisets, graphs, hypergraphs, and others. We
 release an implementation that can easily be adapted to different data types an
d statistical models, and demonstrate that our implementation achieves state-of-
the-art compression rates on a range of graph datasets including molecular data.
****
Jeonghye Kim,Suyoung Lee,Woojun Kim,Youngchul Sung
Decision ConvFormer: Local Filtering in MetaFormer is Sufficient for Decision Ma
king
The recent success of Transformer in natural language processing has sparked its
 use in various domains. In offline reinforcement learning (RL), Decision Transf
ormer (DT) is emerging as a promising model based on Transformer. However, we di
scovered that the attention module of DT is not appropriate to capture the inher
ent local dependence pattern in trajectories of RL modeled as a Markov decision
process. To overcome the limitations of DT, we propose a novel action sequence p
redictor, named Decision ConvFormer (DC), based on the architecture of MetaForme
r, which is a general structure to process multiple entities in parallel and und
erstand the interrelationship among the multiple entities. DC employs local conv
olution filtering as the token mixer and can effectively capture the inherent lo
cal associations of the RL dataset. In extensive experiments, DC achieved state-
of-the-art performance across various standard RL benchmarks while requiring few
er resources. Furthermore, we show that DC better understands the underlying mea
ning in data and exhibits enhanced generalization capability.
****

Hila Manor,Tomer Michaeli
On the Posterior Distribution in Denoising: Application to Uncertainty Quantific
ation
Denoisers play a central role in many applications, from noise suppression in lo
w-grade imaging sensors, to empowering score-based generative models. The latter
 category of methods makes use of Tweedie's formula, which links the posterior m
ean in Gaussian denoising (*i*.*e*., the minimum MSE denoiser) with the score of
 the data distribution. Here, we derive a fundamental relation between the highe
r-order central moments of the posterior distribution, and the higher-order deri
vatives of the posterior mean. We harness this result for uncertainty quantifica
tion of pre-trained denoisers. Particularly, we show how to efficiently compute
the principal components of the posterior distribution for any desired region of
 an image, as well as to approximate the full marginal distribution along those
(or any other) one-dimensional directions. Our method is fast and memory-efficie
nt, as it does not explicitly compute or store the high-order moment tensors and
 it requires no training or fine tuning of the denoiser. Code and examples are a
vailable on the project [website](https://hilamanor.github.io/GaussianDenoisingP
osterior/).
****
Yulei Niu,Wenliang Guo,Long Chen,Xudong Lin,Shih-Fu Chang
SCHEMA: State CHangEs MAtter for Procedure Planning in Instructional Videos
We study the problem of procedure planning in instructional videos, which aims t
o make a goal-oriented sequence of action steps given partial visual state obser
vations. The motivation of this problem is to learn a structured and plannable s
tate and action space. Recent works succeeded in sequence modeling of steps with
 only sequence-level annotations accessible during training, which overlooked th
e roles of states in the procedures. In this work, we point out that State CHang
Es MAtter (SCHEMA) for procedure planning in instructional videos. We aim to est
ablish a more structured state space by investigating the causal relations betwe
en steps and states in procedures. Specifically, we explicitly represent each st
ep as state changes and track the state changes in procedures. For step represen
tation, we leveraged the commonsense knowledge in large language models (LLMs) t
o describe the state changes of steps via our designed chain-of-thought promptin
g. For state changes tracking, we align visual state observations with language
state descriptions via cross-modal contrastive learning, and explicitly model th
e intermediate states of the procedure using LLM-generated state descriptions. E
xperiments on CrossTask, COIN, and NIV benchmark datasets demonstrate that our p
roposed SCHEMA model achieves state-of-the-art performance and obtains explainab
le visualizations.
****
Yangming Li,Boris van Breugel,Mihaela van der Schaar
Soft Mixture Denoising: Beyond the Expressive Bottleneck of Diffusion Models
Because diffusion models have shown impressive performances in a number of tasks
, such as image synthesis, there is a trend in recent works to prove (with certa
in assumptions) that these models have strong approximation capabilities. In thi
s paper, we show that current diffusion models actually have an expressive bottl
eneck in backward denoising and some assumption made by existing theoretical gua
rantees is too strong. Based on this finding, we prove that diffusion models hav
e unbounded errors in both local and global denoising. In light of our theoretic
al studies, we introduce soft mixture denoising (SMD), an expressive and efficie
nt model for backward denoising. SMD not only permits diffusion models to well a
pproximate any Gaussian mixture distributions in theory, but also is simple and
efficient for implementation. Our experiments on multiple image datasets show th
at SMD significantly improves different types of diffusion models (e.g., DDPM),
espeically in the situation of few backward iterations.
****
Ahmed Hendawy,Jan Peters,Carlo D'Eramo
Multi-Task Reinforcement Learning with Mixture of Orthogonal Experts
Multi-Task Reinforcement Learning (MTRL) tackles the long-standing problem of en
dowing agents with skills that generalize across a variety of problems. To this

end, sharing representations plays a fundamental role in capturing both unique a
nd common characteristics of the tasks. Tasks may exhibit similarities in terms
of skills, objects, or physical properties while leveraging their representation
s eases the achievement of a universal policy. Nevertheless, the pursuit of lear
ning a shared set of diverse representations is still an open challenge. In this
 paper, we introduce a novel approach for representation learning in MTRL that e
ncapsulates common structures among the tasks using orthogonal representations t
o promote diversity. Our method, named Mixture Of Orthogonal Experts (MOORE), le
verages a Gram-Schmidt process to shape a shared subspace of representations gen
erated by a mixture of experts. When task-specific information is provided, MOOR
E generates relevant representations from this shared subspace. We assess the ef
fectiveness of our approach on two MTRL benchmarks, namely MiniGrid and MetaWorl
d, showing that MOORE surpasses related baselines and establishes a new state-of
-the-art result on MetaWorld.
****
Kaixuan Ji,Qingyue Zhao,Jiafan He,Weitong Zhang,Quanquan Gu
Horizon-free Reinforcement Learning in Adversarial Linear Mixture MDPs
Recent studies have shown that the regret of reinforcement learning (RL) can be
polylogarithmic in the planning horizon H. However, it remains an open questio
n whether such a result holds for adversarial RL. In this paper, we answer this
question affirmatively by proposing the first horizon-free policy search algorit
hm. To tackle the challenges caused by exploration and adversarially chosen rewa
rd over episodes, our algorithm employs (1) a variance-uncertainty-aware weighte
d least square estimator for the transition kernel; and (2) an occupancy measure
-based technique for the online search of a stochastic policy. We show that our
algorithm achieves an $\tilde{O}\big((d+\log |\mathcal{S}|)\sqrt{K} + d^2\big)$
regret with full-information feedback, where d is the dimension of a known fea
ture mapping linearly parametrizing the unknown transition kernel of the MDP, $K
$ is the number of episodes, $|\mathcal{S}|$ is the cardinality of the state spa
ce. We also provide hardness results to justify the near optimality of our algor
ithm and the inevitability of $\log|\mathcal{S}|$ in the regret bound.
****
WENLONG LIU,Tianyu Yang,Yuhan Wang,Qizhi Yu,Lei Zhang
Symbol as Points: Panoptic Symbol Spotting via Point-based Representation
This work studies the problem of panoptic symbol spotting, which is to spot and
parse both countable object instances (windows, doors, tables, etc.) and uncount
able stuff (wall, railing, etc.) from computer-aided design (CAD) drawings. Exis
ting methods typically involve either rasterizing the vector graphics into image
s and using image-based methods for symbol spotting, or directly building graphs
 and using graph neural networks for symbol recognition. In this paper, we take
a different approach, which treats graphic primitives as a set of 2D points that
 are locally connected and use point cloud segmentation methods to tackle it. Sp
ecifically, we utilize a point transformer to extract the primitive features and
 append a mask2former-like spotting head to predict the final output. To better
use the local connection information of primitives and enhance their discriminab
ility, we further propose the attention with connection module (ACM) and contras
tive connection learning scheme (CCL). Finally, we propose a KNN interpolation m
echanism for the mask attention module of the spotting head to better handle pri
mitive mask downsampling, which is primitive-level in contrast to pixel-level fo
r the image. Our approach, named SymPoint, is simple yet effective, outperformin
g recent state-of-the-art method GAT-CADNet by an absolute increase of 9.6% PQ a
nd 10.4% RQ on the FloorPlanCAD dataset. The source code and models will be avai
lable at \url{https://github.com/nicehuster/SymPoint}.
****
Gautam Reddy
The mechanistic basis of data dependence and abrupt learning in an in-context cl
assification task
Transformer models exhibit in-context learning: the ability to accurately predic
t the response to a novel query based on illustrative examples in the input sequ
ence, which contrasts with traditional in-weights learning of query-output relat

ionships. What aspects of the training data distribution and architecture favor in-context vs in-weights learning? Recent work has shown that specific distributional properties inherent in language, such as burstiness, large dictionaries and skewed rank-frequency distributions, control the trade-off or simultaneous appearance of these two forms of learning. We first show that these results are recapitulated in a minimal attention-only network trained on a simplified dataset. In-context learning (ICL) is driven by the abrupt emergence of an induction head, which subsequently competes with in-weights learning. By identifying progress measures that precede in-context learning and targeted experiments, we construct a two-parameter model of an induction head which emulates the full data distributional dependencies displayed by the attention-based network. A phenomenological model of induction head formation traces its abrupt emergence to the sequential learning of three nested logits enabled by an intrinsic curriculum. We propose that the sharp transitions in attention-based networks arise due to a specific chain of multi-layer operations necessary to achieve ICL, which is implemented by nested nonlinearities sequentially learned during training.
****
Matan Atzmon,Francis Williams,Jiahui Huang,Or Litany
Approximately Piecewise E(3) Equivariant Point Networks
Integrating a notion of symmetry into point cloud neural networks is a provably effective way to improve their generalization capability. Of particular interest are $E(3)$ equivariant point cloud networks where Euclidean transformations applied to the inputs are preserved in the outputs. Recent efforts aim to extend networks that are equivariant with respect to a single global $E(3)$ transformation, to accommodate inputs made of multiple parts, each of which exhibits local $E(3)$ symmetry.
In practical settings, however, the partitioning into individually transforming regions is unknown a priori.
Errors in the partition prediction would unavoidably map to errors in respecting the true input symmetry. Past works have proposed different ways to predict the partition, which may exhibit uncontrolled errors in their ability to maintain equivariance to the actual partition. To this end, we introduce APEN: a general framework for constructing approximate piecewise-$E(3)$ equivariant point networks. Our framework offers an adaptable design to guaranteed bounds on the resulting piecewise $E(3)$ equivariance approximation errors.
Our primary insight is that functions which are equivariant with respect to a finer partition (compared to the unknown true partition) will also maintain equivariance in relation to the true partition. Leveraging this observation, we propose a compositional design for a partition prediction model. It initiates with a fine partition and incrementally transitions towards a coarser subpartition of the true one, consistently maintaining piecewise equivariance in relation to the current partition.
As a result, the equivariance approximation error can be bounded solely in terms of (i) uncertainty quantification of the partition prediction, and (ii) bounds on the probability of failing to suggest a proper subpartition of the ground truth one.
We demonstrate the practical effectiveness of APEN using two data types exemplifying part-based symmetry: (i) real-world scans of room scenes containing multiple furniture-type objects; and, (ii) human motions, characterized by articulated parts exhibiting rigid movement. Our empirical results demonstrate the advantage of integrating piecewise $E(3)$ symmetry into network design, showing a distinct improvement in generalization over prior works in terms of generalization accuracy for both classification and segmentation tasks.
****
Ahmad Faiz,Sotaro Kaneda,Ruhan Wang,Rita Chukwunyere Osi,Prateek Sharma,Fan Chen,Lei Jiang
LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models
The carbon footprint associated with large language models (LLMs) is a significant concern, encompassing emissions from their training, inference, experimentation, and storage processes, including operational and embodied carbon emissions.

An essential aspect is accurately estimating the carbon impact of emerging LLMs even before their training, which heavily relies on GPU usage. Existing studies have reported the carbon footprint of LLM training, but only one tool, mlco2, can predict the carbon footprint of new neural networks prior to physical training. However, mlco2 has several serious limitations. It cannot extend its estimation to dense or mixture-of-experts (MoE) LLMs, disregards critical architectural parameters, focuses solely on GPUs, and cannot model embodied carbon footprints. Addressing these gaps, we introduce \textit{\carb}, an end-to-end carbon footprint projection model designed for both dense and MoE LLMs. Compared to mlco2, \carb~significantly enhances the accuracy of carbon footprint estimations for various LLMs. The source code is released at \url{https://github.com/SotaroKaneda/MLCarbon}.
****
Yanqi Bao,Tianyu Ding,Jing Huo,Wenbin Li,Yuxin Li,Yang Gao
InsertNeRF: Instilling Generalizability into NeRF with HyperNet Modules
Generalizing Neural Radiance Fields (NeRF) to new scenes is a significant challenge that existing approaches struggle to address without extensive modifications to vanilla NeRF framework. We introduce **InsertNeRF**, a method for **INS**tilling g**E**ne**R**alizabili**T**y into **NeRF**. By utilizing multiple plug-and-play HyperNet modules, InsertNeRF dynamically tailors NeRF's weights to specific reference scenes, transforming multi-scale sampling-aware features into scene-specific representations. This novel design allows for more accurate and efficient representations of complex appearances and geometries. Experiments show that this method not only achieves superior generalization performance but also provides a flexible pathway for integration with other NeRF-like systems, even in sparse input settings.
Code will be available at: https://github.com/bbbbby-99/InsertNeRF.
****
Rabia Gondur,Usama Bin Sikandar,Evan Schaffer,Mikio Christian Aoi,Stephen L Keeley
Multi-modal Gaussian Process Variational Autoencoders for Neural and Behavioral Data
Characterizing the relationship between neural population activity and behavioral data is a central goal of neuroscience. While latent variable models (LVMs) are successful in describing high-dimensional data, they are typically only designed for a single type of data, making it difficult to identify structure shared across different experimental data modalities. Here, we address this shortcoming by proposing an unsupervised LVM which extracts shared and independent latents for distinct, simultaneously recorded experimental modalities. We do this by combining Gaussian Process Factor Analysis (GPFA), an interpretable LVM for neural spiking data with temporally smooth latent space, with Gaussian Process Variational Autoencoders (GP-VAEs), which similarly use a GP prior to characterize correlations in a latent space, but admit rich expressivity due to a deep neural network mapping to observations. We achieve interpretability in our model by partitioning latent variability into components that are either shared between or independent to each modality. We parameterize the latents of our model in the Fourier domain, and show improved latent identification using this approach over standard GP-VAE methods. We validate our model on simulated multi-modal data consisting of Poisson spike counts and MNIST images that scale and rotate smoothly over time. We show that the multi-modal GP-VAE (MM-GPVAE) is able to not only identify the shared and independent latent structure across modalities accurately, but provides good reconstructions of both images and neural rates on held-out trials. Finally, we demonstrate our framework on two real world multi-modal experimental settings: Drosophila whole-brain calcium imaging alongside tracked limb positions, and Manduca sexta spike train measurements from ten wing muscles as the animal tracks a visual stimulus.
****
Xiaoyi Liu,Duxin Chen,Wenjia Wei,Xia Zhu,Wenwu Yu
Interpretable Sparse System Identification: Beyond Recent Deep Learning Techniques on Time-Series Prediction

With the continuous advancement of neural network methodologies, time series pre
diction has attracted substantial interest over the past decades. Nonetheless, t
he interpretability of neural networks is insufficient and the utilization of de
ep learning techniques for prediction necessitates significant computational exp
enditures, rendering its application arduous in numerous scenarios. In order to
tackle this challenge, an interpretable sparse system identification method whic
h does not require a time-consuming training through back-propagation is propose
d in this study. This method integrates advantages from both knowledge-based and
 data-driven approaches, and constructs dictionary functions by leveraging Fouri
er basis and taking into account both the long-term trends and the short-term fl
uctuations behind data. By using the l_1 norm for sparse optimization, predict
ion results can be gained with an explicit sparse expression function and an ext
remely high accuracy. The performance evaluation of the proposed method is condu
cted on comprehensive benchmark datasets, including ETT, Exchange, and ILI. Resu
lts reveal that our proposed method attains a significant overall improvement of
 more than 20\% in accordance with the most recent state-of-the-art deep learnin
g methodologies. Additionally, our method demonstrates the efficient training ca
pability on only CPUs. Therefore, this study may shed some light onto the realm
of time series reconstruction and prediction.
****
Daniil Kirilenko,Vitaliy Vorobyov,Alexey Kovalev,Aleksandr Panov
Object-Centric Learning with Slot Mixture Module
Object-centric architectures usually apply a differentiable module to the entire
 feature map to decompose it into sets of entity representations called slots. S
ome of these methods structurally resemble clustering algorithms, where the clus
ter's center in latent space serves as a slot representation. Slot Attention is
an example of such a method, acting as a learnable analog of the soft k-means al
gorithm. Our work employs a learnable clustering method based on the Gaussian Mi
xture Model. Unlike other approaches, we represent slots not only as centers of
clusters but also incorporate information about the distance between clusters an
d assigned vectors, leading to more expressive slot representations. Our experim
ents demonstrate that using this approach instead of Slot Attention improves per
formance in object-centric scenarios, achieving state-of-the-art results in the
set property prediction task.
****
Kai Chen,Chunwei Wang,Kuo Yang,Jianhua Han,Lanqing HONG,Fei Mi,Hang Xu,Zhengying
 Liu,Wenyong Huang,Zhenguo Li,Dit-Yan Yeung,Lifeng Shang
Gaining Wisdom from Setbacks: Aligning Large Language Models via Mistake Analysi
s
The rapid development of large language models (LLMs) has not only provided nume
rous opportunities but also presented significant challenges. This becomes parti
cularly evident when LLMs inadvertently generate harmful or toxic content, eithe
r unintentionally or because of intentional inducement. Existing alignment metho
ds usually direct LLMs toward the favorable outcomes by utilizing human-annotate
d, flawless instruction-response pairs. Conversely, this study proposes a novel
alignment technique based on mistake analysis, which deliberately exposes LLMs t
o erroneous content to learn the reasons for mistakes and how to avoid them. In
this case, mistakes are repurposed into valuable data for alignment, effectively
 helping to avoid the production of erroneous responses. Without external models
 or human annotations, our method leverages a model's intrinsic ability to disce
rn undesirable mistakes and improves the safety of its generated responses. Expe
rimental results reveal that our method outperforms existing alignment approache
s in enhancing model safety while maintaining the overall utility.
****
Mahdi Alikhasi,Levi Lelis
Unveiling Options with Neural Network Decomposition
In reinforcement learning, agents often learn policies for specific tasks withou
t the ability to generalize this knowledge to related tasks. This paper introduc
es an algorithm that attempts to address this limitation by decomposing neural n
etworks encoding policies for Markov Decision Processes into reusable sub-polici

es, which are used to synthesize temporally extended actions, or options. We consider neural networks with piecewise linear activation functions, so that they can be mapped to an equivalent tree that is similar to oblique decision trees. Since each node in such a tree serves as a function of the input of the tree, each sub-tree is a sub-policy of the main policy. We turn each of these sub-policies into options by wrapping it with while-loops of varied number of iterations. Given the large number of options, we propose a selection mechanism based on minimizing the Levin loss for a uniform policy on these options. Empirical results in two grid-world domains where exploration can be difficult confirm that our method can identify useful options, thereby accelerating the learning process on similar but different tasks.
****
Victor Akinwande,Yiding Jiang,Dylan Sam,J Zico Kolter
Understanding prompt engineering may not require rethinking generalization
Zero-shot learning in prompted vision-language models, the practice of crafting prompts to build classifiers without an explicit training process, has achieved impressive performance in many settings. This success presents a seemingly surprising observation: these methods suffer relatively little from overfitting, i.e., when a prompt is manually engineered to achieve low error on a given training set (thus rendering the method no longer actually zero-shot), the approach still performs well on held-out test data. In this paper, we show that we can explain such performance well via recourse to classical PAC-Bayes bounds. Specifically, we show that the discrete nature of prompts, combined with a PAC-Bayes prior given by a language model, results in generalization bounds that are remarkably tight by the standards of the literature: for instance, the generalization bound of an ImageNet classifier is often within a few percentage points of the true test error. We demonstrate empirically that this holds for existing handcrafted prompts and prompts generated through simple greedy search. Furthermore, the resulting bound is well-suited for model selection: the models with the best bound typically also have the best test performance. This work thus provides a possible justification for the widespread practice of "prompt engineering," even if it seems that such methods could potentially overfit the training data.
****
Shengzhong Zhang,Wenjie Yang,Xinyuan Cao,Hongwei Zhang,Zengfeng Huang
StructComp: Substituting propagation with Structural Compression in Training Graph Contrastive Learning
Graph contrastive learning (GCL) has become a powerful tool for learning graph data, but its scalability remains a significant challenge. In this work, we propose a simple yet effective training framework called Structural Compression (StructComp) to address this issue. Inspired by a sparse low-rank approximation on the diffusion matrix, StructComp trains the encoder with the compressed nodes. This allows the encoder not to perform any message passing during the training stage, and significantly reduces the number of sample pairs in the contrastive loss. We theoretically prove that the original GCL loss can be approximated with the contrastive loss computed by StructComp. Moreover, StructComp can be regarded as an additional regularization term for GCL models, resulting in a more robust encoder. Empirical studies on various datasets show that StructComp greatly reduces the time and memory consumption while improving model performance compared to the vanilla GCL models and scalable training methods.
****
Jongwon Jeong,Hoyeop Lee,Hyui Geon Yoon,Beomyoung Lee,Junhee Heo,Geonsoo Kim,Kim Jin Seon
iGraphMix: Input Graph Mixup Method for Node Classification
Recently, Input Mixup, which augments virtual samples by interpolating input features and corresponding labels, is one of the promising methods to alleviate the over-fitting problem on various domains including image classification and natural language processing because of its ability to generate a variety of virtual samples, and ease of usability and versatility. However, designing Input Mixup for the node classification is still challenging due to the irregularity issue that each node contains a different number of neighboring nodes for input and the

alignment issue that how to align and interpolate two sets of neighboring nodes is not well-defined when two nodes are interpolated. To address the issues, this paper proposes a novel Mixup method, called iGraphMix, tailored to node classification. Our method generates virtual nodes and their edges by interpolating input features and labels, and attaching sampled neighboring nodes. The virtual graphs generated by iGraphMix serve as inputs for graph neural networks (GNNs) training, thereby facilitating its easy application to various GNNs and enabling effective combination with other augmentation methods. We mathematically prove that training GNNs with iGraphMix leads to better generalization performance compared to that without augmentation, and our experiments support the theoretical findings.
****
Michalis Titsias,Alexandre Galashov,Amal Rannen-Triki,Razvan Pascanu,Yee Whye Teh,Jorg Bornschein
Kalman Filter for Online Classification of Non-Stationary Data
In Online Continual Learning (OCL) a learning system receives a stream of data and sequentially performs prediction and training steps. Key challenges in OCL include automatic adaptation to the specific non-stationary structure of the data and maintaining appropriate predictive uncertainty. To address these challenges we introduce a probabilistic Bayesian online learning approach that utilizes a (possibly pretrained) neural representation and a state space model over the linear predictor weights. Non-stationarity in the linear predictor weights is modelled using a “parameter drift” transition density, parametrized by a coefficient that quantifies forgetting. Inference in the model is implemented with efficient Kalman filter recursions which track the posterior distribution over the linear weights, while online SGD updates over the transition dynamics coefficient allow for adaptation to the non-stationarity observed in the data. While the framework is developed assuming a linear Gaussian model, we extend it to deal with classification problems and for fine-tuning the deep learning representation. In a set of experiments in multi-class classification using data sets such as CIFAR-100 and CLOC we demonstrate the model's predictive ability and its flexibility in capturing non-stationarity.
****
Cong Liu,David Ruhe,Floor Eijkelboom,Patrick Forré
Clifford Group Equivariant Simplicial Message Passing Networks
We introduce Clifford Group Equivariant Simplicial Message Passing Networks, a method for steerable $\mathrm{E}(n)$-equivariant message passing on simplicial complexes. Our method integrates the expressivity of Clifford group-equivariant layers with simplicial message passing, which is topologically more intricate than regular graph message passing. Clifford algebras include higher-order objects such as bivectors and trivectors, which express geometric features (e.g., areas, volumes) derived from vectors. Using this knowledge, we represent simplex features through geometric products of their vertices. To achieve efficient simplicial message passing, we share the parameters of the message network across different dimensions. Additionally, we restrict the final message to an aggregation of the incoming messages from different dimensions, leading to what we term *shared* simplicial message passing. Experimental results show that our method is able to outperform both equivariant and simplicial graph neural networks on a variety of geometric tasks.
****
Luciano Dyballa,Samuel Lang,Alexandra Haslund-Gourley,Eviatar Yemini,Steven W. Zucker
Learning dynamic representations of the functional connectome in neurobiological networks
The static synaptic connectivity of neuronal circuits stands in direct contrast to the dynamics of their function. As in changing community interactions, different neurons can participate actively in various combinations to effect behaviors at different times. We introduce an unsupervised approach to learn the dynamic affinities between neurons in live, behaving animals, and to reveal which communities form among neurons at different times. The inference occurs in two major s

teps. First, pairwise non-linear affinities between neuronal traces from brain-wide calcium activity are organized by non-negative tensor factorization (NTF). Each factor specifies which groups of neurons are most likely interacting for an inferred interval in time, and for which animals. Finally, a generative model that allows for weighted community detection is applied to the functional motifs produced by NTF to reveal a dynamic functional connectome. Since time codes the different experimental variables (e.g., application of chemical stimuli), this provides an atlas of neural motifs active during separate stages of an experiment (e.g., stimulus application or spontaneous behaviors). Results from our analysis are experimentally validated, confirming that our method is able to robustly predict causal interactions between neurons to generate behavior.
****
Jisu Nam,Gyuseong Lee,Sunwoo Kim,Hyeonsu Kim,Hyoungwon Cho,Seyeon Kim,Seungryong Kim
Diffusion Model for Dense Matching
The objective for establishing dense correspondence between paired images con- sists of two terms: a data term and a prior term. While conventional techniques focused on defining hand-designed prior terms, which are difficult to formulate, re- cent approaches have focused on learning the data term with deep neural networks without explicitly modeling the prior, assuming that the model itself has the capacity to learn an optimal prior from a large-scale dataset. The performance improvement was obvious, however, they often fail to address inherent ambiguities of matching, such as textureless regions, repetitive patterns, large displacements, or noises. To address this, we propose DiffMatch, a novel conditional diffusion-based framework designed to explicitly model both the data and prior terms for dense matching. This is accomplished by leveraging a conditional denoising diffusion model that explic- itly takes matching cost and injects the prior within generative process. However, limited input resolution of the diffusion model is a major hindrance. We address this with a cascaded pipeline, starting with a low-resolution model, followed by a super-resolution model that successively upsamples and incorporates finer details to the matching field. Our experimental results demonstrate significant performance improvements of our method over existing approaches, and the ablation studies validate our design choices along with the effectiveness of each component. Code and pretrained weights are available at https://ku-cvlab.github.io/DiffMatch.
****
Jiawei Zhou,Xiaoguang Li,Lifeng Shang,Xin Jiang,Qun Liu,Lei Chen
Retrieval-based Disentangled Representation Learning with Natural Language Supervision
Disentangled representation learning remains challenging as the underlying factors of variation in the data do not naturally exist. The inherent complexity of real-world data makes it unfeasible to exhaustively enumerate and encapsulate all its variations within a finite set of factors. However, it is worth noting that most real-world data have linguistic equivalents, typically in the form of textual descriptions. These linguistic counterparts can represent the data and effortlessly decomposed into distinct tokens. In light of this, we present Vocabulary Disentangled Retrieval (VDR), a retrieval-based framework that harnesses natural language as proxies of the underlying data variation to drive disentangled representation learning. Our approach employ a bi-encoder model to represent both data and natural language in a vocabulary space, enabling the model to distinguish dimensions that capture intrinsic characteristics within data through its natural language counterpart, thus facilitating disentanglement. We extensively assess the performance of VDR across 15 retrieval benchmark datasets, covering text-to-text and cross-modal retrieval scenarios, as well as human evaluation. Our experimental results compellingly demonstrate the superiority of VDR over previous bi-encoder retrievers with comparable model size and training costs, achieving an impressive 8.7% improvement in NDCG@10 on the BEIR benchmark, a 5.3\% increase on MS COCO, and a 6.0% increase on Flickr30k in terms of mean recall in the zero-shot setting. Moreover, The results from human evaluation indicate that interpretability of our method is on par with SOTA captioning models.

****
Chao Chen,Kai Liu,Ze Chen,Yi Gu,Yue Wu,Mingyuan Tao,Zhihang Fu,Jieping Ye
INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection
Knowledge hallucination have raised widespread concerns for the security and reliability of deployed LLMs. Previous efforts in detecting hallucinations have been employed at logit-level uncertainty estimation or language-level self-consistency evaluation, where the semantic information is inevitably lost during the token-decoding procedure. Thus, we propose to explore the dense semantic information retained within LLMs' \textbf{IN}ternal \textbf{S}tates for halluc\textbf{I}nation \textbf{DE}tection (\textbf{INSIDE}). In particular, a simple yet effective \textbf{EigenScore} metric is proposed to better evaluate responses' self-consistency, which exploits the eigenvalues of responses' covariance matrix to measure the semantic consistency/diversity in the dense embedding space. Furthermore, from the perspective of self-consistent hallucination detection, a test time feature clipping approach is explored to truncate extreme activations in the internal states, which reduces overconfident generations and potentially benefits the detection of overconfident hallucinations. Extensive experiments and ablation studies are performed on several popular LLMs and question-answering (QA) benchmarks, showing the effectiveness of our proposal.
****
Marten Lienen,David Lüdke,Jan Hansen-Palmus,Stephan Günnemann
From Zero to Turbulence: Generative Modeling for 3D Flow Simulation
Simulations of turbulent flows in 3D are one of the most expensive simulations in computational fluid dynamics (CFD). Many works have been written on surrogate models to replace numerical solvers for fluid flows with faster, learned, autoregressive models. However, the intricacies of turbulence in three dimensions necessitate training these models with very small time steps, while generating realistic flow states requires either long roll-outs with many steps and significant error accumulation or starting from a known, realistic flow state—something we aimed to avoid in the first place. Instead, we propose to approach turbulent flow simulation as a generative task directly learning the manifold of all possible turbulent flow states without relying on any initial flow state. For our experiments, we introduce a challenging 3D turbulence dataset of high-resolution flows and detailed vortex structures caused by various objects and derive two novel sample evaluation metrics for turbulent flows. On this dataset, we show that our generative model captures the distribution of turbulent flows caused by unseen objects and generates high-quality, realistic samples amenable for downstream applications without access to any initial state.
****
Qingyue Zhao,Banghua Zhu
Towards the Fundamental Limits of Knowledge Transfer over Finite Domains
We characterize the statistical efficiency of knowledge transfer through n samples from a teacher to a probabilistic student classifier with input space $\mathcal{S}$ over labels $\mathcal{A}$. We show that privileged information at three progressive levels accelerates the transfer. At the first level, only samples with hard labels are known, via which the maximum likelihood estimator attains the minimax rate $\sqrt{{|\mathcal{S}||\mathcal{A}|}/{n}}$. The second level has the teacher probabilities of sampled labels available in addition, which turns out to boost the convergence rate lower bound to ${{|\mathcal{S}||\mathcal{A}|}/{n}}$. However, under this second data acquisition protocol, minimizing a naive adaptation of the cross-entropy loss results in an asymptotically biased student. We overcome this limitation and achieve the fundamental limit by using a novel empirical variant of the squared error logit loss. The third level further equips the student with the soft labels (complete logits) on $\mathcal{A}$ given every sampled input, thereby provably enables the student to enjoy a rate ${|\mathcal{S}|}/{n}$ free of $|\mathcal{A}|$. We find any Kullback-Leibler divergence minimizer to be optimal in the last case. Numerical simulations distinguish the four learners and corroborate our theory.
****
Dominique Beaini,Shenyang Huang,Joao Alex Cunha,Zhiyi Li,Gabriela Moisescu-Parej

a,Oleksandr Dymov,Samuel Maddrell-Mander,Callum McLean,Frederik Wenkel,Luis Müller,Jama Hussein Mohamud,Ali Parviz,Michael Craig,Micha■ Koziarski,Jiarui Lu,Zhaocheng Zhu,Cristian Gabellini,Kerstin Klaser,Josef Dean,Cas Wognum,Maciej Sypetkowski,Guillaume Rabusseau,Reihaneh Rabbany,Jian Tang,Christopher Morris,Mirco Ravanelli,Guy Wolf,Prudencio Tossou,Hadrien Mary,Therence Bois,Andrew William Fitzgibbon,Blazej Banaszewski,Chad Martin,Dominic Masters
Towards Foundational Models for Molecular Learning on Large-Scale Multi-Task Datasets
Recently, pre-trained foundation models have enabled significant advancements in multiple fields. In molecular machine learning, however, where datasets are often hand-curated, and hence typically small, the lack of datasets with labeled features, and codebases to manage those datasets, has hindered the development of foundation models. In this work, we present seven novel datasets categorized by size into three distinct categories: ToyMix, LargeMix and UltraLarge. These datasets push the boundaries in both the scale and the diversity of supervised labels for molecular learning. They cover nearly 100 million molecules and over 3000 sparsely defined tasks, totaling more than 13 billion individual labels of both quantum and biological nature. In comparison, our datasets contain 300 times more data points than the widely used OGB-LSC PCQM4Mv2 dataset, and 13 times more than the quantum-only QM1B dataset. In addition, to support the development of foundational models based on our proposed datasets, we present the Graphium graph machine learning library which simplifies the process of building and training molecular machine learning models for multi-task and multi-level molecular datasets. Finally, we present a range of baseline results as a starting point of multi-task and multi-level training on these datasets. Empirically, we observe that performance on low-resource biological datasets show improvement by also training on large amounts of quantum data. This indicates that there may be potential in multi-task and multi-level training of a foundation model and fine-tuning it to resource-constrained downstream tasks. The Graphium library is publicly available on Github and the dataset links are available in Part 1 and Part 2.
****
Haixin Wang,Jiaxin LI,Anubhav Dwivedi,Kentaro Hara,Tailin Wu
BENO: Boundary-embedded Neural Operators for Elliptic PDEs
Elliptic partial differential equations (PDEs) are a major class of time-independent PDEs that play a key role in many scientific and engineering domains such as fluid dynamics, plasma physics, and solid mechanics. Recently, neural operators have emerged as a promising technique to solve elliptic PDEs more efficiently by directly mapping the input to solutions. However, existing networks typically neglect complex geometries and inhomogeneous boundary values present in the real world. Here we introduce Boundary-Embedded Neural Operators (BENO), a novel neural operator architecture that embeds the complex geometries and inhomogeneous boundary values into the solving of elliptic PDEs. Inspired by classical Green's function, BENO consists of two Graph Neural Networks (GNNs) for interior source term and boundary values, respectively. Furthermore, a Transformer encoder maps the global boundary geometry into a latent vector which influences each message passing layer of the GNNs. We test our model and strong baselines extensively in elliptic PDEs with complex boundary conditions. We show that all existing baseline methods fail to learn the solution operator. In contrast, our model, endowed with boundary-embedded architecture, outperforms state-of-the-art neural operators and strong baselines by an average of 60.96%.
****
Yochai Yemini,Aviv Shamsian,Lior Bracha,Sharon Gannot,Ethan Fetaya
LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading
Lip-to-speech involves generating a natural-sounding speech synchronized with a soundless video of a person talking. Despite recent advances, current methods still cannot produce high-quality speech with high levels of intelligibility for challenging and realistic datasets such as LRS3. In this work, we present LipVoicer, a novel method that generates high-quality speech, even for in-the-wild and rich datasets, by incorporating the text modality. Given a silent video, we first predict the spoken text using a pre-trained lip-reading network. We then condi

tion a diffusion model on the video and use the extracted text through a classif
ier-guidance mechanism where a pre-trained automatic speech recognition (ASR) s
erves as the classifier. LipVoicer outperforms multiple lip-to-speech baselines
on LRS2 and LRS3, which are in-the-wild datasets with hundreds of unique speaker
s in their test set and an unrestricted vocabulary. Moreover, our experiments sh
ow that the inclusion of the text modality plays a major role in the intelligibi
lity of the produced speech, readily perceptible while listening, and is empiric
ally reflected in the substantial reduction of the word error rate (WER) metri
c. We demonstrate the effectiveness of LipVoicer through human evaluation, which
 shows that it produces more natural and synchronized speech signals compared to
 competing methods. Finally, we created a demo showcasing LipVoicer’s superiorit
y in producing natural, synchronized, and intelligible speech, providing additio
nal evidence of its effectiveness. Project page and code: https://github.com/yoc
haiye/LipVoicer
****
Ganchao Wei
Bayesian Bi-clustering of Neural Spiking Activity with Latent Structures
Modern neural recording techniques allow neuroscientists to obtain spiking activ
ity of multiple neurons from different brain regions over long time periods, whi
ch requires new statistical methods to be developed for understanding structure
of the large-scale data. In this paper, we develop a bi-clustering method to clu
ster the neural spiking activity spatially and temporally, according to their lo
w-dimensional latent structures. The spatial (neuron) clusters are defined by th
e latent trajectories within each neural population, while the temporal (state)
clusters are defined by (populationally) synchronous local linear dynamics share
d with different periods. To flexibly extract the bi-clustering structure, we bu
ild the model non-parametrically, and develop an efficient Markov chain Monte Ca
rlo (MCMC) algorithm to sample the posterior distributions of model parameters.
Validating our proposed MCMC algorithm through simulations, we find the method c
an recover unknown parameters and true bi-clustering structures successfully. We
 then apply the proposed bi-clustering method to multi-regional neural recording
s under different experiment settings, where we find that simultaneously conside
ring latent trajectories and spatial-temporal clustering structures can provide
us with a more accurate and interpretable result. Overall, the proposed method p
rovides scientific insights for large-scale (counting) time series with elongate
d recording periods, and it can potentially have application beyond neuroscience
.
****
Brandon Trabucco,Kyle Doherty,Max A Gurinas,Ruslan Salakhutdinov
Effective Data Augmentation With Diffusion Models
Data augmentation is one of the most prevalent tools in deep learning, underpinn
ing many recent advances, including those from classification, generative models
, and representation learning. The standard approach to data augmentation combin
es simple transformations like rotations and flips to generate new images from e
xisting ones. However, these new images lack diversity along key semantic axes p
resent in the data. Current augmentations cannot alter the high-level semantic a
ttributes, such as animal species present in a scene, to enhance the diversity o
f data. We address the lack of diversity in data augmentation with image-to-imag
e transformations parameterized by pre-trained text-to-image diffusion models. O
ur method edits images to change their semantics using an off-the-shelf diffusio
n model, and generalizes to novel visual concepts from a few labelled examples.
We evaluate our approach on few-shot image classification tasks, and on a real-w
orld weed recognition task, and observe an improvement in accuracy in tested dom
ains.
****
Shikai Fang,Xin Yu,Zheng Wang,Shibo Li,Mike Kirby,Shandian Zhe
Functional Bayesian Tucker Decomposition for Continuous-indexed Tensor Data
Tucker decomposition is a powerful tensor model to handle multi-aspect data. It
demonstrates the low-rank property by decomposing the grid-structured data as in
teractions between a core tensor and a set of object representations (factors).

A fundamental assumption of such decomposition is that there are finite objects in each aspect or mode, corresponding to discrete indexes of data entries. Howe ver, real-world data is often not naturally posed in this setting. For example, geographic data is represented as continuous indexes of latitude and longitude coordinates, and cannot fit tensor models directly. To generalize Tucker decompo sition to such scenarios, we propose Functional Bayesian Tucker Decomposition (F unBaT). We treat the continuous-indexed data as the interaction between the Tuck er core and a group of latent functions. We use Gaussian processes (GP) as func tional priors to model the latent functions. Then, we convert each GP into a sta te-space prior by constructing an equivalent stochastic differential equation (S DE) to reduce computational cost. An efficient inference algorithm is developed for scalable posterior approximation based on advanced message-passing techniqu es. The advantage of our method is shown in both synthetic data and several real -world applications. We release the code of FunBaT at {https://github.com/xuangu -fang/Functional-Bayesian-Tucker-Decomposition}.
****
Zhihe YANG,Yunjian Xu
DMBP: Diffusion model-based predictor for robust offline reinforcement learning against state observation perturbations
Offline reinforcement learning (RL), which aims to fully explore offline dataset s for training without interaction with environments, has attracted growing rece nt attention. A major challenge for the real-world application of offline RL ste ms from the robustness against state observation perturbations, e.g., as a resul t of sensor errors or adversarial attacks. Unlike online robust RL, agents canno t be adversarially trained in the offline setting. In this work, we propose Diff usion Model-Based Predictor (DMBP) in a new framework that recovers the actual s tates with conditional diffusion models for state-based RL tasks. To mitigate th e error accumulation issue in model-based estimation resulting from the classica l training of conventional diffusion models, we propose a non-Markovian training objective to minimize the sum entropy of denoised states in RL trajectory. Expe riments on standard benchmark problems demonstrate that DMBP can significantly e nhance the robustness of existing offline RL algorithms against different scales of ran- dom noises and adversarial attacks on state observations. Further, the proposed framework can effectively deal with incomplete state observations with random combinations of multiple unobserved dimensions in the test. Our implement ation is available at https://github.com/zhyang2226/DMBP.
****
Ananya Kumar,Ruoqi Shen,Sebastien Bubeck,Suriya Gunasekar
How to Fine-Tune Vision Models with SGD
SGD and AdamW are the two most used optimizers for fine-tuning large neural netw orks in computer vision. When the two methods perform the same, SGD is preferabl e because it uses less memory (12 bytes/parameter with momentum and 8 bytes/para meter without) than AdamW (16 bytes/parameter). However, on a suite of downstrea m tasks, especially those with distribution shifts, we find that fine-tuning wit h AdamW performs substantially better than SGD on modern Vision Transformer and ConvNeXt models. We find that large gaps in performance between SGD and AdamW oc cur when the fine-tuning gradients in the first "embedding" layer are much large r than in the rest of the model. Our analysis suggests an easy fix that works co nsistently across datasets and models: freezing the embedding layer (less than 1 % of the parameters) leads to SGD with or without momentum performing slightly b etter than AdamW while using less memory (e.g., on ViT-L, SGD uses 33% less GPU memory). Our insights result in state-of-the-art accuracies on five popular dist ribution shift benchmarks: WILDS-FMoW, WILDS-Camelyon, BREEDS-Living-17, Waterbi rds, and DomainNet.
****
Mingxiao Li,Tingyu Qu,Ruicong Yao,Wei Sun,Marie-Francine Moens
Alleviating Exposure Bias in Diffusion Models through Sampling with Shifted Time Steps
Diffusion Probabilistic Models (DPM) have shown remarkable efficacy in the synth esis of high-quality images. However, their inference process characteristically

requires numerous, potentially hundreds, of iterative steps, which could exaggerate the problem of exposure bias due to the training and inference discrepancy. Previous work has attempted to mitigate this issue by perturbing inputs during training, which consequently mandates the retraining of the DPM. In this work, we conduct a systematic study of exposure bias in DPM and, intriguingly, we find that the exposure bias could be alleviated with a novel sampling method that we propose, without retraining the model. We empirically and theoretically show that, during inference, for each backward time step t and corresponding state ˆxt, there might exist another time step t_s which exhibits superior coupling with $\hat{x}_t$. Based on this finding, we introduce a sampling method
named Time-Shift Sampler. Our framework can be seamlessly integrated to existing sampling algorithms, such as DDPM, DDIM and other high-order solvers, inducing merely minimal additional computations. Experimental results show our method brings significant and consistent improvements in FID scores on different datasets and sampling methods. For example, integrating Time-Shift Sampler to F-PNDM yields a FID=3.88, achieving 44.49% improvements as compared to F-PNDM, on CIFAR-10 with 10 sampling steps, which is more performant than the vanilla DDIM with 100 sampling steps. Our code is available at https://github.com/Mingxiao-Li/TS-DPM.
****
Ori Yoran,Tomer Wolfson,Ori Ram,Jonathan Berant
Making Retrieval-Augmented Language Models Robust to Irrelevant Context
Retrieval-augmented language models (RALMs) hold promise to produce language understanding systems that are are factual, efficient, and up-to-date. An important desideratum of RALMs, is that retrieved information helps model performance when it is relevant, and does not harm performance when it is not. This is particularly important in multi-hop reasoning scenarios, where misuse of irrelevant evidence can lead to cascading errors. However, recent work has shown that retrieval augmentation can sometimes have a negative effect on performance. In this work, we present a thorough analysis on five open-domain question answering benchmarks, characterizing cases when retrieval reduces accuracy. We then propose two methods to mitigate this issue. First, a simple baseline that filters out retrieved passages that do not entail question-answer pairs according to a natural language inference (NLI) model. This is effective in preventing performance reduction, but at a cost of also discarding relevant passages. Thus, we propose a method for automatically generating data to fine-tune the language model to properly leverage retrieved passages, using a mix of relevant and irrelevant contexts at training time. We empirically show that even 1,000 examples suffice to train the model to be robust to irrelevant contexts while maintaining high performance on examples with relevant ones.
****
Takuya Furusawa
Mean Field Theory in Deep Metric Learning
In this paper, we explore the application of mean field theory, a technique from statistical physics, to deep metric learning and address the high training complexity commonly associated with conventional metric learning loss functions.
By adapting mean field theory for deep metric learning, we develop an approach to design classification-based loss functions from pair-based ones, which can be considered complementary to the proxy-based approach.
Applying the mean field theory to two pair-based loss functions, we derive two new loss functions, MeanFieldContrastive and MeanFieldClassWiseMultiSimilarity losses, with reduced training complexity.
We extensively evaluate these derived loss functions on three image-retrieval datasets and demonstrate that our loss functions outperform baseline methods in two out of the three datasets.
****
Weigao Sun,Zhen Qin,Weixuan Sun,Shidi Li,Dong Li,Xuyang Shen,Yu Qiao,Yiran Zhong
Efficient Distributed Training with Full Communication-Computation Overlap
The fundamental success of large language models hinges upon the efficacious implementation of large-scale distributed training techniques. Nevertheless, building a vast, high-performance cluster featuring high-speed communication interconn

ectivity is prohibitively costly, and accessible only to prominent entities. In this work, we aim to lower this barrier and democratize large-scale training wit h limited bandwidth clusters. We propose a new approach called CO2 that introduc es local-updating and asynchronous communication to the distributed data-paralle l training, thereby facilitating the full overlap of COmunication with COmputati on. CO2 is able to attain a remarkable 100\% scalability even on extensive multi -node clusters constrained by very limited communication bandwidth. We further p ropose the staleness gap penalty and outer momentum clipping techniques together with CO2 to bolster its convergence and training stability. Besides, CO2 exhibi ts seamless integration with well-established ZeRO-series optimizers which mitig ate memory consumption of model states with large model training. We also provid e a mathematical proof of convergence, accompanied by the establishment of a str ingent upper bound. Furthermore, we validate our findings through an extensive s et of practical experiments encompassing a wide range of tasks in the fields of computer vision and natural language processing. These experiments serve to demo nstrate the capabilities of CO2 in terms of convergence, generalization, and sca lability when deployed across configurations comprising up to 128 A100 GPUs. The outcomes emphasize the outstanding capacity of CO2 to achieve perfect 100\% sca lability, no matter on clusters with 800Gbps RDMA or 80Gbps TCP/IP inter-node co nnections.
****
Tianhao Wu,Chuanxia Zheng,Tat-Jen Cham
PanoDiffusion: 360-degree Panorama Outpainting via Diffusion
Generating complete 360\textdegree{} panoramas from narrow field of view images is ongoing research as omnidirectional RGB data is not readily available. Existi ng GAN-based approaches face some barriers to achieving higher quality output, a nd have poor generalization performance over different mask types. In this paper , we present our 360\textdegree{} indoor RGB panorama outpainting model using la tent diffusion models (LDM), called PanoDiffusion. We introduce a new bi-modal l atent diffusion structure that utilizes both RGB and depth panoramic data during training, which works surprisingly well to outpaint depth-free RGB images durin g inference. We further propose a novel technique of introducing progressive cam era rotations during each diffusion denoising step, which leads to substantial i mprovement in achieving panorama wraparound consistency. Results show that our P anoDiffusion not only significantly outperforms state-of-the-art methods on RGB panorama outpainting by producing diverse well-structured results for different types of masks, but can also synthesize high-quality depth panoramas to provide realistic 3D indoor models.
****
Ashutosh Singh,Ricardo Augusto Borsoi,Deniz Erdogmus,Tales Imbiriba
Learning semilinear neural operators: A unified recursive framework for predicti on and data assimilation.
Recent advances in the theory of Neural Operators (NOs) have enabled fast and ac curate computation of the solutions to complex systems described by partial diff erential equations (PDEs). Despite their great success, current NO-based solutio ns face important challenges when dealing with spatio-temporal PDEs over long ti me scales. Specifically, the current theory of NOs does not present a systematic framework to perform data assimilation and efficiently correct the evolution of PDE solutions over time based on sparsely sampled noisy measurements. In this p aper, we propose a learning-based state-space approach to compute the solution o perators to infinite-dimensional semilinear PDEs. Exploiting the structure of se milinear PDEs and the theory of nonlinear observers in function spaces, we devel op a flexible recursive method that allows for both prediction and data assimila tion by combining prediction and correction operations. The proposed framework i s capable of producing fast and accurate predictions over long time horizons, de aling with irregularly sampled noisy measurements to correct the solution, and b enefits from the decoupling between the spatial and temporal dynamics of this cl ass of PDEs. We show through experiments on the Kuramoto-Sivashinsky, Navier-Sto kes and Korteweg-de Vries equations that the proposed model is robust to noise a nd can leverage arbitrary amounts of measurements to correct its prediction over

a long time horizon with little computational overhead.
****
Puja Trivedi,Mark Heimann,Rushil Anirudh,Danai Koutra,Jayaraman J. Thiagarajan
Accurate and Scalable Estimation of Epistemic Uncertainty for Graph Neural Networks
While graph neural networks (GNNs) are widely used for node and graph representation learning tasks, the reliability of GNN uncertainty estimates under distribution shifts remains relatively under-explored. Indeed, while post-hoc calibration strategies can be used to improve in-distribution calibration, they need not also improve calibration under distribution shift. However, techniques which produce GNNs with better intrinsic uncertainty estimates are particularly valuable, as they can always be combined with post-hoc strategies later. Therefore, in this work, we propose G-ΔUQ, a novel training framework designed to improve intrinsic GNN uncertainty estimates. Our framework adapts the principle of stochastic data centering to graph data through novel graph anchoring strategies, and is able to support partially stochastic GNNs. While, the prevalent wisdom is that fully stochastic networks are necessary to obtain reliable estimates, we find that the functional diversity induced by our anchoring strategies when sampling hypotheses renders this unnecessary and allows us to support G-ΔUQ on pretrained models. Indeed, through extensive evaluation under covariate, concept and graph size shifts, we show that G-ΔUQ leads to better calibrated GNNs for node and graph classification. Further, it also improves performance on the uncertainty-based tasks of out-of-distribution detection and generalization gap estimation. Overall, our work provides insights into uncertainty estimation for GNNs, and demonstrates the utility of G-ΔUQ in obtaining reliable estimates.
****
Shiqiang Wang,Mingyue Ji
A Lightweight Method for Tackling Unknown Participation Statistics in Federated Averaging
In federated learning (FL), clients usually have diverse participation statistics that are unknown a priori, which can significantly harm the performance of FL if not handled properly. Existing works aiming at addressing this problem are usually based on global variance reduction, which requires a substantial amount of additional memory in a multiplicative factor equal to the total number of clients. An important open problem is to find a lightweight method for FL in the presence of clients with unknown participation rates. In this paper, we address this problem by adapting the aggregation weights in federated averaging (FedAvg) based on the participation history of each client. We first show that, with heterogeneous participation statistics, FedAvg with non-optimal aggregation weights can diverge from the optimal solution of the original FL objective, indicating the need of finding optimal aggregation weights. However, it is difficult to compute the optimal weights when the participation statistics are unknown. To address this problem, we present a new algorithm called FedAU, which improves FedAvg by adaptively weighting the client updates based on online estimates of the optimal weights without knowing the statistics of client participation. We provide a theoretical convergence analysis of FedAU using a novel methodology to connect the estimation error and convergence. Our theoretical results reveal important and interesting insights, while showing that FedAU converges to an optimal solution of the original objective and has desirable properties such as linear speedup. Our experimental results also verify the advantage of FedAU over baseline methods with various participation patterns.
****
LINHAO LUO,Yuan-Fang Li,Reza Haf,Shirui Pan
Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning
Large language models (LLMs) have demonstrated impressive reasoning abilities in complex tasks. However, they lack up-to-date knowledge and experience hallucinations during reasoning, which can lead to incorrect reasoning processes and diminish their performance and trustworthiness. Knowledge graphs (KGs), which capture vast amounts of facts in a structured format, offer a reliable source of knowl

edge for reasoning. Nevertheless, existing KG-based LLM reasoning methods only t
reat KGs as factual knowledge bases and overlook the importance of their structu
ral information for reasoning. In this paper, we propose a novel method called r
easoning on graphs (RoG) that synergizes LLMs with KGs to enable faithful and in
terpretable reasoning. Specifically, we present a planning-retrieval-reasoning f
ramework, where RoG first generates relation paths grounded by KGs as faithful p
lans. These plans are then used to retrieve valid reasoning paths from the KGs f
or LLMs to conduct faithful reasoning. Furthermore, RoG not only distills knowle
dge from KGs to improve the reasoning ability of LLMs through training but also
allows seamless integration with any arbitrary LLMs during inference. Extensive
experiments on two benchmark KGQA datasets demonstrate that RoG achieves state-o
f-the-art performance on KG reasoning tasks and generates faithful and interpret
able reasoning results.
****
Qingyan Guo,Rui Wang,Junliang Guo,Bei Li,Kaitao Song,Xu Tan,Guoqing Liu,Jiang Bi
an,Yujiu Yang
Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Pr
ompt Optimizers
Large Language Models (LLMs) excel in various tasks, but they rely on carefully
crafted prompts that often demand substantial human effort. To automate this pro
cess, in this paper, we propose a novel framework for discrete prompt optimizati
on, called EvoPrompt, which borrows the idea of evolutionary algorithms (EAs) as
 they exhibit good performance and fast convergence. To enable EAs to work on di
screte prompts, which are natural language expressions that need to be coherent
and human-readable, we connect LLMs with EAs. This approach allows us to simulta
neously leverage the powerful language processing capabilities of LLMs and the e
fficient optimization performance of EAs. Specifically, abstaining from any grad
ients or parameters, EvoPrompt starts from a population of prompts and iterative
ly generates new prompts with LLMs based on the evolutionary operators, improvin
g the population based on the development set. We optimize prompts for both clos
ed- and open-source LLMs including GPT-3.5 and Alpaca, on 31 datasets covering l
anguage understanding, generation tasks, as well as BIG-Bench Hard (BBH) tasks.
EvoPrompt significantly outperforms human-engineered prompts and existing method
s for automatic prompt generation (e.g., up to 25% on BBH). Furthermore, EvoProm
pt demonstrates that connecting LLMs with EAs creates synergies, which could ins
pire further research on the combination of LLMs and conventional algorithms.
****
Hsiao-Ru Pan,Bernhard Schölkopf
Skill or Luck? Return Decomposition via Advantage Functions
Learning from off-policy data is essential for sample-efficient reinforcement le
arning. In the present work, we build on the insight that the advantage function
 can be understood as the causal effect of an action on the return, and show tha
t this allows us to decompose the return of a trajectory into parts caused by th
e agent’s actions (skill) and parts outside of the agent’s control (luck). Furth
ermore, this decomposition enables us to naturally extend Direct Advantage Estim
ation (DAE) to off-policy settings (Off-policy DAE). The resulting method can le
arn
from off-policy trajectories without relying on importance sampling techniques o
r truncating off-policy actions. We draw connections between Off-policy DAE and
previous methods to demonstrate how it can speed up learning and when the propos
ed off-policy corrections are important. Finally, we use the MinAtar environment
s to illustrate how ignoring off-policy corrections can lead to suboptimal polic
y optimization performance.
****
Hanmin Li,Avetik Karagulyan,Peter Richtárik
Det-CGD: Compressed Gradient Descent with Matrix Stepsizes for Non-Convex Optimi
zation
This paper introduces a new method for minimizing matrix-smooth non-convex objec
tives through the use of novel Compressed Gradient Descent (CGD) algorithms enha
nced with a matrix-valued stepsize.

The proposed algorithms are theoretically analyzed first in the single-node and subsequently in the distributed settings. Our theoretical results reveal that the matrix stepsize in CGD can capture the objective's structure and lead to faster convergence compared to a scalar stepsize.
As a byproduct of our general results, we emphasize the importance of selecting the compression mechanism and the matrix stepsize in a layer-wise manner, taking advantage of model structure.
Moreover, we provide theoretical guarantees for free compression, by designing specific layer-wise compressors for the non-convex matrix smooth objectives. Our findings are supported with empirical evidence.
****
Ashutosh Baheti,Ximing Lu,Faeze Brahman,Ronan Le Bras,Maarten Sap,Mark Riedl
Leftover-Lunch: Advantage-based Offline Reinforcement Learning for Language Models
Reinforcement Learning with Human Feedback (RLHF) is the most prominent method for Language Model (LM) alignment. However, RLHF is an unstable and data-hungry process that continually requires new high-quality LM-generated data for finetuning. We introduce Advantage-Leftover Lunch RL (A-LoL), a new class of offline policy gradient algorithms that enable RL training on any pre-existing data. By assuming the entire LM output sequence as a single action, A-LoL allows incorporating sequence-level classifiers or human-designed scoring functions as
rewards. Subsequently, by using LM's value estimate, A-LoL only trains on positive advantage (leftover) data points, making it resilient to noise. Overall, A-LoL is an easy-to-implement, sample-efficient, and stable LM training recipe.

We demonstrate the effectiveness of A-LoL and its variants with a set of four different language generation tasks. We compare against both online RL (PPO) and recent preference-based (DPO, PRO) and reward-based (GOLD) offline RL baselines. On the commonly-used RLHF benchmark, Helpful and Harmless Assistant (HHA), LMs trained with A-LoL methods achieve the highest diversity while also being rated more safe and helpful than the baselines according to humans. Additionally, in the remaining three tasks, A-LoL could optimize multiple distinct reward functions even when using noisy or suboptimal training data.
****
Xiongye Xiao,Gengshuo Liu,Gaurav Gupta,Defu Cao,Shixuan Li,Yaxing Li,Tianqing Fang,Mingxi Cheng,Paul Bogdan
Neuro-Inspired Information-Theoretic Hierarchical Perception for Multimodal Learning
Integrating and processing information from various sources or modalities are critical for obtaining a comprehensive and accurate perception of the real world. Drawing inspiration from neuroscience, we develop the Information-Theoretic Hierarchical Perception (ITHP) model, which utilizes the concept of information bottleneck. Different from most traditional fusion models that incorporate all modalities identically in neural networks, our model designates a prime modality and regards the remaining modalities as detectors in the information pathway, serving to distill the flow of information. Our proposed perception model focuses on constructing an effective and compact information flow by achieving a balance between the minimization of mutual information between the latent state and the input modal state, and the maximization of mutual information between the latent states and the remaining modal states. This approach leads to compact latent state representations that retain relevant information while minimizing redundancy, thereby substantially enhancing the performance of multimodal representation learning. Experimental evaluations on the MUStARD, CMU-MOSI, and CMU-MOSEI datasets demonstrate that our model consistently distills crucial information in multimodal learning scenarios, outperforming state-of-the-art benchmarks. Remarkably, on the CMU-MOSI dataset, ITHP-DeBERTa surpasses human-level performance in the multimodal sentiment binary classification task across all evaluation metrics (i.e., Binary Accuracy, F1 Score, Mean Absolute Error, and Pearson Correlation).
****
Wei Wang,Dongqi Han,Xufang Luo,Dongsheng Li

Addressing Signal Delay in Deep Reinforcement Learning
Despite the notable advancements in deep reinforcement learning (DRL) in recent years, a prevalent issue that is often overlooked is the impact of signal delay. Signal delay occurs when there is a lag between an agent's perception of the environment and its corresponding actions. In this paper, we first formalize delayed-observation Markov decision processes (DOMDP) by extending the standard MDP framework to incorporate signal delays. Next, we elucidate the challenges posed by the presence of signal delay in DRL, showing that trivial DRL algorithms and generic methods for partially observable tasks suffer greatly from delays. Lastly, we propose effective strategies to overcome these challenges. Our methods achieve remarkable performance in continuous robotic control tasks with large delays, yielding results comparable to those in non-delayed cases. Overall, our work contributes to a deeper understanding of DRL in the presence of signal delays and introduces novel approaches to address the associated challenges.
****
Jonah Philion,Xue Bin Peng,Sanja Fidler
Trajeglish: Learning the Language of Driving Scenarios
A longstanding challenge for self-driving development is the ability to simulate dynamic driving scenarios seeded from recorded driving logs. Given an initial scene observed during a test drive, we seek the ability to sample plausible scene-consistent future trajectories for all agents in the scene, even when the actions for a subset of agents are chosen by an external source, such as a black-box self-driving planner. In order to model the complicated spatial and temporal interaction across agents in driving scenarios, we propose to tokenize the motion of dynamic agents and use tools from language modeling to model the full sequence of multi-agent actions. Our traffic model explicitly captures intra-timestep dependence between agents, which we show is essential for simulation given only a single frame of historical scene context, as well as enabling improvements when provided longer historical context. We demonstrate competitive results sampling scenarios given initializations from the Waymo Open Dataset with full autonomy as well as partial autonomy, and show that the representations learned by our model can quickly be adapted to improve performance on nuScenes, a considerably smaller dataset. We additionally use density estimates from our model to quantify the saliency of context length and intra-timestep interaction for the traffic modeling task.
****
Paul Hagemann,Johannes Hertrich,Fabian Altekrüger,Robert Beinert,Jannis Chemseddine,Gabriele Steidl
Posterior Sampling Based on Gradient Flows of the MMD with Negative Distance Kernel
We propose conditional flows of the maximum mean discrepancy (MMD) with the negative distance kernel for posterior sampling and conditional generative modelling. This MMD, which is also known as energy distance, has several advantageous properties like efficient computation via slicing and sorting. We approximate the joint distribution of the ground truth and the observations using discrete Wasserstein gradient flows and establish an error bound for the posterior distributions. Further, we prove that our particle flow is indeed a Wasserstein gradient flow of an appropriate functional. The power of our method is demonstrated by numerical examples including conditional image generation and inverse problems like superresolution, inpainting and computed tomography in low-dose and limited-angle settings.
****
Lijia Zhou,James B Simon,Gal Vardi,Nathan Srebro
An Agnostic View on the Cost of Overfitting in (Kernel) Ridge Regression
We study the cost of overfitting in noisy kernel ridge regression (KRR), which we define as the ratio between the test error of the interpolating ridgeless model and the test error of the optimally-tuned model. We take an ``agnostic'' view in the following sense: we consider the cost as a function of sample size for any target function, even if the sample size is not large enough for consistency or the target is outside the RKHS. We analyze the cost of overfitting under a Gau

ssian universality ansatz using recently derived (non-rigorous) risk estimates in terms of the task eigenstructure. Our analysis provides a more refined characterization of benign, tempered and catastrophic overfitting (cf. Mallinar et al. 2022).
****

Koichi Namekata,Amirmojtaba Sabour,Sanja Fidler,Seung Wook Kim
EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models
Diffusion models have recently received increasing research attention for their remarkable transfer abilities in semantic segmentation tasks. However, generating fine-grained segmentation masks with diffusion models often requires additional training on annotated datasets, leaving it unclear to what extent pre-trained diffusion models alone understand the semantic relations of their generated images. To address this question, we leverage the semantic knowledge extracted from Stable Diffusion (SD) and aim to develop an image segmentor capable of generating fine-grained segmentation maps without any additional training. The primary difficulty stems from the fact that semantically meaningful feature maps typically exist only in the spatially lower-dimensional layers, which poses a challenge in directly extracting pixel-level semantic relations from these feature maps. To overcome this issue, our framework identifies semantic correspondences between image pixels and spatial locations of low-dimensional feature maps by exploiting SD's generation process and utilizes them for constructing image-resolution segmentation maps. In extensive experiments, the produced segmentation maps are demonstrated to be well delineated and capture detailed parts of the images, indicating the existence of highly accurate pixel-level semantic knowledge in diffusion models.
Project page: https://kmcode1.github.io/Projects/EmerDiff/
****

Rohan Sharma,Kaiyi Ji,zhiqiang xu,Changyou Chen
AUC-CL: A Batchsize-Robust Framework for Self-Supervised Contrastive Representation Learning
Self-supervised learning through contrastive representations is an emergent and promising avenue, aiming at alleviating the availability of labeled data. Recent research in the field also demonstrates its viability for several downstream tasks, henceforth leading to works that implement the contrastive principle through innovative loss functions and methods. However, despite achieving impressive progress, most methods depend on prohibitively large batch sizes and compute requirements for good performance.
In this work, we propose the $\textbf{AUC}$-$\textbf{C}$ontrastive $\textbf{L}$earning, a new approach to contrastive learning that demonstrates robust and competitive performance in compute-limited regimes.
We propose to incorporate the contrastive objective within the AUC-maximization framework, by noting that the AUC metric is maximized upon enhancing the probability of the network's binary prediction difference between positive and negative samples which inspires adequate embedding space arrangements in representation learning. Unlike standard contrastive methods, when performing stochastic optimization, our method maintains unbiased stochastic gradients and thus is more robust to batchsizes as opposed to standard stochastic optimization problems.
Remarkably, our method with a batch size of 256, outperforms several state-of-the-art methods that may need much larger batch sizes (e.g., 4096), on ImageNet and other standard datasets. Experiments on transfer learning, few-shot learning, and other downstream tasks also demonstrate the viability of our method.
****

Shashanka Venkataramanan,Mamshad Nayeem Rizve,Joao Carreira,Yuki M Asano,Yannis Avrithis
Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video
Self-supervised learning has unlocked the potential of scaling up pretraining to billions of images, since annotation is unnecessary. But are we making the best use of data? How more economical can we be? In this work, we attempt to answer this question by making two contributions. First, we investigate first-person vi

deos and introduce a ``Walking Tours'' dataset. These videos are high-resolution, hours-long, captured in a single uninterrupted take, depicting a large number of objects and actions with natural scene transitions. They are unlabeled and uncurated, thus realistic for self-supervision and comparable with human learning.

Second, we introduce a novel self-supervised image pretraining method tailored for learning from continuous videos. Existing methods typically adapt image-based pretraining approaches to incorporate more frames. Instead, we advocate a ``tracking to learn to recognize'' approach. Our method called DoRA, leads to attention maps that **D**isc**O**ver and t**RA**ck objects over time in an end-to-end manner, using transformer cross-attention. We derive multiple views from the tracks and use them in a classical self-supervised distillation loss. Using our novel approach, a single Walking Tours video remarkably becomes a strong competitor to ImageNet for several image and video downstream tasks.
****
Scott Sussex,Pier Giuseppe Sessa,Anastasia Makarova,Andreas Krause
Adversarial Causal Bayesian Optimization
In Causal Bayesian Optimization (CBO), an agent intervenes on a structural causal model with known graph but unknown mechanisms to maximize a downstream reward variable. In this paper, we consider the generalization where other agents or external events also intervene on the system, which is key for enabling adaptiveness to non-stationarities such as weather changes, market forces, or adversaries. We formalize this generalization of CBO as Adversarial Causal Bayesian Optimization (ACBO) and introduce the first algorithm for ACBO with bounded regret: Causal Bayesian Optimization with Multiplicative Weights (CBO-MW). Our approach combines a classical online learning strategy with causal modeling of the rewards. To achieve this, it computes optimistic counterfactual reward estimates by propagating uncertainty through the causal graph. We derive regret bounds for CBO-MW that naturally depend on graph-related quantities. We further propose a scalable implementation for the case of combinatorial interventions and submodular rewards. Empirically, CBO-MW outperforms non-causal and non-adversarial Bayesian optimization methods on synthetic environments and environments based on real-word data. Our experiments include a realistic demonstration of how CBO-MW can be used to learn users' demand patterns in a shared mobility system and reposition vehicles in strategic areas.
****
Siqi Kou,Lei Gan,Dequan Wang,Chongxuan Li,Zhijie Deng
BayesDiff: Estimating Pixel-wise Uncertainty in Diffusion via Bayesian Inference
Diffusion models have impressive image generation capability, but low-quality generations still exist, and their identification remains challenging due to the lack of a proper sample-wise metric. To address this, we propose BayesDiff, a pixel-wise uncertainty estimator for generations from diffusion models based on Bayesian inference. In particular, we derive a novel uncertainty iteration principle to characterize the uncertainty dynamics in diffusion, and leverage the last-layer Laplace approximation for efficient Bayesian inference. The estimated pixel-wise uncertainty can not only be aggregated into a sample-wise metric to filter out low-fidelity images but also aids in augmenting successful generations and rectifying artifacts in failed generations in text-to-image tasks. Extensive experiments demonstrate the efficacy of BayesDiff and its promise for practical applications.
****
Roger Creus Castanyer,Joshua Romoff,Glen Berseth
Improving Intrinsic Exploration by Creating Stationary Objectives
Exploration bonuses in reinforcement learning guide long-horizon exploration by defining custom intrinsic objectives. Count-based methods use the frequency of state visits to derive an exploration bonus. In this paper, we identify that any intrinsic reward function derived from count-based methods is non-stationary and hence induces a difficult objective to optimize for the agent. The key contribution of our work lies in transforming the original non-stationary rewards into s

tationary rewards through an augmented state representation. For this purpose, w
e introduce the Stationary Objectives For Exploration (SOFE) framework. SOFE req
uires *identifying* sufficient statistics for different exploration bonuses and
finding an *efficient* encoding of these statistics to use as input to a deep ne
twork. SOFE is based on proposing state augmentations that expand the state spac
e but hold the promise of simplifying the optimization of the agent's objective.
 Our experiments show that SOFE improves the agents' performance in challenging
exploration problems, including sparse-reward tasks, pixel-based observations, 3
D navigation, and procedurally generated environments.
****
Yiyang Chen,Zhedong Zheng,Wei Ji,Leigang Qu,Tat-Seng Chua
Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regula
rization
We investigate composed image retrieval with text feedback. Users gradually look
 for the target of interest by moving from coarse to fine-grained feedback. How
ever, existing methods merely focus on the latter, i.e., fine-grained search, by
 harnessing positive and negative pairs during training. This pair-based paradig
m only considers the one-to-one distance between a pair of specific points, whic
h is not aligned with the one-to-many coarse-grained retrieval process and compr
omises the recall rate.
In an attempt to fill this gap, we introduce a unified learning approach to simu
ltaneously modeling the coarse- and fine-grained retrieval by considering the mu
lti-grained uncertainty.
The key idea underpinning the proposed method is to integrate fine- and coarse-g
rained retrieval as matching data points with small and large fluctuations, resp
ectively.
Specifically, our method contains two modules: uncertainty modeling and uncertai
nty regularization.
(1) The uncertainty modeling simulates the multi-grained queries by introducing
identically distributed fluctuations in the feature space.
(2) Based on the uncertainty modeling, we further introduce uncertainty regulari
zation to adapt the matching objective according to the fluctuation range.
Compared with existing methods, the proposed strategy explicitly prevents the mo
del from pushing away potential candidates in the early stage and thus improves
the recall rate. On the three public datasets, i.e., FashionIQ, Fashion200k, an
d Shoes, the proposed method has achieved +4.03%, + 3.38%, and + 2.40% Recall@50
 accuracy over a strong baseline, respectively.
****
Ziping Xu,Zifan Xu,Runxuan Jiang,Peter Stone,Ambuj Tewari
Sample Efficient Myopic Exploration Through Multitask Reinforcement Learning wit
h Diverse Tasks
Multitask Reinforcement Learning (MTRL) approaches have gained increasing attent
ion for its wide applications in many important Reinforcement Learning (RL) task
s. However, while recent advancements in MTRL theory have focused on the improve
d statistical efficiency by assuming a shared structure across tasks, exploratio
n--a crucial aspect of RL--has been largely overlooked. This paper addresses thi
s gap by showing that when an agent is trained on a sufficiently diverse set of
tasks, a generic policy-sharing algorithm with myopic exploration design like $
\epsilon$-greedy that are inefficient in general can be sample-efficient for MTR
L. To the best of our knowledge, this is the first theoretical demonstration of
the "exploration benefits" of MTRL. It may also shed light on the enigmatic succ
ess of the wide applications of myopic exploration in practice. To validate the
role of diversity, we conduct experiments on synthetic robotic control environme
nts, where the diverse task set aligns with the task selection by automatic curr
iculum learning, which is empirically shown to improve sample-efficiency.
****
Bingchen Zhao,Haoqin Tu,Chen Wei,Jieru Mei,Cihang Xie
Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning
This paper introduces an efficient strategy to transform Large Language Models (
LLMs) into Multi-Modal Large Language Models.

By conceptualizing this transformation as a domain adaptation process, \ie, tran sitioning from text understanding to embracing multiple modalities, we intriguin gly note that, within each attention block, tuning LayerNorm suffices to yield s trong performance.
Moreover, when benchmarked against other tuning approaches like full parameter f inetuning or LoRA, its benefits on efficiency are substantial.
For example, when compared to LoRA on a 13B model scale, performance can be enha nced by an average of over 20\% across five multi-modal tasks, and meanwhile, results in a significant reduction of trainable parameters by 41.9\% and a decre ase in GPU memory usage by 17.6\%. On top of this LayerNorm strategy, we showcas e that selectively tuning only with conversational data can improve efficiency f urther.
Beyond these empirical outcomes, we provide a comprehensive analysis to explore the role of LayerNorm in adapting LLMs to the multi-modal domain and improving t he expressive power of the model.
****
Jannik Kossen,Yarin Gal,Tom Rainforth
In-Context Learning Learns Label Relationships but Is Not Conventional Learning
The predictions of Large Language Models (LLMs) on downstream tasks often improv e significantly when including examples of the input–label relationship in the c ontext. However, there is currently no consensus about how this in-context learn ing (ICL) ability of LLMs works. For example, while Xie et al. (2022) liken ICL to a general-purpose learning algorithm, Min et al. (2022b) argue ICL does not e ven learn label relationships from in-context examples. In this paper, we provid e novel insights into how ICL leverages label information, revealing both capabi lities and limitations. To ensure we obtain a comprehensive picture of ICL behav ior, we study probabilistic aspects of ICL predictions and thoroughly examine th e dynamics of ICL as more examples are provided. Our experiments show that ICL p redictions almost always depend on in-context labels and that ICL can learn trul y novel tasks in-context. However, we also find that ICL struggles to fully over come prediction preferences acquired from pre-training data and, further, that I CL does not consider all in-context information equally.
****
Gabriel Della Maggiora,Luis Alberto Croquevielle,Nikita Deshpande,Harry Horsley, Thomas Heinis,Artur Yakimovich
Conditional Variational Diffusion Models
Inverse problems aim to determine parameters from observations, a crucial task i n engineering and science. Lately, generative models, especially diffusion model s, have gained popularity in this area for their ability to produce realistic so lutions and their good mathematical properties. Despite their success, an import ant drawback of diffusion models is their sensitivity to the choice of variance schedule, which controls the dynamics of the diffusion process. Fine-tuning this schedule for specific applications is crucial but time-consuming and does not g uarantee an optimal result. We propose a novel approach for learning the schedul e as part of the training process. Our method supports probabilistic conditionin g on data, provides high-quality solutions, and is flexible, proving able to ada pt to different applications with minimum overhead. This approach is tested in t wo unrelated inverse problems: super-resolution microscopy and quantitative phas e imaging, yielding comparable or superior results to previous methods and fine-tuned diffusion models. We conclude that fine-tuning the schedule by experimenta tion should be avoided because it can be learned during training in a stable way that yields better results. The code is available on https://github.com/casus/c vdm
****
Tianci Liu,Haoyu Wang,Feijie Wu,Hengtong Zhang,Pan Li,Lu Su,Jing Gao
Towards Poisoning Fair Representations
Fair machine learning seeks to mitigate model prediction bias against certain de mographic subgroups such as elder and female.
Recently, fair representation learning (FRL) trained by deep neural networks has demonstrated superior performance, whereby representations containing no demogr

aphic information are inferred from the data and then used as the input to class ification or other downstream tasks.
Despite the development of FRL methods, their vulnerability under data poisoning attack, a popular protocol to benchmark model robustness under adversarial scen arios, is under-explored. Data poisoning attacks have been developed for classic al fair machine learning methods which incorporate fairness constraints into sha llow-model classifiers.
Nonetheless, these attacks fall short in FRL due to notably different fairness g oals and model architectures.
This work proposes the first data poisoning framework attacking FRL. We induce t he model to output unfair representations that contain as much demographic infor mation as possible by injecting carefully crafted poisoning samples into the tra ining data.
This attack entails a prohibitive bilevel optimization, wherefore an effective a pproximated solution is proposed. A theoretical analysis on the needed number of poisoning samples is derived and sheds light on defending against the attack. E xperiments on benchmark fairness datasets and state-of-the-art fair representati on learning models demonstrate the superiority of our attack.
****
Juno Kim,Kakei Yamamoto,Kazusato Oko,Zhuoran Yang,Taiji Suzuki
Symmetric Mean-field Langevin Dynamics for Distributional Minimax Problems
In this paper, we extend mean-field Langevin dynamics to minimax optimization ov er probability distributions for the first time with symmetric and provably conv ergent updates. We propose \emph{mean-field Langevin averaged gradient} (MFL-AG) , a single-loop algorithm that implements gradient descent ascent in the distrib ution spaces with a novel weighted averaging, and establish average-iterate conv ergence to the mixed Nash equilibrium. We also study both time and particle disc retization regimes and prove a new uniform-in-time propagation of chaos result w hich accounts for the dependency of the particle interactions on all previous di stributions. Furthermore, we propose \emph{mean-field Langevin anchored best res ponse} (MFL-ABR), a symmetric double-loop algorithm based on best response dynam ics with linear last-iterate convergence. Finally, we study applications to zero -sum Markov games and conduct simulations demonstrating long-term optimality.
****
Victor Geadah,International Brain Laboratory,Jonathan W. Pillow
Parsing neural dynamics with infinite recurrent switching linear dynamical syste ms
Unsupervised methods for dimensionality reduction of neural activity and behavio r have provided unprecedented insights into the underpinnings of neural informat ion processing. One popular approach involves the recurrent switching linear dyn amical system (rSLDS) model, which describes the latent dynamics of neural spike train data using discrete switches between a finite number of low-dimensional l inear dynamical systems. However, a few properties of rSLDS model limit its depl oyability on trial-varying data, such as a fixed number of states over trials, a nd no latent structure or organization of states. Here we overcome these limitat ions by endowing the rSLDS model with a semi-Markov discrete state process, with latent geometry, that captures key properties of stochastic processes over part itions with flexible state cardinality. We leverage partial differential equatio ns (PDE) theory to derive an efficient, semi-parametric formulation for dynamica l sufficient statistics to the discrete states. This process, combined with swit ching dynamics, defines our infinite recurrent switching linear dynamical system (irSLDS) model class. We first validate and demonstrate the capabilities of our model on synthetic data. Next, we turn to the analysis of mice electrophysiolog ical data during decision-making, and uncover strong non-stationary processes un derlying both within-trial and trial-averaged neural activity.
****
Shurui Gui,Xiner Li,Shuiwang Ji
Active Test-Time Adaptation: Theoretical Analyses and An Algorithm
Test-time adaptation (TTA) addresses distribution shifts for streaming test data in unsupervised settings. Currently, most TTA methods can only deal with minor

shifts and rely heavily on heuristic and empirical studies.

To advance TTA under domain shifts, we propose the novel problem setting of active test-time adaptation (ATTA) that integrates active learning within the fully TTA setting.

We provide a learning theory analysis, demonstrating that incorporating limited labeled test instances enhances overall performances across test domains with a theoretical guarantee. We also present a sample entropy balancing for implementing ATTA while avoiding catastrophic forgetting (CF). We introduce a simple yet effective ATTA algorithm, known as SimATTA, using real-time sample selection techniques. Extensive experimental results confirm consistency with our theoretical analyses and show that the proposed ATTA method yields substantial performance improvements over TTA methods while maintaining efficiency and shares similar effectiveness to the more demanding active domain adaptation (ADA) methods. Our code is available at https://github.com/divelab/ATTA.

****

Defu Cao,Furong Jia,Sercan O Arik,Tomas Pfister,Yixiang Zheng,Wen Ye,Yan Liu

TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting

The past decade has witnessed significant advances in time series modeling with deep learning. While achieving state-of-the-art results, the best-performing architectures vary highly across applications and domains. On the other hand, for natural language processing, Generative Pre-trained Transformer (GPT) has demonstrated impressive performance via training one general-purpose model across various textual datasets. It is intriguing to explore whether GPT-type architectures can be effective for time series, capturing the intrinsic dynamic attributes and leading to significant accuracy improvements. In this paper, we propose a novel framework, TEMPO, that can effectively learn time series representations. We focus on utilizing two essential inductive biases of the time series task for pre-trained models: (i) decomposition of the complex interaction between trend, seasonal, and residual components; and (ii) introducing the selection-based prompts to facilitate distribution adaptation in non-stationary time series. TEMPO expands the capability for dynamically modeling real-world temporal phenomena from data within diverse domains. Our experiments demonstrate the superior performance of TEMPO, with 20%-60% improvement over state-of-the-art methods on a number of time series benchmark datasets. This performance gain is observed not only in standard supervised learning settings but also in scenarios involving previously unseen datasets. This compelling finding highlights TEMPO’s potential to constitute a foundational model building framework.

****

Zinan Lin,Sivakanth Gopi,Janardhan Kulkarni,Harsha Nori,Sergey Yekhanin

Differentially Private Synthetic Data via Foundation Model APIs 1: Images

Generating differentially private (DP) synthetic data that closely resembles the original private data is a scalable way to mitigate privacy concerns in the current data-driven world. In contrast to current practices that train customized models for this task, we aim to generate DP Synthetic Data via APIs (DPSDA), where we treat foundation models as blackboxes and only utilize their inference APIs. Such API-based, training-free approaches are easier to deploy as exemplified by the recent surge in the number of API-based apps. These approaches can also leverage the power of large foundation models which are only accessible via their inference APIs. However, this comes with greater challenges due to strictly more restrictive model access and the need to protect privacy from the API provider.

In this paper, we present a new framework called Private Evolution (PE) to solve this problem and show its initial promise on synthetic images. Surprisingly, PE can match or even outperform state-of-the-art (SOTA) methods without any model training. For example, on CIFAR10 (with ImageNet as the public data), we achieve FID ≤ 7.9 with privacy cost $\varepsilon = 0.67$, significantly improving the previous SOTA from $\varepsilon = 32$. We further demonstrate the promise of applying PE on large foundation models such as Stable Diffusion to tackle challenging private datasets with

a small number of high-resolution images. The code and data are released at http
s://github.com/microsoft/DPSDA.
****
Robert Peach,Matteo Vinao-Carl,Nir Grossman,Michael David,Emma Mallas,David J. S
harp,Paresh A. Malhotra,Pierre Vandergheynst,Adam Gosztolai
Implicit Gaussian process representation of vector fields over arbitrary latent
manifolds
Gaussian processes (GPs) are popular nonparametric statistical models for learni
ng unknown functions and quantifying the spatiotemporal uncertainty in data. Rec
ent works have extended GPs to model scalar and vector quantities distributed ov
er non-Euclidean domains, including smooth manifolds, appearing in numerous fiel
ds such as computer vision, dynamical systems, and neuroscience. However, these
approaches assume that the manifold underlying the data is known, limiting their
 practical utility. We introduce RVGP, a generalisation of GPs for learning vect
or signals over latent Riemannian manifolds. Our method uses positional encoding
 with eigenfunctions of the connection Laplacian, associated with the tangent bu
ndle, readily derived from common graph-based approximation of data. We demonstr
ate that RVGP possesses global regularity over the manifold, which allows it to
super-resolve and inpaint vector fields while preserving singularities. Furtherm
ore, we use RVGP to reconstruct high-density neural dynamics derived from low-de
nsity EEG recordings in healthy individuals and Alzheimer's patients. We show th
at vector field singularities are important disease markers and that their recon
struction leads to a comparable classification accuracy of disease states to hig
h-density recordings. Thus, our method overcomes a significant practical limitat
ion in experimental and clinical applications.
****
Kevin Black,Michael Janner,Yilun Du,Ilya Kostrikov,Sergey Levine
Training Diffusion Models with Reinforcement Learning
Diffusion models are a class of flexible generative models trained with an appro
ximation to the log-likelihood objective. However, most use cases of diffusion m
odels are not concerned with likelihoods, but instead with downstream objectives
 such as human-perceived image quality or drug effectiveness. In this paper, we
investigate reinforcement learning methods for directly optimizing diffusion mod
els for such objectives. We describe how posing denoising as a multi-step decisi
on-making problem enables a class of policy gradient algorithms, which we refer
to as denoising diffusion policy optimization (DDPO), that are more effective th
an alternative reward-weighted likelihood approaches. Empirically, DDPO is able
to adapt text-to-image diffusion models to objectives that are difficult to expr
ess via prompting, such as image compressibility, and those derived from human f
eedback, such as aesthetic quality. Finally, we show that DDPO can improve promp
t-image alignment using feedback from a vision-language model without the need f
or additional data collection or human annotation.
****
Antoine Bambade,Fabian Schramm,Adrien Taylor,Justin Carpentier
Leveraging augmented-Lagrangian techniques for differentiating over infeasible q
uadratic programs in machine learning
Optimization layers within neural network architectures have become increasingly
 popular for their ability to solve a wide range of machine learning tasks and t
o model domain-specific knowledge. However, designing optimization layers requir
es careful consideration as the underlying optimization problems might be infeas
ible during training.
Motivated by applications in learning, control and robotics, this work focuses o
n convex quadratic programming (QP) layers. The specific structure of this type
of optimization layer can be efficiently exploited for faster computations while
 still allowing rich modeling capabilities. We leverage primal-dual augmented La
grangian techniques for computing derivatives of both feasible and infeasible QP
 solutions.
More precisely, we propose a unified approach which tackles the differentiabilit
y of the closest feasible QP solutions in a classical ℓ_2 sense. We then ha
rness this approach to enrich the expressive capabilities of existing QP layers.

 More precisely, we show how differentiating through infeasible QPs during train
ing enables to drive towards feasibility at test time a new range of QP layers.
These layers notably demonstrate superior predictive performance in some convent
ional learning tasks. Additionally, we present alternative formulations that enh
ance numerical robustness, speed, and accuracy for training such layers.
Along with these contributions, we provide an open-source C++ software package c
alled QPLayer for differentiating feasible and infeasible convex QPs and which c
an be interfaced with modern learning frameworks.
****
Moonseok Choi,Hyungi Lee,Giung Nam,Juho Lee
Sparse Weight Averaging with Multiple Particles for Iterative Magnitude Pruning
Given the ever-increasing size of modern neural networks, the significance of sp
arse architectures has surged due to their accelerated inference speeds and mini
mal memory demands. When it comes to global pruning techniques, Iterative Magnit
ude Pruning (IMP) still stands as a state-of-the-art algorithm despite its simpl
e nature, particularly in extremely sparse regimes. In light of the recent findi
ng that the two successive matching IMP solutions are linearly connected without
 a loss barrier, we propose Sparse Weight Averaging with Multiple Particles (SWA
MP), a straightforward modification of IMP that achieves performance comparable
to an ensemble of two IMP solutions. For every iteration, we concurrently train
multiple sparse models, referred to as particles, using different batch orders y
et the same matching ticket, and then weight average such models to produce a si
ngle mask. We demonstrate that our method consistently outperforms existing base
lines across different sparsities through extensive experiments on various neura
l network structures and data.
****
Oscar Sainz,Iker García-Ferrero,Rodrigo Agerri,Oier Lopez de Lacalle,German Riga
u,Eneko Agirre
GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction
Large Language Models (LLMs) combined with instruction tuning have made signific
ant progress when generalizing to unseen tasks. However, they have been less suc
cessful in Information Extraction (IE), lagging behind task-specific models. Typ
ically, IE tasks are characterized by complex annotation guidelines which descri
be the task and give examples to humans. Previous attempts to leverage such info
rmation have failed, even with the largest models, as they are not able to follo
w the guidelines out-of-the-box. In this paper we propose GoLLIE (Guideline-foll
owing Large Language Model for IE), a model able to improve zero-shot results on
 unseen IE tasks by virtue of being fine-tuned to comply with annotation guideli
nes. Comprehensive evaluation empirically demonstrates that GoLLIE is able to ge
neralize to and follow unseen guidelines, outperforming previous attempts at zer
o-shot information extraction. The ablation study shows that detailed guidelines
 is key for good results. Code, data and models will be made publicly available.
****
Xiangxin Zhou,Xiwei Cheng,Yuwei Yang,Yu Bao,Liang Wang,Quanquan Gu
DecompOpt: Controllable and Decomposed Diffusion Models for Structure-based Mole
cular Optimization
Recently, 3D generative models have shown promising performances in structure-ba
sed drug design by learning to generate ligands given target binding sites. Howe
ver, only modeling the target-ligand distribution can hardly fulfill one of the
main goals in drug discovery -- designing novel ligands with desired properties,
 e.g., high binding affinity, easily synthesizable, etc. This challenge becomes
particularly pronounced when the target-ligand pairs used for training do not al
ign with these desired properties. Moreover, most existing methods aim at solvin
g de novo design task, while many generative scenarios requiring flexible contro
llability, such as R-group optimization and scaffold hopping, have received litt
le attention. In this work, we propose DecompOpt, a structure-based molecular op
timization method based on a controllable and decomposed diffusion model. Decomp
Opt presents a new generation paradigm which combines optimization with conditio
nal diffusion models to achieve desired properties while adhering to the molecul
ar grammar. Additionally, DecompOpt offers a unified framework covering both de

novo design and controllable generation. To achieve so, ligands are decomposed i
nto substructures which allows fine-grained control and local optimization. Expe
riments show that DecompOpt can efficiently generate molecules with improved pro
perties than strong de novo baselines, and demonstrate great potential in contro
llable generation tasks.
****
Meraj Hashemizadeh,Juan Ramirez,Rohan Sukumaran,Golnoosh Farnadi,Simon Lacoste-J
ulien,Jose Gallego-Posada
Balancing Act: Constraining Disparate Impact in Sparse Models
Model pruning is a popular approach to enable the deployment of large deep learn
ing models on edge devices with restricted computational or storage capacities.
Although sparse models achieve performance comparable to that of their dense cou
nterparts at the level of the entire dataset, they exhibit high accuracy drops f
or some data sub-groups. Existing methods to mitigate this disparate impact indu
ced by pruning (i) rely on surrogate metrics that address the problem indirectly
 and have limited interpretability; or (ii) scale poorly with the number of prot
ected sub-groups in terms of computational cost. We propose a constrained optimi
zation approach that _directly addresses the disparate impact of pruning_: our f
ormulation bounds the accuracy change between the dense and sparse models, for e
ach sub-group. This choice of constraints provides an interpretable success crit
erion to determine if a pruned model achieves acceptable disparity levels. Exper
imental results demonstrate that our technique scales reliably to problems invol
ving large models and hundreds of protected sub-groups.
****
Zequn Yang,Yake Wei,Ce Liang,Di Hu
Quantifying and Enhancing Multi-modal Robustness with Modality Preference
Multi-modal models have shown a promising capability to effectively integrate in
formation from various sources, yet meanwhile, they are found vulnerable to perv
asive perturbations, such as uni-modal attacks and missing conditions. To counte
r these perturbations, robust multi-modal representations are highly expected, w
hich are positioned well away from the discriminative multi-modal decision bound
ary. In this paper, different from conventional empirical studies, we focus on a
 commonly used joint multi-modal framework and theoretically discover that large
r uni-modal representation margins and more reliable integration for modalities
are essential components for achieving higher robustness. This discovery can fur
ther explain the limitation of multi-modal robustness and the phenomenon that mu
lti-modal models are often vulnerable to attacks on the specific modality. Moreo
ver, our analysis reveals how the widespread issue, that the model has different
 preferences for modalities, limits the multi-modal robustness by influencing th
e essential components and could lead to attacks on the specific modality highly
 effective. Inspired by our theoretical finding, we introduce a training procedu
re called Certifiable Robust Multi-modal Training (CRMT), which can alleviate th
is influence from modality preference and explicitly regulate essential componen
ts to significantly improve robustness in a certifiable manner. Our method demon
strates substantial improvements in performance and robustness compared with exi
sting methods. Furthermore, our training procedure can be easily extended to enh
ance other robust training strategies, highlighting its credibility and flexibil
ity.
****
Jaroslaw Blasiok,Preetum Nakkiran
Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing
Calibration measures and reliability diagrams are two fundamental tools for meas
uring and interpreting the calibration of probabilistic predictors. Calibration
measures quantify the degree of miscalibration, and reliability diagrams visuali
ze the structure of this miscalibration. However, the most common constructions
of reliability diagrams and calibration measures --- binning and ECE --- both su
ffer from well-known flaws (e.g. discontinuity). We show that a simple modificat
ion fixes both constructions: first smooth the observations using an RBF kernel,
 then compute the Expected Calibration Error (ECE) of this smoothed function. We
 prove that with a careful choice of bandwidth, this method yields a calibration

 measure that is well-behaved in the sense of (Blasiok, Gopalan, Hu, and Nakkira
n 2023) --- a consistent calibration measure. We call this measure the SmoothECE
. Moreover, the reliability diagram obtained from this smoothed function visuall
y encodes the SmoothECE, just as binned reliability diagrams encode the BinnedEC
E.   We also release a Python package with simple, hyperparameter-free methods f
or measuring and plotting calibration: "pip install relplot."
Code at: https://github.com/apple/ml-calibration
****
Wenxuan Zhang,Youssef Mohamed,Bernard Ghanem,Philip Torr,Adel Bibi,Mohamed Elhos
einy
Continual Learning on a Diet: Learning from Sparsely Labeled Streams Under Cons
trained Computation
We propose and study a realistic Continual Learning (CL) setting where learning
algorithms are granted a restricted computational budget per time step while tra
ining. We apply this setting to large-scale semi-supervised Continual Learning s
cenarios with sparse label rate. Previous proficient CL methods perform very po
orly in this challenging setting. Overfitting to the sparse labeled data and ins
ufficient computational budget are the two main culprits for such a poor perform
ance. Our new setting encourages learning methods to effectively and efficiently
 utilize the unlabeled data during training. To that end, we propose a simple bu
t highly effective baseline, DietCL, which utilizes both unlabeled and labeled d
ata jointly. DietCL meticulously allocates computational budget for both types o
f data. We validate our baseline, at scale, on several datasets, e.g., CLOC, Ima
geNet10K, and CGLM, under constraint budget setup. DietCL outperforms, by a larg
e margin, all existing supervised CL algorithms as well as more recent continual
 semi-supervised methods. Our extensive analysis and ablations demonstrate that
DietCL is stable under a full spectrum of label sparsity, computational budget a
nd various other ablations.
****
Kaifeng Lyu,Jikai Jin,Zhiyuan Li,Simon Shaolei Du,Jason D. Lee,Wei Hu
Dichotomy of Early and Late Phase Implicit Biases Can Provably Induce Grokking
Recent work by Power et al. (2022) highlighted a surprising "grokking" phenomeno
n in learning arithmetic tasks: a neural net first "memorizes" the training set,
 resulting in perfect training accuracy but near-random test accuracy, and after
 training for sufficiently longer, it suddenly transitions to perfect test accur
acy. This paper studies the grokking phenomenon in theoretical setups and shows
that it can be induced by a dichotomy of early and late phase implicit biases. S
pecifically, when training homogeneous neural nets with large initialization and
 small weight decay on both classification and regression tasks, we prove that t
he training process gets trapped at a solution corresponding to a kernel predict
or for a long time, and then a very sharp transition to min-norm/max-margin pred
ictors occurs, leading to a dramatic change in test accuracy. Even in the absenc
e of weight decay, we show that grokking can still happen when the late phase im
plicit bias is driven by other regularization mechanisms, such as implicit margi
n maximization or sharpness reduction.
****
Yizhou Jiang,Kunlin Hu,Tianren Zhang,Haichuan Gao,Yuqian Liu,Ying Fang,Feng Chen
Spatio-Temporal Approximation: A Training-Free SNN Conversion for Transformers
Spiking neural networks (SNNs) are energy-efficient and hold great potential for
 large-scale inference. Since training SNNs from scratch is costly and has limit
ed performance, converting pretrained artificial neural networks (ANNs) to SNNs
is an attractive approach that retains robust performance without additional tra
ining data and resources. However, while existing conversion methods work well o
n convolution networks, emerging Transformer models introduce unique mechanisms
like self-attention and test-time normalization, leading to non-causal non-linea
r interactions unachievable by current SNNs. To address this, we approximate the
se operations in both temporal and spatial dimensions, thereby providing the fir
st SNN conversion pipeline for Transformers. We propose \textit{Universal Group
Operators} to approximate non-linear operations spatially and a \textit{Temporal
-Corrective Self-Attention Layer} that approximates spike multiplications at inf

erence through an estimation-correction approach. Our algorithm is implemented on a pretrained ViT-B/32 from CLIP, inheriting its zero-shot classification capabilities, while improving control over conversion losses. To our knowledge, this is the first direct training-free conversion of a pretrained Transformer to a purely event-driven SNN, promising for neuromorphic hardware deployment.
****

Wufei Ma,Qihao Liu,Jiahao Wang,Angtian Wang,Xiaoding Yuan,Yi Zhang,Zihao Xiao,Guofeng Zhang,Beijia Lu,Ruxiao Duan,Yongrui Qi,Adam Kortylewski,Yaoyao Liu,Alan Yuille

Generating Images with 3D Annotations Using Diffusion Models

Diffusion models have emerged as a powerful generative method, capable of producing stunning photo-realistic images from natural language descriptions. However, these models lack explicit control over the 3D structure in the generated images. Consequently, this hinders our ability to obtain detailed 3D annotations for the generated images or to craft instances with specific poses and distances. In this paper, we propose 3D Diffusion Style Transfer (3D-DST), which incorporates 3D geometry control into diffusion models. Our method exploits ControlNet, which extends diffusion models by using visual prompts in addition to text prompts. We generate images of the 3D objects taken from 3D shape repositories~(e.g., ShapeNet and Objaverse), render them from a variety of poses and viewing directions, compute the edge maps of the rendered images, and use these edge maps as visual prompts to generate realistic images. With explicit 3D geometry control, we can easily change the 3D structures of the objects in the generated images and obtain ground-truth 3D annotations automatically. This allows us to improve a wide range of vision tasks, e.g., classification and 3D pose estimation, in both in-distribution (ID) and out-of-distribution (OOD) settings. We demonstrate the effectiveness of our method through extensive experiments on ImageNet-100/200, ImageNet-R, PASCAL3D+, ObjectNet3D, and OOD-CV. The results show that our method significantly outperforms existing methods, e.g., 3.8 percentage points on ImageNet-100 using DeiT-B. Our code is available at <https://ccvl.jhu.edu/3D-DST/>
****

Xianjun Yang,Wei Cheng,Yue Wu,Linda Ruth Petzold,William Yang Wang,Haifeng Chen

DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text

Large language models (LLMs) have notably enhanced the fluency and diversity of machine-generated text. However, this progress also presents a significant challenge in detecting the origin of a given text, and current research on detection methods lags behind the rapid evolution of LLMs. Conventional training-based methods have limitations in flexibility, particularly when adapting to new domains, and they often lack explanatory power. To address this gap, we propose a novel training-free detection strategy called Divergent N-Gram Analysis (DNA-GPT). Given a text, we first truncate it in the middle and then use only the preceding portion as input to the LLMs to regenerate the new remaining parts. By analyzing the differences between the original and new remaining parts through N-gram analysis in black-box or probability divergence in white-box, we can clearly illustrate significant discrepancies between machine-generated and human-written text. We conducted extensive experiments on the most advanced LLMs from OpenAI, including text-davinci-003, GPT-3.5-turbo, and GPT-4, as well as open-source models such as GPT-NeoX-20B and LLaMa-13B. Results show that our zero-shot approach exhibits state-of-the-art performance in distinguishing between human and GPT-generated text on four English and one German dataset, outperforming OpenAI's own classifier, which is trained on millions of text. Additionally, our methods provide reasonable explanations and evidence to support our claim, which is a unique feature of explainable detection. Our method is also robust under the revised text attack and can additionally solve model sourcing.
****

Ziteng Sun,Ananda Theertha Suresh,Aditya Krishna Menon

The importance of feature preprocessing for differentially private linear optimization

Training machine learning models with differential privacy (DP) has received inc

reasing interest in recent years. One of the most popular algorithms for trainin
g differentially private models is differentially private stochastic gradient de
scent (DPSGD) and its variants, where at each step gradients are clipped and com
bined with some noise. Given the increasing usage of DPSGD, we ask the question:
 is DPSGD alone sufficient to find a good minimizer for every dataset under priv
acy constraints?

As a first step towards answering this question, we show that even for the simpl
e case of linear classification, unlike non-private optimization, (private) feat
ure preprocessing is vital for differentially private optimization. In detail,
we first show theoretically that there exists an example where without feature p
reprocessing, DPSGD incurs a privacy error proportional to the maximum norm of f
eatures over all samples. We then propose an algorithm called *DPSGD-F*, which c
ombines DPSGD with feature preprocessing and prove that for classification tasks
, it incurs a privacy error proportional to the diameter of the features $\max_{
x, x' \in D} \|x - x'\|_2$. We then demonstrate the practicality of our algorith
m on image classification benchmarks.
****
Chongyi Zheng,Benjamin Eysenbach,Homer Rich Walke,Patrick Yin,Kuan Fang,Ruslan S
alakhutdinov,Sergey Levine
Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Da
ta
Robotic systems that rely primarily on self-supervised learning have the potenti
al to decrease the amount of human annotation and engineering effort required to
 learn control strategies. In the same way that prior robotic systems have lever
aged self-supervised techniques from computer vision (CV) and natural language p
rocessing (NLP), our work builds on prior work showing that the reinforcement le
arning (RL) itself can be cast as a self-supervised problem: learning to reach a
ny goal without human-specified rewards or labels. Despite the seeming appeal, l
ittle (if any) prior work has demonstrated how self-supervised RL methods can be
 practically deployed on robotic systems. By first studying a challenging simula
ted version of this task, we discover design decisions about architectures and h
yperparameters that increase the success rate by $2 \times$. These findings lay
the groundwork for our main result: we demonstrate that a self-supervised RL alg
orithm based on contrastive learning can solve real-world, image-based robotic m
anipulation tasks, with tasks being specified by a single goal image provided af
ter training.
****
Siyuan Li,Zedong Wang,Zicheng Liu,Cheng Tan,Haitao Lin,Di Wu,Zhiyuan Chen,Jiangb
in Zheng,Stan Z. Li
MogaNet: Multi-order Gated Aggregation Network
By contextualizing the kernel as global as possible, Modern ConvNets have shown
great potential in computer vision tasks. However, recent progress on \textit{mu
lti-order game-theoretic interaction} within deep neural networks (DNNs) reveals
 the representation bottleneck of modern ConvNets, where the expressive interact
ions have not been effectively encoded with the increased kernel size. To tackle
 this challenge, we propose a new family of modern ConvNets, dubbed MogaNet, for
 discriminative visual representation learning in pure ConvNet-based models with
 favorable complexity-performance trade-offs. MogaNet encapsulates conceptually
simple yet effective convolutions and gated aggregation into a compact module, w
here discriminative features are efficiently gathered and contextualized adaptiv
ely. MogaNet exhibits great scalability, impressive efficiency of parameters, an
d competitive performance compared to state-of-the-art ViTs and ConvNets on Imag
eNet and various downstream vision benchmarks, including COCO object detection,
ADE20K semantic segmentation, 2D\&3D human pose estimation, and video prediction
. Notably, MogaNet hits 80.0\% and 87.8\% accuracy with 5.2M and 181M parameters
 on ImageNet-1K, outperforming ParC-Net and ConvNeXt-L, while saving 59\% FLOPs
and 17M parameters, respectively. The source code is available at https://github
.com/Westlake-AI/MogaNet.
****

Brian DuSell,David Chiang
Stack Attention: Improving the Ability of Transformers to Model Hierarchical Pat
terns
Attention, specifically scaled dot-product attention, has proven effective for n
atural language, but it does not have a mechanism for handling hierarchical patt
erns of arbitrary nesting depth, which limits its ability to recognize certain s
yntactic structures. To address this shortcoming, we propose stack attention: an
 attention operator that incorporates stacks, inspired by their theoretical conn
ections to context-free languages (CFLs). We show that stack attention is analog
ous to standard attention, but with a latent model of syntax that requires no sy
ntactic supervision. We propose two variants: one related to deterministic pushd
own automata (PDAs) and one based on nondeterministic PDAs, which allows transfo
rmers to recognize arbitrary CFLs. We show that transformers with stack attentio
n are very effective at learning CFLs that standard transformers struggle on, ac
hieving strong results on a CFL with theoretically maximal parsing difficulty. W
e also show that stack attention is more effective at natural language modeling
under a constrained parameter budget, and we include results on machine translat
ion.
****
Gabriele Sarti,Grzegorz Chrupa■a,Malvina Nissim,Arianna Bisazza
Quantifying the Plausibility of Context Reliance in Neural Machine Translation
Establishing whether language models can use contextual information in a human-p
lausible way is important to ensure their safe adoption in real-world settings.
However, the questions of $\textit{when}$ and $\textit{which parts}$ of the cont
ext affect model generations are typically tackled separately, and current plaus
ibility evaluations are practically limited to a handful of artificial benchmark
s. To address this, we introduce $\textbf{P}$lausibility $\textbf{E}$valuation o
f $\textbf{Co}$ntext $\textbf{Re}$liance (PECoRe), an end-to-end interpretabilit
y framework designed to quantify context usage in language models' generations.
Our approach leverages model internals to (i) contrastively identify context-sen
sitive target tokens in generated texts and (ii) link them to contextual cues ju
stifying their prediction. We use PECoRe to quantify the plausibility of context
-aware machine translation models, comparing model rationales with human annotat
ions across several discourse-level phenomena. Finally, we apply our method to u
nannotated model translations to identify context-mediated predictions and highl
ight instances of (im)plausible context usage throughout generation.
****
Erik Schultheis,Wojciech Kotlowski,Marek Wydmuch,Rohit Babbar,Strom Borman,Krzys
ztof Dembczynski
Consistent algorithms for multi-label classification with macro-at-k metrics
We consider the optimization of complex performance metrics in multi-label class
ification under the population utility framework. We mainly focus on metrics lin
early decomposable into a sum of binary classification utilities applied separat
ely to each label with an additional requirement of exactly k labels predicted
 for each instance. These "macro-at-k" metrics possess desired properties for
extreme classification problems with long tail labels. Unfortunately, the at-k
 constraint couples the otherwise independent binary classification tasks, leadi
ng to a much more challenging optimization problem than standard macro-averages.
 We provide a statistical framework to study this problem, prove the existence a
nd the form of the optimal classifier, and propose a statistically consistent an
d practical learning algorithm based on the Frank-Wolfe method. Interestingly, o
ur main results concern even more general metrics being non-linear functions of
label-wise confusion matrices. Empirical results provide evidence for the compet
itive performance of the proposed approach.
****
Awni Altabaa,Taylor Whittington Webb,Jonathan D. Cohen,John Lafferty
Abstractors and relational cross-attention: An inductive bias for explicit relat
ional reasoning in Transformers
An extension of Transformers is proposed that enables explicit relational reason
ing through a novel module called the *Abstractor*. At the core of the Abstracto

r is a variant of attention called *relational cross-attention*. The approach is motivated by an architectural inductive bias for relational learning that disentangles relational information from object-level features. This enables explicit relational reasoning, supporting abstraction and generalization from limited data. The Abstractor is first evaluated on simple discriminative relational tasks and compared to existing relational architectures. Next, the Abstractor is evaluated on purely relational sequence-to-sequence tasks, where dramatic improvements are seen in sample efficiency compared to standard Transformers. Finally, Abstractors are evaluated on a collection of tasks based on mathematical problem solving, where consistent improvements in performance and sample efficiency are observed.
****
Mintong Kang,Nezihe Merve Gürel,Linyi Li,Bo Li
COLEP: Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits
Conformal prediction has shown spurring performance in constructing statistically rigorous prediction sets for arbitrary black-box machine learning models, assuming the data is exchangeable. However, even small adversarial perturbations during the inference can violate the exchangeability assumption, challenge the coverage guarantees, and result in a subsequent decline in empirical coverage. In this work, we propose a certifiably robust learning-reasoning conformal prediction framework (COLEP) via probabilistic circuits, which comprise a data-driven learning component that trains statistical models to learn different semantic concepts, and a reasoning component that encodes knowledge and characterizes the relationships among the trained models for logic reasoning. To achieve exact and efficient reasoning, we employ probabilistic circuits (PCs) within the reasoning component. Theoretically, we provide end-to-end certification of prediction coverage for COLEP in the presence of ℓ_2 bounded adversarial perturbations. We also provide certified coverage considering the finite size of the calibration set. Furthermore, we prove that COLEP achieves higher prediction coverage and accuracy over a single model as long as the utilities of knowledge models are non-trivial. Empirically, we show the validity and tightness of our certified coverage, demonstrating the robust conformal prediction of COLEP on various datasets, including GTSRB, CIFAR10, and AwA2. We show that COLEP achieves up to 12% improvement in certified coverage on GTSRB, 9% on CIFAR-10, and 14% on AwA2.
****
Alexander Robey,Fabian Latorre,George J. Pappas,Hamed Hassani,Volkan Cevher
Adversarial Training Should Be Cast as a Non-Zero-Sum Game
One prominent approach toward resolving the adversarial vulnerability of deep neural networks is the two-player zero-sum paradigm of adversarial training, in which predictors are trained against adversarially chosen perturbations of data. Despite the promise of this approach, algorithms based on this paradigm have not engendered sufficient levels of robustness and suffer from pathological behavior like robust overfitting. To understand this shortcoming, we first show that the commonly used surrogate-based relaxation used in adversarial training algorithms voids all guarantees on the robustness of trained classifiers. The identification of this pitfall informs a novel non-zero-sum bilevel formulation of adversarial training, wherein each player optimizes a different objective function. Our formulation yields a simple algorithmic framework that matches and in some cases outperforms state-of-the-art attacks, attains comparable levels of robustness to standard adversarial training algorithms, and does not suffer from robust overfitting.
****
Jianhui Li,Shilong Liu,Zidong Liu,Yikai Wang,Kaiwen Zheng,Jinghui Xu,Jianmin Li,Jun Zhu
InstructPix2NeRF: Instructed 3D Portrait Editing from a Single Image
With the success of Neural Radiance Field (NeRF) in 3D-aware portrait editing, a variety of works have achieved promising results regarding both quality and 3D consistency. However, these methods heavily rely on per-prompt optimization when handling natural language as editing instructions. Due to the lack of labeled h

uman face 3D datasets and effective architectures, the area of human-instructed
3D-aware editing for open-world portraits in an end-to-end manner remains under-
explored. To solve this problem, we propose an end-to-end diffusion-based framew
ork termed $\textbf{InstructPix2NeRF}$, which enables instructed 3D-aware portra
it editing from a single open-world image with human instructions. At its core
lies a conditional latent 3D diffusion process that lifts 2D editing to 3D space
 by learning the correlation between the paired images' difference and the instr
uctions via triplet data. With the help of our proposed token position randomiza
tion strategy, we could even achieve multi-semantic editing through one single p
ass with the portrait identity well-preserved. Besides, we further propose an id
entity consistency module that directly modulates the extracted identity signals
 into our diffusion process, which increases the multi-view 3D identity consiste
ncy. Extensive experiments verify the effectiveness of our method and show its s
uperiority against strong baselines quantitatively and qualitatively. Source cod
e and pretrained models can be found on our project page: https://mybabyyh.githu
b.io/InstructPix2NeRF.
****
Assaf Shocher,Amil V Dravid,Yossi Gandelsman,Inbar Mosseri,Michael Rubinstein,Al
exei A Efros
Idempotent Generative Network
We propose a new approach for generative modeling based on training a neural net
work to be idempotent. An idempotent operator is one that can be applied sequent
ially without changing the result beyond the initial application, namely $f(f(z)
)=f(z)$. The proposed model f is trained to map a source distribution (e.g, Ga
ussian noise) to a target distribution (e.g. realistic images) using the followi
ng objectives:
(1) Instances from the target distribution should map to themselves, namely $f(x
)=x$. We define the target manifold as the set of all instances that f maps to
 themselves.
(2) Instances that form the source distribution should map onto the defined targ
et manifold. This is achieved by optimizing the idempotence term, $f(f(z))=f(z)$
 which encourages the range of $f(z)$ to be on the target manifold. Under ideal
assumptions such a process provably converges to the target distribution. This s
trategy results in a model capable of generating an output in one step, maintain
ing a consistent latent space, while also allowing sequential applications for r
efinement. Additionally, we find that by processing inputs from both target and
source distributions, the model adeptly projects corrupted or modified data back
 to the target manifold. This work is a first step towards a ``global projector'
' that enables projecting any input into a target data distribution.
****
Tomasz Limisiewicz,David Mare■ek,Tomáš Musil
Debiasing Algorithm through Model Adaptation
Large language models are becoming the go-to solution for the ever-growing numbe
r of tasks.
However, with growing capacity, models are prone to rely on spurious correlation
s stemming from biases and stereotypes present in the training data.
This work proposes a novel method for detecting and mitigating gender bias in la
nguage models.
We perform causal analysis to identify problematic model components and discover
 that mid-upper feed-forward layers are most prone to convey bias.
Based on the analysis results, we intervene in the model by applying a linear pr
ojection to the weight matrices of these layers.
Our titular method DAMA, significantly decreases bias as measured by diverse met
rics while maintaining the model's performance on downstream tasks.
We release code for our method and models, which retrain LLaMA's state-of-the-ar
t performance while being significantly less biased.
****
Sascha Marton,Stefan Lüdtke,Christian Bartelt,Heiner Stuckenschmidt
GRANDE: Gradient-Based Decision Tree Ensembles for Tabular Data
Despite the success of deep learning for text and image data, tree-based ensembl

e models are still state-of-the-art for machine learning with heterogeneous tabular data. However, there is a significant need for tabular-specific gradient-based methods due to their high flexibility. In this paper, we propose GRANDE, GRAdie$\text{N}$t-Based Decision Tree Ensembles, a novel approach for learning hard, axis-aligned decision tree ensembles using end-to-end gradient descent. GRANDE is based on a dense representation of tree ensembles, which affords to use backpropagation with a straight-through operator to jointly optimize all model parameters. Our method combines axis-aligned splits, which is a useful inductive bias for tabular data, with the flexibility of gradient-based optimization. Furthermore, we introduce an advanced instance-wise weighting that facilitates learning representations for both, simple and complex relations, within a single model. We conducted an extensive evaluation on a predefined benchmark with 19 classification datasets and demonstrate that our method outperforms existing gradient-boosting and deep learning frameworks on most datasets. The method is available under: https://github.com/s-marton/GRANDE
****
Xianghao Kong,Ollie Liu,Han Li,Dani Yogatama,Greg Ver Steeg
Interpretable Diffusion via Information Decomposition
Denoising diffusion models enable conditional generation and density modeling of complex relationships like images and text.
However, the nature of the learned relationships is opaque making it difficult to understand precisely what relationships between words and parts of an image are captured, or to predict the effect of an intervention. We illuminate the fine-grained relationships learned by diffusion models by noticing a precise relationship between diffusion and information decomposition. Exact expressions for mutual information and conditional mutual information can be written in terms of the denoising model. Furthermore, ${pointwise}$ estimates can be easily estimated as well, allowing us to ask questions about the relationships between specific images and captions. Decomposing information even further to understand which variables in a high-dimensional space carry information is a long-standing problem. For diffusion models, we show that a natural non-negative decomposition of mutual information emerges, allowing us to quantify informative relationships between words and pixels in an image. We exploit these new relations to measure the compositional understanding of diffusion models, to do unsupervised localization of objects in images, and to measure effects when selectively editing images through prompt interventions.
****
Xu Han,Caihua Shan,Yifei Shen,Can Xu,Han Yang,Xiang Li,Dongsheng Li
Training-free Multi-objective Diffusion Model for 3D Molecule Generation
Searching for novel and diverse molecular candidates is a critical undertaking in drug and material discovery. Existing approaches have successfully adapted the diffusion model, the most effective generative model in image generation, to create 1D SMILES strings, 2D chemical graphs, or 3D molecular conformers. However, these methods are not efficient and flexible enough to generate 3D molecules with multiple desired properties, as they require additional training for the models for each new property or even a new combination of existing properties. Moreover, some properties may potentially conflict, making it impossible to find a molecule that satisfies all of them simultaneously. To address these challenges, we present a training-free conditional 3D molecular generation algorithm based on off-the-shelf unconditional diffusion models and property prediction models. The key techniques include modeling the loss of property prediction models as energy functions, considering the property relation between multiple conditions as a probabilistic graph, and developing a stable posterior estimation for computing the conditional score function. We conducted experiments on both single-objective and multi-objective 3D molecule generation, focusing on quantum properties, and compared our approach with the trained or fine-tuned diffusion models. Our proposed model achieves superior performance in generating molecules that meet the conditions, without any additional training cost.
****
Vishaal Udandarao,Max F Burg,Samuel Albanie,Matthias Bethge

Visual Data-Type Understanding does not emerge from scaling Vision-Language Mode
ls
Recent advances in the development of vision-language models (VLMs) are yielding
 remarkable success in recognizing visual semantic content, including impressive
 instances of compositional image understanding. Here, we introduce the novel ta
sk of Visual Data-Type Identification, a basic perceptual skill with implication
s for data curation (e.g., noisy data-removal from large datasets, domains pecif
ic retrieval) and autonomous vision (e.g., distinguishing changing weather condi
tions from camera lens staining). We develop two datasets consisting of animal i
mages altered across a diverse set of 27 visual data-types, spanning four broad
categories. An extensive zero-shot evaluation of 39 VLMs, ranging from 100M to 8
0B parameters, shows a nuanced performance landscape. While VLMs are reasonably
good at identifying certain stylistic data-types, such as cartoons and sketches,
 they struggle with simpler data-types arising from basic manipulations like ima
ge rotations or additive noise. Our findings reveal that (i) model scaling alone
 yields marginal gains for contrastively-trained models like CLIP, and (ii) ther
e is a pronounced drop in performance for the largest auto-regressively trained
VLMs like OpenFlamingo. This finding points to a blind spot in current frontier
VLMs: they excel in recognizing semantic content but fail to acquire an
understanding of visual data-types through scaling. By analyzing the pre-trainin
g distributions of these models and incorporating data-type information into the
 captions during fine-tuning, we achieve a significant enhancement in performanc
e. By exploring this previously uncharted task, we aim to set the stage for furt
her advancing VLMs to equip them with visual data-type understanding.
****
Soichiro Kumano,Hiroshi Kera,Toshihiko Yamasaki
Theoretical Understanding of Learning from Adversarial Perturbations
It is not fully understood why adversarial examples can deceive neural networks
and transfer between different networks. To elucidate this, several studies have
 hypothesized that adversarial perturbations, while appearing as noises, contain
 class features. This is supported by empirical evidence showing that networks t
rained on mislabeled adversarial examples can still generalize well to correctly
 labeled test samples. However, a theoretical understanding of how perturbations
 include class features and contribute to generalization is limited. In this stu
dy, we provide a theoretical framework for understanding learning from perturbat
ions using a one-hidden-layer network trained on mutually orthogonal samples. Ou
r results highlight that various adversarial perturbations, even perturbations o
f a few pixels, contain sufficient class features for generalization. Moreover,
we reveal that the decision boundary when learning from perturbations matches th
at from standard samples except for specific regions under mild conditions. The
code is available at https://github.com/s-kumano/learning-from-adversarial-pertu
rbations.
****
Yuhui Xu,Lingxi Xie,Xiaotao Gu,Xin Chen,Heng Chang,Hengheng Zhang,Zhengsu Chen,X
IAOPENG ZHANG,Qi Tian
QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models
Recently years have witnessed a rapid development of large language models (LLMs
). Despite the strong ability in many language-understanding tasks, the heavy co
mputational burden largely restricts the application of LLMs especially when one
 needs to deploy them onto edge devices. In this paper, we propose a quantizatio
n-aware low-rank adaptation (QA-LoRA) algorithm. The motivation lies in the imba
lanced degrees of freedom of quantization and adaptation, and the solution is to
 use group-wise operators which increase the degree of freedom of quantization m
eanwhile decreasing that of adaptation. QA-LoRA is easily implemented with a few
 lines of code, and it equips the original LoRA with two-fold abilities: (i) dur
ing fine-tuning, the LLM's weights are quantized (e.g., into INT4) to reduce tim
e and memory usage; (ii) after fine-tuning, the LLM and auxiliary weights are na
turally integrated into a quantized model without loss of accuracy. We apply QA-
LoRA to the LLaMA and LLaMA2 model families and validate its effectiveness in di
fferent fine-tuning datasets and downstream scenarios. The code is made availabl

e at https://github.com/yuhuixu1993/qa-lora.
****
Hao Liu,Matei Zaharia,Pieter Abbeel
RingAttention with Blockwise Transformers for Near-Infinite Context
Transformers have emerged as the architecture of choice for many state-of-the-ar t AI models, showcasing exceptional performance across a wide range of AI applic ations. However, the memory demands imposed by Transformers limit their ability to handle long sequences, thereby posing challenges in utilizing videos, actions , and other long-form sequences and modalities in complex environments. We prese nt a novel approach, Blockwise RingAttention, which leverages blockwise computat ion of self-attention and feedforward to distribute long sequences across multip le devices while fully overlapping the communication of key-value blocks with th e computation of blockwise attention. Our approach enables training and inferenc e of sequences that are up to device count times longer than those achievable by prior memory-efficient Transformers, without resorting to approximations or inc urring additional communication and computation overheads. Extensive experiments on language modeling and reinforcement learning tasks demonstrate the effective ness of our approach in allowing millions of tokens context size and improving p erformance.
****
Saptarshi Chakraborty,Peter Bartlett
A Statistical Analysis of Wasserstein Autoencoders for Intrinsically Low-dimensi onal Data
Variational Autoencoders (VAEs) have gained significant popularity among researc hers as a powerful tool for understanding unknown distributions based on limited samples. This popularity stems partly from their impressive performance and par tly from their ability to provide meaningful feature representations in the late nt space. Wasserstein Autoencoders (WAEs), a variant of VAEs, aim to not only im prove model efficiency but also interpretability. However, there has been limite d focus on analyzing their statistical guarantees. The matter is further complic ated by the fact that the data distributions to which WAEs are applied - such as natural images - are often presumed to possess an underlying low-dimensional st ructure within a high-dimensional feature space, which current theory does not a dequately account for, rendering known bounds inefficient. To bridge the gap bet ween the theory and practice of WAEs, in this paper, we show that WAEs can learn the data distributions when the network architectures are properly chosen. We s how that the convergence rates of the expected excess risk in the number of samp les for WAEs are independent of the high feature dimension, instead relying only on the intrinsic dimension of the data distribution.
****
Kibum Kim,Kanghoon Yoon,Yeonjun In,Jinyoung Moon,Donghyun Kim,Chanyoung Park
Adaptive Self-training Framework for Fine-grained Scene Graph Generation
Scene graph generation (SGG) models have suffered from inherent problems regardi ng the benchmark datasets such as the long-tailed predicate distribution and mis sing annotation problems. In this work, we aim to alleviate the long-tailed prob lem of SGG by utilizing unannotated triplets. To this end, we introduce a **S**e lf-**T**raining framework for **SGG** **(ST-SGG)** that assigns pseudo-labels fo r unannotated triplets based on which the SGG models are trained. While there ha s been significant progress in self-training for image recognition, designing a self-training framework for the SGG task is more challenging due to its inherent nature such as the semantic ambiguity and the long-tailed distribution of predi cate classes. Hence, we propose a novel pseudo-labeling technique for SGG, calle d **C**lass-specific **A**daptive **T**hresholding with **M**omentum **(CATM)**, which is a model-agnostic framework that can be applied to any existing SGG mod els. Furthermore, we devise a graph structure learner (GSL) that is beneficial w hen adopting our proposed self-training framework to the state-of-the-art messag e-passing neural network (MPNN)-based SGG models. Our extensive experiments veri fy the effectiveness of ST-SGG on various SGG models, particularly in enhancing the performance on fine-grained predicate classes.
****

Zeyu Yang,Hongye Yang,Zijie Pan,Li Zhang
Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting
Reconstructing dynamic 3D scenes from 2D images and generating diverse views over time is challenging due to scene complexity and temporal dynamics. Despite advancements in neural implicit models, limitations persist: (i) Inadequate Scene Structure: Existing methods struggle to reveal the spatial and temporal structure of dynamic scenes from directly learning the complex 6D plenoptic function. (ii) Scaling Deformation Modeling: Explicitly modeling scene element deformation becomes impractical for complex dynamics. To address these issues, we consider the spacetime as an entirety and propose to approximate the underlying spatio-temporal 4D volume of a dynamic scene by optimizing a collection of 4D primitives, with explicit geometry and appearance modeling. Learning to optimize the 4D primitives enables us to synthesize novel views at any desired time with our tailored rendering routine. Our model is conceptually simple, consisting of a 4D Gaussian parameterized by anisotropic ellipses that can rotate arbitrarily in space and time, as well as view-dependent and time-evolved appearance represented by the coefficient of 4D spherindrical harmonics. This approach offers simplicity, flexibility for variable-length video and end-to-end training, and efficient real-time rendering, making it suitable for capturing complex dynamic scene motions. Experiments across various benchmarks, including monocular and multi-view scenarios, demonstrate our 4DGS model's superior visual quality and efficiency.
****
Alexander Korotin,Nikita Gushchin,Evgeny Burnaev
Light Schrödinger Bridge
Despite the recent advances in the field of computational Schrödinger Bridges (SB), most existing SB solvers are still heavy-weighted and require complex optimization of several neural networks. It turns out that there is no principal solver which plays the role of simple-yet-effective baseline for SB just like, e.g., k-means method in clustering, logistic regression in classification or Sinkhorn algorithm in discrete optimal transport. We address this issue and propose a novel fast and simple SB solver. Our development is a smart combination of two ideas which recently appeared in the field: (a) parameterization of the Schrödinger potentials with sum-exp quadratic functions and (b) viewing the log-Schrödinger potentials as the energy functions. We show that combined together these ideas yield a lightweight, simulation-free and theoretically justified SB solver with a simple straightforward optimization objective. As a result, it allows solving SB in moderate dimensions in a matter of minutes on CPU without a painful hyperparameter selection. Our light solver resembles the Gaussian mixture model which is widely used for density estimation. Inspired by this similarity, we also prove an important theoretical result showing that our light solver is a universal approximator of SBs. Furthemore, we conduct the analysis of the generalization error of our light solver. The code for our solver can be found at https://github.com/ngushchin/LightSB.
****
Yifu Yuan,Jianye HAO,Yi Ma,Zibin Dong,Hebin Liang,Jinyi Liu,Zhixin Feng,Kai Zhao,YAN ZHENG
Uni-RLHF: Universal Platform and Benchmark Suite for Reinforcement Learning with Diverse Human Feedback
Reinforcement Learning with Human Feedback (RLHF) has received significant attention for performing tasks without the need for costly manual reward design by aligning human preferences. It is crucial to consider diverse human feedback types and various learning methods in different environments. However, quantifying progress in RLHF with diverse feedback is challenging due to the lack of standardized annotation platforms and widely used unified benchmarks. To bridge this gap, we introduce **Uni-RLHF**, a comprehensive system implementation tailored for RLHF. It aims to provide a complete workflow from *real human feedback*, fostering progress in the development of practical problems. Uni-RLHF contains three packages: 1) a universal multi-feedback annotation platform, 2) large-scale crowdsourced feedback datasets, and 3) modular offline RLHF baseline implementations. U

ni-RLHF develops a user-friendly annotation interface tailored to various feedba
ck types, compatible with a wide range of mainstream RL environments. We then es
tablish a systematic pipeline of crowdsourced annotations, resulting in large-sc
ale annotated datasets comprising more than 15 million steps across 30 popular t
asks. Through extensive experiments, the results in the collected datasets demon
strate competitive performance compared to those from well-designed manual rewar
ds. We evaluate various design choices and offer insights into their strengths a
nd potential areas of improvement. We wish to build valuable open-source platfor
ms, datasets, and baselines to facilitate the development of more robust and rel
iable RLHF solutions based on realistic human feedback. The website is available
 at https://uni-rlhf.github.io/.
****
Khai Nguyen,Nicola Bariletto,Nhat Ho
Quasi-Monte Carlo for 3D Sliced Wasserstein
Monte Carlo (MC) integration has been employed as the standard approximation met
hod for the Sliced Wasserstein (SW) distance, whose analytical expression involv
es an intractable expectation. However, MC integration is not optimal in terms o
f absolute approximation error. To provide a better class of empirical SW, we pr
opose quasi-sliced Wasserstein (QSW) approximations that rely on Quasi-Monte Car
lo (QMC) methods. For a comprehensive investigation of QMC for SW, we focus on t
he 3D setting, specifically computing the SW between probability measures in thr
ee dimensions. In greater detail, we empirically evaluate various methods to con
struct QMC point sets on the 3D unit-hypersphere, including the Gaussian-based a
nd equal area mappings, generalized spiral points, and optimizing discrepancy en
ergies. Furthermore, to obtain an unbiased estimator for stochastic optimization
, we extend QSW to Randomized Quasi-Sliced Wasserstein (RQSW) by introducing ran
domness in the discussed point sets. Theoretically, we prove the asymptotic conv
ergence of QSW and the unbiasedness of RQSW. Finally, we conduct experiments on
various 3D tasks, such as point-cloud comparison, point-cloud interpolation, ima
ge style transfer, and training deep point-cloud autoencoders, to demonstrate th
e favorable performance of the proposed QSW and RQSW variants.
****
Yeongyeon Na,Minje Park,Yunwon Tae,Sunghoon Joo
Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship o
f Electrocardiogram
Electrocardiograms (ECG) are widely employed as a diagnostic tool for monitoring
 electrical signals originating from a heart. Recent machine learning research e
fforts have focused on the application of screening various diseases using ECG s
ignals. However, adapting to the application of screening disease is challenging
 in that labeled ECG data are limited. Achieving general representation through
self-supervised learning (SSL) is a well-known approach to overcome the scarcity
 of labeled data; however, a naive application of SSL to ECG data, without consi
dering the spatial-temporal relationships inherent in ECG signals, may yield sub
optimal results. In this paper, we introduce ST-MEM (Spatio-Temporal Masked Elec
trocardiogram Modeling), designed to learn spatio-temporal features by reconstru
cting masked 12-lead ECG data. ST-MEM outperforms other SSL baseline methods in
various experimental settings for arrhythmia classification tasks. Moreover, we
demonstrate that ST-MEM is adaptable to various lead combinations. Through quant
itative and qualitative analysis, we show a spatio-temporal relationship within
ECG data. Our code is available at https://github.com/bakqui/ST-MEM.
****
Shangbin Feng,Weijia Shi,Yuyang Bai,Vidhisha Balachandran,Tianxing He,Yulia Tsve
tkov
Knowledge Card: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language M
odels
By design, large language models (LLMs) are static general-purpose models, expen
sive to retrain or update frequently. As they are increasingly adopted for knowl
edge-intensive tasks, it becomes evident that these design choices lead to failu
res to generate factual, relevant, and up-to-date knowledge. To this end, we pro
pose Knowledge Card, a modular framework to plug in new factual and relevant kno

wledge into general-purpose LLMs. We first introduce knowledge cards---specialized language models trained on corpora from specific domains and sources. Knowledge cards serve as parametric repositories that are selected at inference time to generate background knowledge for the base LLM. We then propose three content selectors to dynamically select and retain information in documents generated by knowledge cards, specifically controlling for relevance, brevity, and factuality of outputs. Finally, we propose two complementary integration approaches to augment the base LLM with the (relevant, factual) knowledge curated from the specialized LMs. Through extensive experiments, we demonstrate that Knowledge Card achieves state-of-the-art performance on six benchmark datasets. Ultimately, Knowledge Card framework enables dynamic synthesis and updates of knowledge from diverse domains. Its modularity will ensure that relevant knowledge can be continuously updated through the collective efforts of the research community.
****
Xiangming Zhu,Huayu Deng,Haochen Yuan,Yunbo Wang,Xiaokang Yang
Latent Intuitive Physics: Learning to Transfer Hidden Physics from A 3D Video
We introduce latent intuitive physics, a transfer learning framework for physics simulation that can infer hidden properties of fluids from a single 3D video and simulate the observed fluid in novel scenes. Our key insight is to use latent features drawn from a learnable prior distribution conditioned on the underlying particle states to capture the invisible and complex physical properties. To achieve this, we train a parametrized prior learner given visual observations to approximate the visual posterior of inverse graphics, and both the particle states and the visual posterior are obtained from a learned neural renderer. The converged prior learner is embedded in our probabilistic physics engine, allowing us to perform novel simulations on unseen geometries, boundaries, and dynamics without knowledge of the true physical parameters. We validate our model in three ways: (i) novel scene simulation with the learned visual-world physics, (ii) future prediction of the observed fluid dynamics, and (iii) supervised particle simulation. Our model demonstrates strong performance in all three tasks.
****
Tim De Ryck,Florent Bonnet,Siddhartha Mishra,Emmanuel de Bezenac
An operator preconditioning perspective on training in physics-informed machine learning
In this paper, we investigate the behavior of gradient descent algorithms in physics-informed machine learning methods like PINNs, which minimize residuals connected to partial differential equations (PDEs). Our key result is that the difficulty in training these models is closely related to the conditioning of a specific differential operator. This operator, in turn, is associated to the Hermitian square of the differential operator of the underlying PDE. If this operator is ill-conditioned, it results in slow or infeasible training. Therefore, preconditioning this operator is crucial. We employ both rigorous mathematical analysis and empirical evaluations to investigate various strategies, explaining how they better condition this critical operator, and consequently improve training.
****
Seunghan Lee,Taeyoung Park,Kibok Lee
Learning to Embed Time Series Patches Independently
Masked time series modeling has recently gained much attention as a self-supervised representation learning strategy for time series.
Inspired by masked image modeling in computer vision, recent works first patchify and partially mask out time series, and then train Transformers to capture the dependencies between patches by predicting masked patches from unmasked patches.
However, we argue that capturing such patch dependencies might not be an optimal strategy for time series representation learning;
rather, learning to embed patches independently results in better time series representations.
Specifically, we propose to use 1) the simple patch reconstruction task, which autoencode each patch without looking at other patches, and 2) the simple patch-wise MLP that embeds each patch independently.

In addition, we introduce complementary contrastive learning to hierarchically c
apture adjacent time series information efficiently.
Our proposed method improves time series forecasting and classification performa
nce compared to state-of-the-art Transformer-based models, while it is more effi
cient in terms of the number of parameters and training time.
Code is available at this repository: https://github.com/seunghan96/pits.
****
Orren Karniol-Tambour,David M. Zoltowski,E. Mika Diamanti,Lucas Pinto,Carlos D B
rody,David W. Tank,Jonathan W. Pillow
Modeling state-dependent communication between brain regions with switching nonl
inear dynamical systems
Understanding how multiple brain regions interact to produce behavior is a major
 challenge in systems neuroscience, with many regions causally implicated in com
mon tasks such as sensory processing and decision making. A precise description
of interactions between regions remains an open problem. Moreover, neural dynami
cs are nonlinear and non-stationary. Here, we propose MR-SDS, a multiregion, swi
tching nonlinear state space model that decomposes global dynamics into local an
d cross-communication components in the latent space. MR-SDS includes directed i
nteractions between brain regions, allowing for estimation of state-dependent co
mmunication signals, and accounts for sensory inputs effects, history effects, a
nd heterogeneity across days and animals. We show that our model accurately reco
vers latent trajectories, vector fields underlying switching nonlinear dynamics,
 and cross-region communication profiles in three simulations. We then apply our
 method to two large-scale, multi-region neural datasets involving mouse decisio
n making. The first includes hundreds of neurons per region, recorded simultaneo
usly at single-cell-resolution across 3 distant cortical regions. The second is
a mesoscale widefield dataset of 8 adjacent cortical regions imaged across both
hemispheres. On these multi-region datasets, our model outperforms existing piec
e-wise linear multi-region models and reveals multiple distinct dynamical states
 and a rich set of cross-region communication profiles.
****
Congpei Qiu,Tong Zhang,Yanhao Wu,Wei Ke,Mathieu Salzmann,Sabine Süsstrunk
Mind Your Augmentation: The Key to Decoupling Dense Self-Supervised Learning
Dense Self-Supervised Learning (SSL) creates positive pairs by building positive
 paired regions or points, thereby aiming to preserve local features, for exampl
e of individual objects. However, existing approaches tend to couple objects by
leaking information from the neighboring contextual regions when the pairs have
a limited overlap. In this paper, we first quantitatively identify and confirm t
he existence of such a coupling phenomenon. We then address it by developing a r
emarkably simple yet highly effective solution comprising a novel augmentation m
ethod, Region Collaborative Cutout (RCC), and a corresponding decoupling branch.
 Importantly, our design is versatile and can be seamlessly integrated into exis
ting SSL frameworks, whether based on Convolutional Neural Networks (CNNs) or Vi
sion Transformers (ViTs). We conduct extensive experiments, incorporating our so
lution into two CNN-based and two ViT-based methods, with results confirming the
 effectiveness of our approach. Moreover, we provide empirical evidence that our
 method significantly contributes to the disentanglement of feature representati
ons among objects, both in quantitative and qualitative terms.
****
Katherine Tian,Eric Mitchell,Huaxiu Yao,Christopher D Manning,Chelsea Finn
Fine-Tuning Language Models for Factuality
The fluency and creativity of large pre-trained language models (LLMs) have led
to their widespread use, sometimes even as a replacement for traditional search
engines. However, language models are prone to making convincing but factually i
naccurate claims, often referred to as 'hallucinations', which can harmfully per
petuate myths and misconceptions. Further, manual fact-checking of model respons
es is a time-consuming process, making human factuality labels expensive to acqu
ire. In this work, we leverage two key recent innovations in NLP to fine-tune la
nguage models to be more factual without human labeling, targeting more open-end
ed generation settings than past work. First, several recent works have proposed

methods for scoring the factuality of open-ended text derived from consistency with an external knowledge base or simply a large model's confidence scores. Second, the Direct Preference Optimization algorithm enables straightforward fine-tuning of language models on objectives other than supervised imitation, using a preference ranking over possible model responses. We show that learning from preference rankings generated by either automated criterion significantly improves the factuality of Llama-2 on held-out topics (percent of generated claims that are correct) compared with existing RLHF procedures or decoding strategies targeted at factuality, showing over 50% and 20-30% error reduction for biographies and medical questions respectively.
****

Tianyu Huang,Yihan Zeng,Bowen Dong,Hang Xu,Songcen Xu,Rynson W. H. Lau,Wangmeng Zuo
TextField3D: Towards Enhancing Open-Vocabulary 3D Generation with Noisy Text Fields
Recent works learn 3D representation explicitly under text-3D guidance. However, limited text-3D data restricts the vocabulary scale and text control of generations. Generators may easily fall into a stereotype concept for certain text prompts, thus losing open-vocabulary generation ability. To tackle this issue, we introduce a conditional 3D generative model, namely TextField3D. Specifically, rather than using the text prompts as input directly, we suggest to inject dynamic noise into the latent space of given text prompts, i.e., Noisy Text Fields (NTFs). In this way, limited 3D data can be mapped to the appropriate range of textual latent space that is expanded by NTFs. To this end, an NTFGen module is proposed to model general text latent code in noisy fields. Meanwhile, an NTFBind module is proposed to align view-invariant image latent code to noisy fields, further supporting image-conditional 3D generation. To guide the conditional generation in both geometry and texture, multi-modal discrimination is constructed with a text-3D discriminator and a text-2.5D discriminator. Compared to previous methods, TextField3D includes three merits: 1) large vocabulary, 2) text consistency, and 3) low latency. Extensive experiments demonstrate that our method achieves a potential open-vocabulary 3D generation capability.
****

Yang Song,Prafulla Dhariwal
Improved Techniques for Training Consistency Models
Consistency models are a nascent family of generative models that can sample high quality data in one step without the need for adversarial training. Current consistency models achieve optimal sample quality by distilling from pre-trained diffusion models, and employing learned metrics such as LPIPS. However, distillation limits the quality of consistency models to that of the pre-trained diffusion model, and LPIPS causes undesirable bias in evaluation. To tackle these challenges, we present improved techniques for consistency training, where consistency models learn directly from data without distillation. We delve into the theory behind consistency training and identify a previously overlooked flaw, which we address by eliminating Exponential Moving Average from the teacher consistency model. To replace learned metrics like LPIPS, we borrow Pseudo-Huber losses from robust statistics. Additionally, we introduce a new noise schedule for the consistency training objective, and propose a new curriculum for total discretization steps. Collectively, these modifications enable consistency models to achieve FID scores of 2.62 and 3.91 on CIFAR-10 and ImageNet 64×64 respectively in a single sampling step. These scores mark a 3.3$\times$ improvement compared to prior consistency training approaches. Through two-step sampling, we further reduce FID scores to 2.28 and 3.64, surpassing those obtained via distillation in both one-step and two-step settings, while narrowing the gap between consistency models and state-of-the-art generative models on both datasets.
****

Zhantao Yang,Ruili Feng,Han Zhang,Yujun Shen,Kai Zhu,Lianghua Huang,Yifei Zhang,Yu Liu,Deli Zhao,Jingren Zhou,Fan Cheng
Lipschitz Singularities in Diffusion Models
Diffusion models, which employ stochastic differential equations to sample image

s through integrals, have emerged as a dominant class of generative models. Howe
ver, the rationality of the diffusion process itself receives limited attention,
 leaving the question of whether the problem is well-posed and well-conditioned.
 In this paper, we uncover a vexing propensity of diffusion models: they frequen
tly exhibit the infinite Lipschitz near the zero point of timesteps. We provide
theoretical proofs to illustrate the presence of infinite Lipschitz constants an
d empirical results to confirm it. The Lipschitz singularities pose a threat to
the stability and accuracy during both the training and inference processes of d
iffusion models. Therefore, the mitigation of Lipschitz singularities holds grea
t potential for enhancing the performance of diffusion models. To address this c
hallenge, we propose a novel approach, dubbed E-TSDM, which alleviates the Lipsc
hitz singularities of the diffusion model near the zero point. Remarkably, our t
echnique yields a substantial improvement in performance. Moreover, as a byprodu
ct of our method, we achieve a dramatic reduction in the Fréchet Inception Dista
nce of acceleration methods relying on network Lipschitz, including DDIM and DPM
-Solver, by over 33\%. Extensive experiments on diverse datasets validate our th
eory and method. Our work may advance the understanding of the general diffusion
 process, and also provide insights for the design of diffusion models.
****
Prithvijit Chattopadhyay,Bharat Goyal,Boglarka Ecsedi,Viraj Uday Prabhu,Judy Hof
fman
AUGCAL: Improving Sim2Real Adaptation by Uncertainty Calibration on Augmented Sy
nthetic Images
Synthetic data (Sim) drawn from simulators have emerged as a popular alternative
for training models where acquiring annotated real-world images is difficult. Ho
wever, transferring models trained on synthetic images to real-world application
scan be challenging due to appearance disparities. A commonly employed solution
to counter this Sim2Real gap is unsupervised domain adaptation, where models are
 trained using labeled Sim data and unlabeled Real data. Mispredictions made by
such Sim2Real adapted models are often associated with miscalibration – stemming
 from overconfident predictions on real data. In this paper, we introduce AUGCAL
, a simple training-time patch for unsupervised adaptation that improves Sim2Rea
l adapted models by – (1) reducing overall miscalibration, (2) reducing overconf
idence in incorrect predictions and (3) improving confidence score reliability b
y better guiding misclassification detection – all while retaining or improving
Sim2Real performance. Given a base Sim2Real adaptation algorithm, at training ti
me, AUGCAL involves replacing vanilla Sim images with strongly augmented views (
AUG intervention) and additionally optimizing for a training time calibration lo
ss on augmented Sim predictions (CAL intervention). We motivate AUGCAL using a b
rief analytical justification of how to reduce miscalibration on unlabeled REAL
data. Through our experiments, we empirically show the efficacy of AUGCAL across
 multiple adaptation methods, backbones, tasks and shifts.
****
Tianqi Du,Yifei Wang,Yisen Wang
On the Role of Discrete Tokenization in Visual Representation Learning
In the realm of self-supervised learning (SSL), masked image modeling (MIM) has
gained popularity alongside contrastive learning methods. MIM involves reconstru
cting masked regions of input images using unmasked portions. A notable subset o
f MIM methodologies employs discrete visual tokens as reconstruction target. Thi
s study explores the role of discrete visual tokens in MIM, with the aim of deco
ding their potential benefits and inherent constraints. Building upon the connec
tion between MIM and contrastive learning, we provide comprehensive explanations
 on how discrete tokenization affects generalization performance of MIM. Further
more, we introduce a novel metric designed to quantify the proficiency of discre
te visual tokens in the MIM framework. Inspired by this metric, we contribute an
 accessible tokenizer design and demonstrate its superior performance across var
ious benchmark datasets and ViT backbones.
****
Francesco Bacchiocchi,Matteo Castiglioni,Alberto Marchesi,Nicola Gatti
Learning Optimal Contracts: How to Exploit Small Action Spaces

We study principal-agent problems in which a principal commits to an outcome-dependent payment scheme---called contract---in order to induce an agent to take a costly, unobservable action leading to favorable outcomes. We consider a generalization of the classical (single-round) version of the problem in which the principal interacts with the agent by committing to contracts over multiple rounds. The principal has no information about the agent, and they have to learn an optimal contract by only observing the outcome realized at each round. We focus on settings in which the size of the agent's action space is small. We design an algorithm that learns an approximately-optimal contract with high probability in a number of rounds polynomial in the size of the outcome space, when the number of actions is constant. Our algorithm solves an open problem by Zhu et al. [2022]. Moreover, it can also be employed to provide a $\widetilde{\mathcal{O}}(T^{4/5})$ regret bound in the related online learning setting in which the principal aims at maximizing their cumulative utility, thus considerably improving previously-known regret bounds.
****

Xiaoqi Wang,Han Wei Shen
GNNBoundary: Towards Explaining Graph Neural Networks through the Lens of Decision Boundaries
While Graph Neural Networks (GNNs) have achieved remarkable performance on various machine learning tasks on graph data, they also raised questions regarding their transparency and interpretability. Recently, there have been extensive research efforts to explain the decision-making process of GNNs. These efforts often focus on explaining why a certain prediction is made for a particular instance, or what discriminative features the GNNs try to detect for each class. However, to the best of our knowledge, there is no existing study on understanding the decision boundaries of GNNs, even though the decision-making process of GNNs is directly determined by the decision boundaries. To bridge this research gap, we propose a model-level explainability method called GNNBoundary, which attempts to gain deeper insights into the decision boundaries of graph classifiers. Specifically, we first develop an algorithm to identify the pairs of classes whose decision regions are adjacent. For an adjacent class pair, the near-boundary graphs between them are effectively generated by optimizing a novel objective function specifically designed for boundary graph generation. Thus, by analyzing the nearboundary graphs, the important characteristics of decision boundaries can be uncovered. To evaluate the efficacy of GNNBoundary, we conduct experiments on both synthetic and public real-world datasets. The results demonstrate that, via the analysis of faithful near-boundary graphs generated by GNNBoundary, we can thoroughly assess the robustness and generalizability of the explained GNNs. The official implementation can be found at https://github.com/yolandalalala/GNNBoundary.
****

Daniel Geng,Andrew Owens
Motion Guidance: Diffusion-Based Image Editing with Differentiable Motion Estimators
Diffusion models are capable of generating impressive images conditioned on text descriptions, and extensions of these models allow users to edit images at a relatively coarse scale. However, the ability to precisely edit the layout, position, pose, and shape of objects in images with diffusion models is still difficult. To this end, we propose {\it motion guidance}, a zero-shot technique that allows a user to specify dense, complex motion fields that indicate where each pixel in an image should move. Motion guidance works by steering the diffusion sampling process with the gradients through an off-the-shelf optical flow network. Specifically, we design a guidance loss that encourages the sample to have the desired motion, as estimated by a flow network, while also being visually similar to the source image. By simultaneously sampling from a diffusion model and guiding the sample to have low guidance loss, we can obtain a motion-edited image. We demonstrate that our technique works on complex motions and produces high quality edits of real and generated images.
****

Antonis Antoniades,Yiyi Yu,Joe S Canzano,William Yang Wang,Spencer Smith

Neuroformer: Multimodal and Multitask Generative Pretraining for Brain Data
State-of-the-art systems neuroscience experiments yield large-scale multimodal data, and these data sets require new tools for analysis. Inspired by the success of large pretrained models in vision and language domains, we reframe the analysis of large-scale, cellular-resolution neuronal spiking data into an auto-regressive spatiotemporal generation problem. Neuroformer is a multimodal, multitask generative pre-trained transformer (GPT) model that is specifically designed to handle the intricacies of data in systems neuroscience. It scales linearly with feature size, can process an arbitrary number of modalities, and is adaptable to downstream tasks, such as predicting behavior. We first trained Neuroformer on simulated datasets, and found that it both accurately predicted simulated neuronal circuit activity, and also intrinsically inferred the underlying neural circuit connectivity, including direction. When pretrained to decode neural responses, the model predicted the behavior of a mouse with only few-shot fine-tuning, suggesting that the model begins learning how to do so directly from the neural representations themselves, without any explicit supervision. We used an ablation study to show that joint training on neuronal responses and behavior boosted performance, highlighting the model's ability to associate behavioral and neural representations in an unsupervised manner. These findings show that Neuroformer can analyze neural datasets and their emergent properties, informing the development of models and hypotheses associated with the brain.
****
Athul Paul Jacob,Abhishek Gupta,Jacob Andreas
Modeling Boundedly Rational Agents with Latent Inference Budgets
We study the problem of modeling a population of agents pursuing unknown goals subject to unknown computational constraints. In standard models of bounded rationality, sub-optimal decision-making is simulated by adding homoscedastic noise to optimal decisions rather than actually simulating constrained inference. In this work, we introduce a latent inference budget model (L-IBM) that models these constraints explicitly, via a latent variable (inferred jointly with a model of agents’ goals) that controls the runtime of an iterative inference algorithm. L-IBMs make it possible to learn agent models using data from diverse populations of suboptimal actors. In three modeling tasks—inferring navigation goals from routes, inferring communicative intents from human utterances, and predicting next moves in human chess games—we show that L-IBMs match or outperforms Boltzmann models of decision-making under uncertainty. Moreover, the inferred inference budgets are themselves meaningful, efficient to compute, and correlated with measures of player skill, partner skill and task difficulty.
****
Huan He,William hao,Yuanzhe Xi,Yong Chen,Bradley Malin,Joyce Ho
A Flexible Generative Model for Heterogeneous Tabular EHR with Missing Modality
Realistic synthetic electronic health records (EHRs) can be leveraged to acceler- ate methodological developments for research purposes while mitigating privacy concerns associated with data sharing. However, the training of Generative Ad- versarial Networks remains challenging, often resulting in issues like mode col- lapse. While diffusion models have demonstrated progress in generating qual- ity synthetic samples for tabular EHRs given ample denoising steps, their perfor- mance wanes when confronted with missing modalities in heterogeneous tabular EHRs data. For example, some EHRs contain solely static measurements, and some contain only contain temporal measurements, or a blend of both data types. To bridge this gap, we introduce FLEXGEN-EHR– a versatile diffusion model tai- lored for heterogeneous tabular EHRs, equipped with the capability of handling missing modalities in an integrative learning framework. We define an optimal transport module to align and accentuate the common feature space of hetero- geneity of EHRs. We empirically show that our model consistently outperforms existing state-of-the-art synthetic EHR generation methods both in fidelity by up to 3.10% and utility by up to 7.16%. Additionally, we show that our method can be successfully used in privacy-sensitive settings, where the original patient-level data cannot be shared.
****

Enshu Liu,Xuefei Ning,Huazhong Yang,Yu Wang
A Unified Sampling Framework for Solver Searching of Diffusion Probabilistic Mod
els
Recent years have witnessed the rapid progress and broad application of diffusio
n probabilistic models (DPMs). Sampling from DPMs can be viewed as solving an or
dinary differential equation (ODE). Despite the promising performance, the gener
ation of DPMs usually consumes much time due to the large number of function eva
luations (NFE). Though recent works have accelerated the sampling to around 20 s
teps with high-order solvers, the sample quality with less than 10 NFE can still
 be improved. In this paper, we propose a unified sampling framework (USF) to st
udy the optional strategies for solver. Under this framework, we further reveal
that taking different solving strategies at different timesteps may help further
 decrease the truncation error, and a carefully designed \emph{solver schedule}
has the potential to improve the sample quality by a large margin. Therefore, we
 propose a new sampling framework based on the exponential integral formulation
that allows free choices of solver strategy at each step and design specific dec
isions for the framework. Moreover, we propose S^3, a predictor-based search m
ethod that automatically optimizes the solver schedule to get a better time-qual
ity trade-off of sampling. We demonstrate that S^3 can find outstanding solver
 schedules which outperform the state-of-the-art sampling methods on CIFAR-10, C
elebA, ImageNet-64, and LSUN-Bedroom datasets. Specifically, we achieve 2.69 FID
 with 9 NFE and 6.86 FID with 5 NFE on CIFAR-10 dataset, outperforming the SOTA
method significantly. We further apply S^3 to Stable-Diffusion model and get a
n acceleration ratio of 2$\times$, showing the feasibility of sampling in very f
ew steps without retraining of the neural network.
****
Leo Feng,Frederick Tung,Hossein Hajimirsadeghi,Yoshua Bengio,Mohamed Osama Ahmed
Tree Cross Attention
Cross Attention is a popular method for retrieving information from a set of con
text tokens for making predictions. At inference time, for each prediction, Cros
s Attention scans the full set of $\mathcal{O}(N)$ tokens. In practice, however,
 often only a small subset of tokens are required for good performance.
Methods such as Perceiver IO are cheap at inference as they distill the informat
ion to a smaller-sized set of latent tokens $L < N$ on which cross attention is
then applied, resulting in only $\mathcal{O}(L)$ complexity.
However, in practice, as the number of input tokens and the amount of informatio
n to distill increases, the number of latent tokens needed also increases signif
icantly.
In this work, we propose Tree Cross Attention (TCA) - a module based on Cross At
tention that only retrieves information from a logarithmic $\mathcal{O}(\log(N))
$ number of tokens for performing inference.
TCA organizes the data in a tree structure and performs a tree search at inferen
ce time to retrieve the relevant tokens for prediction.
Leveraging TCA, we introduce ReTreever, a flexible architecture for token-effici
ent inference.
We show empirically that Tree Cross Attention (TCA) performs comparable to Cross
 Attention across various classification and uncertainty regression tasks while
being significantly more token-efficient.
Furthermore, we compare ReTreever against Perceiver IO, showing significant gain
s while using the same number of tokens for inference.
****
Sirui Hong,Mingchen Zhuge,Jonathan Chen,Xiawu Zheng,Yuheng Cheng,Jinlin Wang,Cey
ao Zhang,Zili Wang,Steven Ka Shing Yau,Zijuan Lin,Liyang Zhou,Chenyu Ran,Lingfen
g Xiao,Chenglin Wu,Jürgen Schmidhuber
MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework
Recently, remarkable progress has been made on automated problem solving through
 societies of agents based on large language models (LLMs). Previous LLM-based m
ulti-agent systems can already solve simple dialogue tasks. More complex tasks,
however, face challenges through logic inconsistencies due to cascading hallucin
ations caused by naively chaining LLMs. Here we introduce MetaGPT, an innovative

meta-programming framework incorporating efficient human workflows into LLM-based multi-agent collaborations. MetaGPT encodes Standardized Operating Procedures (SOPs) into prompt sequences for more streamlined workflows, thus allowing agents with human-like domain expertise to verify intermediate results and reduce errors. MetaGPT utilizes an assembly line paradigm to assign diverse roles to various agents, efficiently breaking down complex tasks into subtasks involving many agents working together. On collaborative software engineering benchmarks, MetaGPT generates more coherent solutions than previous chat-based multi-agent systems.
****
Suhas Kotha,Jacob Mitchell Springer,Aditi Raghunathan
Understanding Catastrophic Forgetting in Language Models via Implicit Inference
We lack a systematic understanding of the effects of fine-tuning (via methods such as instruction-tuning or reinforcement learning from human feedback), particularly on tasks outside the narrow fine-tuning distribution. In a simplified scenario, we demonstrate that improving performance on tasks within the fine-tuning data distribution comes at the expense of capabilities on other tasks. We hypothesize that language models implicitly infer the task of the prompt and that fine-tuning skews this inference towards tasks in the fine-tuning distribution. To test this, we propose Conjugate Prompting, which artificially makes the task look farther from the fine-tuning distribution while requiring the same capability, and we find that this recovers some of the pretraining capabilities on our synthetic setup. Since real-world fine-tuning distributions are predominantly English, we apply conjugate prompting to recover pretrained capabilities in LLMs by simply translating the prompts to different languages. This allows us to recover the in-context learning abilities lost via instruction tuning, and more concerningly, recover harmful content generation suppressed by safety fine-tuning in chatbots like ChatGPT.
****
Noel Loo,Ramin Hasani,Mathias Lechner,Alexander Amini,Daniela Rus
Understanding Reconstruction Attacks with the Neural Tangent Kernel and Dataset Distillation
Modern deep learning requires large volumes of data, which could contain sensitive or private information that cannot be leaked. Recent work has shown for homogeneous neural networks a large portion of this training data could be reconstructed with only access to the trained network parameters. While the attack was shown to work empirically, there exists little formal understanding of its effective regime and which datapoints are susceptible to reconstruction. In this work, we first build a stronger version of the dataset reconstruction attack and show how it can provably recover the \emph{entire training set} in the infinite width regime. We then empirically study the characteristics of this attack on two-layer networks and reveal that its success heavily depends on deviations from the frozen infinite-width Neural Tangent Kernel limit. Next, we study the nature of easily-reconstructed images. We show that both theoretically and empirically, reconstructed images tend to ``outliers'' in the dataset, and that these reconstruction attacks can be used for \textit{dataset distillation}, that is, we can retrain on reconstructed images and obtain high predictive accuracy.
****
Haoheng Lan,Jindong Gu,Philip Torr,Hengshuang Zhao
Influencer Backdoor Attack on Semantic Segmentation
When a small number of poisoned samples are injected into the training dataset of a deep neural network, the network can be induced to exhibit malicious behavior during inferences, which poses potential threats to real-world applications. While they have been intensively studied in classification, backdoor attacks on semantic segmentation have been largely overlooked. Unlike classification, semantic segmentation aims to classify every pixel within a given image. In this work, we explore backdoor attacks on segmentation models to misclassify all pixels of a victim class by injecting a specific trigger on non-victim pixels during inferences, which is dubbed Influencer Backdoor Attack (IBA). IBA is expected to maintain the classification accuracy of non-victim pixels and mislead classificatio

ns of all victim pixels in every single inference and could be easily applied to real-world scenes. Based on the context aggregation ability of segmentation models, we proposed a simple, yet effective, Nearest-Neighbor trigger injection strategy. We also introduce an innovative Pixel Random Labeling strategy which maintains optimal performance even when the trigger is placed far from the victim pixels. Our extensive experiments reveal that current segmentation models do suffer from backdoor attacks, demonstrate IBA real-world applicability, and show that our proposed techniques can further increase attack performance.
****
Jiajun He,Gergely Flamich,Zongyu Guo,José Miguel Hernández-Lobato
RECOMBINER: Robust and Enhanced Compression with Bayesian Implicit Neural Representations
COMpression with Bayesian Implicit NEural Representations (COMBINER) is a recent data compression method that addresses a key inefficiency of previous Implicit Neural Representation (INR)-based approaches: it avoids quantization and enables direct optimization of the rate-distortion performance. However, COMBINER still has significant limitations: 1) it uses factorized priors and posterior approximations that lack flexibility; 2) it cannot effectively adapt to local deviations from global patterns in the data; and 3) its performance can be susceptible to modeling choices and the variational parameters' initializations. Our proposed method, Robust and Enhanced COMBINER (RECOMBINER), addresses these issues by 1) enriching the variational approximation while retaining a low computational cost via a linear reparameterization of the INR weights, 2) augmenting our INRs with learnable positional encodings that enable them to adapt to local details and 3) splitting high-resolution data into patches to increase robustness and utilizing expressive hierarchical priors to capture dependency across patches. We conduct extensive experiments across several data modalities, showcasing that RECOMBINER achieves competitive results with the best INR-based methods and even outperforms autoencoder-based codecs on low-resolution images at low bitrates. Our PyTorch implementation is available at https://github.com/cambridge-mlg/RECOMBINER/.
****
Kai Huang,Hanyun Yin,Heng Huang,Wei Gao
Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation
Fine-tuning is essential to adapting pre-trained large language models to downstream applications. With the increasing popularity of LLM-enabled applications, fine-tuning has been performed intensively worldwide, incurring a tremendous amount of computing costs that correspond to big carbon footprint and environmental impact. Mitigating such environmental impact directly correlates to reducing the fine-tuning FLOPs. Existing fine-tuning schemes focus on either saving memory or reducing the overhead of computing weight updates, but cannot achieve sufficient FLOPs reduction due to their ignorance of the training cost in backpropagation. To address this limitation, in this paper we present GreenTrainer, a new technique that minimizes the FLOPs of LLM fine-tuning via adaptive backpropagation, which adaptively selects the most appropriate set of LLM tensors for fine-tuning based on their importance and backpropagation cost in training. Experiment results show that GreenTrainer can save up to 64\% training FLOPs compared to full fine-tuning, without any noticeable accuracy loss. Compared to the existing schemes such as Prefix Tuning and LoRA, GreenTrainer can achieve up to 4\% improvement of model accuracy, with on-par FLOPs reduction.
****
Pratyush Maini,Sachin Goyal,Zachary Chase Lipton,J Zico Kolter,Aditi Raghunathan
T-MARS: Improving Visual Representations by Circumventing Text Feature Learning
Large web-crawled multimodal datasets have powered a slew of new methods for learning general-purpose visual representations, advancing the state of the art in computer vision and revolutionizing zero- and few-shot recognition. One crucial decision facing practitioners is how, if at all, to curate these ever-larger datasets. For example, the creators of the LAION-5B dataset chose to retain only image-caption pairs whose CLIP similarity score exceeded a designated threshold. I

n this paper, we propose a new state-of-the-art data filtering approach motivated by our observation that nearly 40% of LAION's images contain text that overlaps significantly with the caption. Intuitively, such data could be wasteful as it incentivizes models to perform optical character recognition rather than learning visual features. However, naively removing all such data could also be wasteful, as it throws away images that contain visual features (in addition to overlapping text). Our simple and scalable approach, T-MARS (Text Masking and Re-Scoring), filters out only those pairs where the text dominates the remaining visual features---by first masking out the text and then filtering out those with a low CLIP similarity score of the masked image with original captions. Experimentally, T-MARS is the top ranked approach on Imagenet at ``medium scale'' of DataComp (a data filtering benchmark), and outperforms CLIP filtering by a margin of 6.5% on ImageNet and 4.7% on VTAB. Additionally, we show that the accuracy gains enjoyed by T-MARS linearly increase as data and compute are scaled exponentially.
****
Meng Liu,Yue Liu,KE LIANG,Wenxuan Tu,Siwei Wang,sihang zhou,Xinwang Liu
Deep Temporal Graph Clustering
Deep graph clustering has recently received significant attention due to its ability to enhance the representation learning capabilities of models in unsupervised scenarios. Nevertheless, deep clustering for temporal graphs, which could capture crucial dynamic interaction information, has not been fully explored. It means that in many clustering-oriented real-world scenarios, temporal graphs can only be processed as static graphs. This not only causes the loss of dynamic information but also triggers huge computational consumption. To solve the problem, we propose a general framework for deep Temporal Graph Clustering called TGC, which introduces deep clustering techniques to suit the interaction sequence-based batch-processing pattern of temporal graphs. In addition, we discuss differences between temporal graph clustering and static graph clustering from several levels. To verify the superiority of the proposed framework TGC, we conduct extensive experiments. The experimental results show that temporal graph clustering enables more flexibility in finding a balance between time and space requirements, and our framework can effectively improve the performance of existing temporal graph learning methods. The code is released: https://github.com/MGitHubL/Deep-Temporal-Graph-Clustering.
****
Katja Schwarz,Seung Wook Kim,Jun Gao,Sanja Fidler,Andreas Geiger,Karsten Kreis
WildFusion: Learning 3D-Aware Latent Diffusion Models in View Space
Modern learning-based approaches to 3D-aware image synthesis achieve high photorealism and 3D-consistent viewpoint changes for the generated images. Existing approaches represent instances in a shared canonical space. However, for in-the-wild datasets a shared canonical system can be difficult to define or might not even exist. In this work, we instead model instances in view space, alleviating the need for posed images and learned camera distributions. We find that in this setting, existing GAN-based methods are prone to generating flat geometry and struggle with distribution coverage. We hence propose WildFusion, a new approach to 3D-aware image synthesis based on latent diffusion models (LDMs). We first train an autoencoder that infers a compressed latent representation, which additionally
captures the images’ underlying 3D structure and enables not only reconstruction but also novel view synthesis. To learn a faithful 3D representation, we leverage cues from monocular depth prediction. Then, we train a diffusion model in the 3D-aware latent space, thereby enabling synthesis of high-quality 3D-consistent image samples, outperforming recent state-of-the-art GAN-based methods. Importantly,
our 3D-aware LDM is trained without any direct supervision from multiview images or 3D geometry and does not require posed images or learned pose or camera distributions. It directly learns a 3D representation without relying on canonical camera coordinates. This opens up promising research avenues for scalable 3D-aware image synthesis and 3D content creation from in-the-wild image data.

****
Johannes Hertrich,Christian Wald,Fabian Altekrüger,Paul Hagemann
Generative Sliced MMD Flows with Riesz Kernels
Maximum mean discrepancy (MMD) flows suffer from high computational costs in lar
ge scale computations.
In this paper, we show that MMD flows with Riesz kernels $K(x,y) = - \|x-y\|^r$,
 $r \in (0,2)$
have exceptional properties which allow their efficient computation.
We prove that the MMD of Riesz kernels, which is also known as energy distance,
coincides with the MMD of their sliced version.
As a consequence, the computation of gradients of MMDs can be performed in the o
ne-dimensional setting.
Here, for $r=1$, a simple sorting algorithm can be applied to reduce the complex
ity
from $O(MN+N^2)$ to $O((M+N)\log(M+N))$ for two measures with M and N suppo
rt points.
As another interesting follow-up result, the MMD of compactly supported measures
can be estimated from above and below by the Wasserstein-1 distance.
For the implementations we approximate the gradient of the sliced MMD by using o
nly a finite number P of slices.
We show that the resulting error has complexity \smash{$O(\sqrt{d/P})$}, where $
d$ is the data dimension.
These results enable us to train generative models by approximating MMD gradient
 flows by neural networks even
for image applications. We demonstrate the efficiency of our model by image gene
ration on MNIST, FashionMNIST and CIFAR10.
****
Haoyue Bai,Yifei Ming,Julian Katz-Samuels,Yixuan Li
HYPO: Hyperspherical Out-Of-Distribution Generalization
Out-of-distribution (OOD) generalization is critical for machine learning models
 deployed in the real world. However, achieving this can be fundamentally challe
nging, as it requires the ability to learn invariant features across different d
omains or environments. In this paper, we propose a novel framework HYPO (HYPers
pherical OOD generalization) that provably learns domain-invariant representatio
ns in a hyperspherical space. In particular, our hyperspherical learning algorit
hm is guided by intra-class variation and inter-class separation principles—ensu
ring that features from the same class (across different training domains) are c
losely aligned with their class prototypes, while different class prototypes are
 maximally separated. We further provide theoretical justifications on how our p
rototypical learning objective improves the OOD generalization bound. Through ex
tensive experiments on challenging OOD benchmarks, we demonstrate that our appro
ach outperforms competitive baselines and achieves superior performance. Code is
 available at https://github.com/deeplearning-wisc/hypo.
****
Yihang Chen,Lukas Mauch
Order-Preserving GFlowNets
Generative Flow Networks (GFlowNets) have been introduced as a method to sample
a diverse set of candidates with probabilities proportional to a given reward. H
owever, GFlowNets can only be used with a predefined scalar reward, which can be
 either computationally expensive or not directly accessible, in the case of mul
ti-objective optimization (MOO) tasks for example. Moreover, to prioritize ident
ifying high-reward candidates, the conventional practice is to raise the reward
to a higher exponent, the optimal choice of which may vary across different envi
ronments. To address these issues, we propose Order-Preserving GFlowNets (OP-GFN
s), which sample with probabilities in proportion to a learned reward function t
hat is consistent with a provided (partial) order on the candidates, thus elimin
ating the need for an explicit formulation of the reward function. We theoretica
lly prove that the training process of OP-GFNs gradually sparsifies the learned
reward landscape in single-objective maximization tasks. The sparsification conc
entrates on candidates of a higher hierarchy in the ordering, ensuring explorati

on at the beginning and exploitation towards the end of the training. We demonst rate OP-GFN's state-of-the-art performance in single-objective maximization (tot ally ordered) and multi-objective Pareto front approximation (partially ordered) tasks, including synthetic datasets, molecule generation, and neural architectu re search.
****
Xilie Xu,Keyi Kong,Ning Liu,Lizhen Cui,Di Wang,Jingfeng Zhang,Mohan Kankanhalli
An LLM can Fool Itself: A Prompt-Based Adversarial Attack
The wide-ranging applications of large language models (LLMs), especially in saf ety-critical domains, necessitate the proper evaluation of the LLM's adversarial robustness. This paper proposes an efficient tool to audit the LLM's adversaria l robustness via a prompt-based adversarial attack (PromptAttack). PromptAttack converts adversarial textual attacks into an attack prompt that can cause the vi ctim LLM to output the adversarial sample to fool itself. The attack prompt is c omposed of three important components: (1) original input (OI) including the ori ginal sample and its ground-truth label, (2) attack objective (AO) illustrating a task description of generating a new sample that can fool itself without chang ing the semantic meaning, and (3) attack guidance (AG) containing the perturbati on instructions to guide the LLM on how to complete the task by perturbing the o riginal sample at character, word, and sentence levels, respectively. Besides, w e use a fidelity filter to ensure that PromptAttack maintains the original seman tic meanings of the adversarial examples. Further, we enhance the attack power o f PromptAttack by ensembling adversarial examples at different perturbation leve ls. Comprehensive empirical results using Llama2 and GPT-3.5 validate that Promp tAttack consistently yields a much higher attack success rate compared to AdvGLU E and AdvGLUE++. Interesting findings include that a simple emoji can easily mis lead GPT-3.5 to make wrong predictions. Our source code is available at https:// github.com/GodXuxilie/PromptAttack.
****
William Yang,Byron Zhang,Olga Russakovsky
ImageNet-OOD: Deciphering Modern Out-of-Distribution Detection Algorithms
The task of out-of-distribution (OOD) detection is notoriously ill-defined. Earl ier works focused on new-class detection, aiming to identify label-altering data distribution shifts, also known as "semantic shift." However, recent works argu e for a focus on failure detection, expanding the OOD evaluation framework to ac count for label-preserving data distribution shifts, also known as "covariate sh ift." Intriguingly, under this new framework, complex OOD detectors that were pr eviously considered state-of-the-art now perform similarly to, or even worse tha n the simple maximum softmax probability baseline. This raises the question: wha t are the latest OOD detectors actually detecting? Deciphering the behavior of O OD detection algorithms requires evaluation datasets that decouples semantic shi ft and covariate shift. To aid our investigations, we present ImageNet-OOD, a cl ean semantic shift dataset that minimizes the interference of covariate shift. T hrough comprehensive experiments, we show that OOD detectors are more sensitive to covariate shift than to semantic shift, and the benefits of recent OOD detect ion algorithms on semantic shift detection is minimal. Our dataset and analyses provide important insights for guiding the design of future OOD detectors.
****
Carlos E Jimenez,John Yang,Alexander Wettig,Shunyu Yao,Kexin Pei,Ofir Press,Kart hik R Narasimhan
SWE-bench: Can Language Models Resolve Real-world Github Issues?
Language models have outpaced our ability to evaluate them effectively, but for their future development it is essential to study the frontier of their capabili ties. We find real-world software engineering to be a rich, sustainable, and cha llenging testbed for evaluating the next generation of language models. To this end, we introduce SWE-bench, an evaluation framework consisting of 2,294 softwar e engineering problems drawn from real GitHub issues and corresponding pull requ ests across 12 popular Python repositories. Given a codebase along with a descri ption of an issue to be resolved, a language model is tasked with editing the co debase to address the issue. Resolving issues in SWE-bench frequently requires u

nderstanding and coordinating changes across multiple functions, classes, and even files simultaneously, calling for models to interact with execution environments, process extremely long contexts and perform complex reasoning that goes far beyond traditional code generation tasks. Our evaluations show that both state-of-the-art proprietary models and our fine-tuned model SWE-Llama can resolve only the simplest issues. The best-performing model, Claude 2, is able to solve a mere 1.96% of the issues. Advances on SWE-bench represent steps towards LMs that are more practical, intelligent, and autonomous.
****
Tian Yu Liu,Aditya Golatkar,Stefano Soatto
Tangent Transformers for Composition,Privacy and Removal
We introduce Tangent Attention Fine-Tuning (TAFT), a method for fine-tuning linearized transformers obtained by computing a First-order Taylor Expansion around a pre-trained initialization. We show that the Jacobian-Vector Product resulting from linearization can be computed efficiently in a single forward pass, reducing training and inference cost to the same order of magnitude as its original non-linear counterpart, while using the same number of parameters. Furthermore, we show that, when applied to various downstream visual classification tasks, the resulting Tangent Transformer fine-tuned with TAFT can perform comparably with fine-tuning the original non-linear network. Since Tangent Transformers are linear with respect to the new set of weights, and the resulting fine-tuning loss is convex, we show that TAFT enjoys several advantages compared to non-linear fine-tuning when it comes to model composition, parallel training, machine unlearning, and differential privacy. Our code is available at: https://github.com/tianyu139/tangent-model-composition
****
Mert Kosan,Samidha Verma,Burouj Armgaan,Khushbu Pahwa,Ambuj Singh,Sourav Medya,Sayan Ranu
GNNX-BENCH: Unravelling the Utility of Perturbation-based GNN Explainers through In-depth Benchmarking
Numerous explainability methods have been proposed to shed light on the inner workings of GNNs. Despite the inclusion of empirical evaluations in all the proposed algorithms, the interrogative aspects of these evaluations lack diversity. As a result, various facets of explainability pertaining to GNNs, such as a comparative analysis of counterfactual reasoners, their stability to variational factors such as different GNN architectures, noise, stochasticity in non-convex loss surfaces, feasibility amidst domain constraints, and so forth, have yet to be formally investigated. Motivated by this need, we present a benchmarking study on perturbation-based explainability methods for GNNs, aiming to systematically evaluate and compare a wide range of explainability techniques. Among the key findings of our study, we identify the Pareto-optimal methods that exhibit superior efficacy and stability in the presence of noise. Nonetheless, our study reveals that
all algorithms are affected by stability issues when faced with noisy data. Furthermore, we have established that the current generation of counterfactual explainers often fails to provide feasible recourses due to violations of topological constraints encoded by domain-specific considerations. Overall, this benchmarking study empowers stakeholders in the field of GNNs with a comprehensive understanding of the state-of-the-art explainability methods, potential research problems for further enhancement, and the implications of their application in real-world scenarios.
****
Jin Peng Zhou,Charles E Staats,Wenda Li,Christian Szegedy,Kilian Q Weinberger,Yuhuai Wu
Don't Trust: Verify -- Grounding LLM Quantitative Reasoning with Autoformalization
Large language models (LLM), such as Google's Minerva and OpenAI's GPT families, are becoming increasingly capable of solving mathematical quantitative reasoning problems. However, they still make unjustified logical and computational errors in their reasoning steps and answers. In this paper, we leverage the fact that

if the training corpus of LLMs contained sufficiently many examples of formal mathematics (e.g. in Isabelle, a formal theorem proving environment), they can be prompted to translate i.e. autoformalize informal mathematical statements into formal Isabelle code --- which can be verified automatically for internal consistency. This provides a mechanism to automatically reject solutions whose formalized versions are inconsistent within themselves or with the formalized problem statement. We evaluate our method on GSM8K, MATH and MultiArith datasets and demonstrate that our approach provides a consistently better heuristic than vanilla majority voting --- the previously best method to identify correct answers, by more than 12\% on GSM8K. In our experiments it improves results consistently across all datasets and LLM model sizes. The code can be found at https://github.com/jinpz/dtv.
****
Zikai Xiao,Zihan Chen,Liyinglan Liu,YANG FENG,Joey Tianyi Zhou,Jian Wu,Wanlu Liu,Howard Hao Yang,Zuozhu Liu
FedLoGe: Joint Local and Generic Federated Learning under Long-tailed Data
Federated Long-Tailed Learning (Fed-LT), a paradigm wherein data collected from decentralized local clients manifests a globally prevalent long-tailed distribution, has garnered considerable attention in recent times. In the context of Fed-LT, existing works have predominantly centered on addressing the data imbalance issue to enhance the efficacy of the generic global model while neglecting the performance at the local level. In contrast, conventional Personalized Federated Learning (pFL) techniques are primarily devised to optimize personalized local models under the presumption of a balanced global data distribution. This paper introduces an approach termed Federated Local and Generic Model Training in Fed-LT (FedLoGe), which enhances both local and generic model performance through the integration of representation learning and classifier alignment within a neural collapse framework. Our investigation reveals the feasibility of employing a shared backbone as a foundational framework for capturing overarching global trends, while concurrently employing individualized classifiers to encapsulate distinct refinements stemming from each client’s local features. Building upon this discovery, we establish the Static Sparse Equiangular Tight Frame Classifier (SSE-C), inspired by neural collapse principles that naturally prune extraneous noisy features and foster the acquisition of potent data representations. Furthermore, leveraging insights from imbalance neural collapse's classifier norm patterns, we develop Global and Local Adaptive Feature Realignment (GLA-FR) via an auxiliary global classifier and personalized Euclidean norm transfer to align global features with client preferences. Extensive experimental results on CIFAR-10/100-LT, ImageNet, and iNaturalist demonstrate the advantage of our method over state-of-the-art pFL and Fed-LT approaches.
****
Kazuki Irie,Anand Gopalakrishnan,Jürgen Schmidhuber
Exploring the Promise and Limits of Real-Time Recurrent Learning
Real-time recurrent learning (RTRL) for sequence-processing recurrent neural networks (RNNs) offers certain conceptual advantages over backpropagation through time (BPTT). RTRL requires neither caching past activations nor truncating context, and enables online learning. However, RTRL's time and space complexity make it impractical. To overcome this problem, most recent work on RTRL focuses on approximation theories, while experiments are often limited to diagnostic settings. Here we explore the practical promise of RTRL in more realistic settings. We study actor-critic methods that combine RTRL and policy gradients, and test them in several subsets of DMLab-30, ProcGen, and Atari-2600 environments. On DMLab memory tasks, our system trained on fewer than 1.2B environmental frames is competitive with or outperforms well-known IMPALA and R2D2 baselines trained on 10B frames. To scale to such challenging tasks, we focus on certain well-known neural architectures with element-wise recurrence, allowing for tractable RTRL without approximation. Importantly, we also discuss rarely addressed limitations of RTRL in real-world applications, such as its complexity in the multi-layer case.
****
Benjie Wang,Joel Jennings,Wenbo Gong

Neural structure learning with stochastic differential equations
Discovering the underlying relationships among variables from temporal observations has been a longstanding challenge in numerous scientific disciplines, including biology, finance, and climate science. The dynamics of such systems are often best described using continuous-time stochastic processes. Unfortunately, most existing structure learning approaches assume that the underlying process evolves in discrete-time and/or observations occur at regular time intervals. These mismatched assumptions can often lead to incorrect learned structures and models. In this work, we introduce a novel structure learning method, SCOTCH, which combines neural stochastic differential equations (SDE) with variational inference to infer a posterior distribution over possible structures. This continuous-time approach can naturally handle both learning from and predicting observations at arbitrary time points. Theoretically, we establish sufficient conditions for an SDE and SCOTCH to be structurally identifiable, and prove its consistency under infinite data limits. Empirically, we demonstrate that our approach leads to improved structure learning performance on both synthetic and real-world datasets compared to relevant baselines under regular and irregular sampling intervals.
****
Jiaxiang Tang,Jiawei Ren,Hang Zhou,Ziwei Liu,Gang Zeng
DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation
Recent advances in 3D content creation mostly leverage optimization-based 3D generation via score distillation sampling (SDS).
Though promising results have been exhibited, these methods often suffer from slow per-sample optimization, limiting their practical usage.
In this paper, we propose DreamGaussian, a novel 3D content generation framework that achieves both efficiency and quality simultaneously.
Our key insight is to design a generative 3D Gaussian Splatting model with companioned mesh extraction and texture refinement in UV space.
In contrast to the occupancy pruning used in Neural Radiance Fields, we demonstrate that the progressive densification of 3D Gaussians converges significantly faster for 3D generative tasks.
To further enhance the texture quality and facilitate downstream applications, we introduce an efficient algorithm to convert 3D Gaussians into textured meshes and apply a fine-tuning stage to refine the details.
Extensive experiments demonstrate the superior efficiency and competitive generation quality of our proposed approach.
Notably, DreamGaussian produces high-quality textured meshes in just 2 minutes from a single-view image, achieving approximately 10 times acceleration compared to existing methods.
****
Tian Yu Liu,Matthew Trager,Alessandro Achille,Pramuditha Perera,Luca Zancato,Stefano Soatto
Meaning Representations from Trajectories in Autoregressive Models
We propose to extract meaning representations from autoregressive language models by considering the distribution of all possible trajectories extending an input text. This strategy is prompt-free, does not require fine-tuning, and is applicable to any pre-trained autoregressive model. Moreover, unlike vector-based representations, distribution-based representations can also model asymmetric relations (e.g., direction of logical entailment, hypernym/hyponym relations) by using algebraic operations between likelihood functions. These ideas are grounded in distributional perspectives on semantics and are connected to standard constructions in automata theory, but to our knowledge they have not been applied to modern language models. We empirically show that the representations obtained from large models align well with human annotations, outperform other zero-shot and prompt-free methods on semantic similarity tasks, and can be used to solve more complex entailment and containment tasks that standard embeddings cannot handle. Finally, we extend our method to represent data from different modalities (e.g., image and text) using multimodal autoregressive models. Our code is available at: https://github.com/tianyu139/meaning-as-trajectories
****

Neria Uzan,Nir Weinberger
A representation-learning game for classes of prediction tasks
We propose a game-based formulation for learning dimensionality-reducing represe
ntations of feature vectors, when only a prior knowledge on future prediction ta
sks is available. In this game, the first player chooses a representation, and t
hen the second player adversarially chooses a prediction task from a given class
, representing the prior knowledge. The first player aims to minimize, and the s
econd player to maximize, the regret: The minimal prediction loss using the repr
esentation, compared to the same loss using the original features. We consider t
he canonical setting in which the representation, the response to predict and th
e predictors are all linear functions, and the loss function is the mean squared
 error. We derive the theoretically optimal representation in pure strategies, w
hich shows the effectiveness of the prior knowledge, and the optimal regret in m
ixed strategies, which shows the usefulness of randomizing the representation. F
or general representation, prediction and loss functions, we propose an efficien
t algorithm to optimize a randomized representation. The algorithm only requires
 the gradients of the loss function, and is based on incrementally adding a repr
esentation rule to a mixture of such rules.
****
Haojie Huang,Owen Lewis Howell,Dian Wang,Xupeng Zhu,Robert Platt,Robin Walters
Fourier Transporter: Bi-Equivariant Robotic Manipulation in 3D
Many complex robotic manipulation tasks can be decomposed as a sequence of pick
and place actions. Training a robotic agent to learn this sequence over many dif
ferent starting conditions typically requires many iterations or demonstrations,
 especially in 3D environments. In this work, we propose Fourier Transporter ($\
text{FourTran}$), which leverages the two-fold $\mathrm{SE}(d)\times\mathrm{SE}(
d)$ symmetry in the pick-place problem to achieve much higher sample efficiency
. FourTran is an open-loop behavior cloning method trained using expert
 demonstrations to predict pick-place actions on new configurations. $\text{Four
Tran}$ is constrained by the symmetries of the pick and place actions independen
tly. Our method utilizes a fiber space Fourier transformation that allows for me
mory-efficient computation. Tests on the RLbench benchmark achieve state-of-the-
art results across various tasks.
****
Hong Wang,Zhongkai Hao,Jie Wang,Zijie Geng,Zhen Wang,Bin Li,Feng Wu
Accelerating Data Generation for Neural Operators via Krylov Subspace Recycling
Learning neural operators for solving partial differential equations (PDEs) has
attracted great attention due to its high inference efficiency.
However, training such operators requires generating a substantial amount of lab
eled data, i.e., PDE problems together with their solutions.
The data generation process is exceptionally time-consuming, as it involves solv
ing numerous systems of linear equations to obtain numerical solutions to the PD
Es.
Many existing methods solve these systems independently without considering thei
r inherent similarities, resulting in extremely redundant computations.
To tackle this problem, we propose a novel method, namely **S**orting **K**rylov
 Recycling (**SKR**), to boost the efficiency of solving these systems, thus
 significantly accelerating data generation for neural operators training.
To the best of our knowledge, SKR is the first attempt to address the time-consu
ming nature of data generation for learning neural operators.
The working horse of SKR is Krylov subspace recycling, a powerful technique for
solving a series of interrelated systems by leveraging their inherent similariti
es.
Specifically, SKR employs a sorting algorithm to arrange these systems in a sequ
ence, where adjacent systems exhibit high similarities.
Then it equips a solver with Krylov subspace recycling to solve the systems sequ
entially instead of independently, thus effectively enhancing the solving effici
ency.
Both theoretical analysis and extensive experiments demonstrate that SKR can sig
nificantly accelerate neural operator data generation, achieving a remarkable sp

eedup of up to 13.9 times.
****
Xiao Hu,Jianxiong Li,Xianyuan Zhan,Qing-Shan Jia,Ya-Qin Zhang
Query-Policy Misalignment in Preference-Based Reinforcement Learning
Preference-based reinforcement learning (PbRL) provides a natural way to align RL agents' behavior with human desired outcomes, but is often restrained by costly human feedback. To improve feedback efficiency, most existing PbRL methods focus on selecting queries to maximally improve the overall quality of the reward model, but counter-intuitively, we find that this may not necessarily lead to improved performance. To unravel this mystery, we identify a long-neglected issue in the query selection schemes of existing PbRL studies: Query-Policy Misalignment. We show that the seemingly informative queries selected to improve the overall quality of reward model actually may not align with RL agents' interests, thus offering little help on policy learning and eventually resulting in poor feedback efficiency. We show that this issue can be effectively addressed via policy-aligned query and a specially designed hybrid experience replay, which together enforce the bidirectional query-policy alignment. Simple yet elegant, our method can be easily incorporated into existing approaches by changing only a few lines of code. We showcase in comprehensive experiments that our method achieves substantial gains in both human feedback and RL sample efficiency, demonstrating the importance of addressing query-policy misalignment in PbRL tasks.
****
Ming Jin,Shiyu Wang,Lintao Ma,Zhixuan Chu,James Y. Zhang,Xiaoming Shi,Pin-Yu Chen,Yuxuan Liang,Yuan-Fang Li,Shirui Pan,Qingsong Wen
Time-LLM: Time Series Forecasting by Reprogramming Large Language Models
Time series forecasting holds significant importance in many real-world dynamic systems and has been extensively studied. Unlike natural language process (NLP) and computer vision (CV), where a single large model can tackle multiple tasks, models for time series forecasting are often specialized, necessitating distinct designs for different tasks and applications. While pre-trained foundation models have made impressive strides in NLP and CV, their development in time series domains has been constrained by data sparsity. Recent studies have revealed that large language models (LLMs) possess robust pattern recognition and reasoning abilities over complex sequences of tokens. However, the challenge remains in effectively aligning the modalities of time series data and natural language to leverage these capabilities. In this work, we present Time-LLM, a reprogramming framework to repurpose LLMs for general time series forecasting with the backbone language models kept intact. We begin by reprogramming the input time series with text prototypes before feeding it into the frozen LLM to align the two modalities. To augment the LLM's ability to reason with time series data, we propose Prompt-as-Prefix (PaP), which enriches the input context and directs the transformation of reprogrammed input patches. The transformed time series patches from the LLM are finally projected to obtain the forecasts. Our comprehensive evaluations demonstrate that \method is a powerful time series learner that outperforms state-of-the-art, specialized forecasting models. Moreover, Time-LLM excels in both few-shot and zero-shot learning scenarios. The code is made available at https://github.com/KimMeen/Time-LLM.
****
Ziyang Luo,Can Xu,Pu Zhao,Qingfeng Sun,Xiubo Geng,Wenxiang Hu,Chongyang Tao,Jing Ma,Qingwei Lin,Daxin Jiang
WizardCoder: Empowering Code Large Language Models with Evol-Instruct
Code Large Language Models (Code LLMs), such as StarCoder, have demonstrated remarkable performance in various code-related tasks. However, different from their counterparts in the general language modeling field, the technique of instruction fine-tuning remains relatively under-researched in this domain. In this paper, we present Code Evol-Instruct, a novel approach that adapts the Evol-Instruct method to the realm of code, enhancing Code LLMs to create novel models, WizardCoder. Through comprehensive experiments on five prominent code generation benchmarks, namely HumanEval, HumanEval+, MBPP, DS-1000, and MultiPL-E, our models showcase outstanding performance. They consistently outperform all other open-sourc

e Code LLMs by a significant margin. Remarkably, WizardCoder 15B even surpasses the well-known closed-source LLMs, including Anthropic's Claude and Google's Bar d, on the HumanEval and HumanEval+ benchmarks. Additionally, WizardCoder 34B not only achieves a HumanEval score comparable to GPT3.5 (ChatGPT) but also surpass es it on the HumanEval+ benchmark. Furthermore, our preliminary exploration high lights the pivotal role of instruction complexity in achieving exceptional codin g performance.
****
Haoyu Lu,Guoxing Yang,Nanyi Fei,Yuqi Huo,Zhiwu Lu,Ping Luo,Mingyu Ding
VDT: General-purpose Video Diffusion Transformers via Mask Modeling
This work introduces Video Diffusion Transformer (VDT), which pioneers the use o f transformers in diffusion-based video generation.
It features transformer blocks with modularized temporal and spatial attention m odules to leverage the rich spatial-temporal representation inherited in transfo rmers. Additionally, we propose a unified spatial-temporal mask modeling mechani sm, seamlessly integrated with the model, to cater to diverse video generation s cenarios.

VDT offers several appealing benefits. (1) It excels at capturing temporal depen dencies to produce temporally consistent video frames and even simulate the phys ics and dynamics of 3D objects over time. (2) It facilitates flexible conditioni ng information, e.g., simple concatenation in the token space, effectively unify ing different token lengths and modalities. (3) Pairing with our proposed spatia l-temporal mask modeling mechanism, it becomes a general-purpose video diffuser for harnessing a range of tasks, including unconditional generation, video predi ction, interpolation, animation, and completion, etc. Extensive experiments on t hese tasks spanning various scenarios, including autonomous driving, natural wea ther, human action, and physics-based simulation, demonstrate the effectiveness of VDT. Moreover, we provide a comprehensive study on the capabilities of VDT in capturing accurate temporal dependencies, handling conditioning information, an d the spatial-temporal mask modeling mechanism. Additionally, we present compreh ensive studies on how VDT handles conditioning information with the mask modelin g mechanism, which we believe will benefit future research and advance the field . Codes and models are available at the https://VDT-2023.github.io.
****
Yefei He,Jing Liu,Weijia Wu,Hong Zhou,Bohan Zhuang
EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Model s
Diffusion models have demonstrated remarkable capabilities in image synthesis an d related generative tasks. Nevertheless, their practicality for low-latency rea l-world applications is constrained by substantial computational costs and laten cy issues. Quantization is a dominant way to compress and accelerate diffusion m odels, where post-training quantization (PTQ) and quantization-aware training (Q AT) are two main approaches, each bearing its own properties. While PTQ exhibits efficiency in terms of both time and data usage, it may lead to diminished perf ormance in low bit-width settings. On the other hand, QAT can help alleviate per formance degradation but comes with substantial demands on computational and dat a resources. To capitalize on the advantages while avoiding their respective dra wbacks, we introduce a data-free, quantization-aware and parameter-efficient fin e-tuning framework for low-bit diffusion models, dubbed EfficientDM, to achieve QAT-level performance with PTQ-like efficiency. Specifically, we propose a quant ization-aware variant of the low-rank adapter (QALoRA) that can be merged with m odel weights and jointly quantized to low bit-width. The fine-tuning process dis tills the denoising capabilities of the full-precision model into its quantized counterpart, eliminating the requirement for training data. To further enhance p erformance, we introduce scale-aware optimization to address ineffective learnin g of QALoRA due to variations in weight quantization scales across different lay ers. We also employ temporal learned step-size quantization to handle notable va riations in activation distributions across denoising steps. Extensive experimen tal results demonstrate that our method significantly outperforms previous PTQ-b

ased diffusion models while maintaining similar time and data efficiency. Specifically, there is only a marginal 0.05 sFID increase when quantizing both weights and activations of LDM-4 to 4-bit on ImageNet 256×256. Compared to QAT-based methods, our EfficientDM also boasts a $16.2\times$ faster quantization speed with comparable generation quality, rendering it a compelling choice for practical applications.
****
Gabriele Corso,Arthur Deng,Nicholas Polizzi,Regina Barzilay,Tommi S. Jaakkola
Deep Confident Steps to New Pockets: Strategies for Docking Generalization
Accurate blind docking has the potential to lead to new biological breakthroughs, but for this promise to be realized, docking methods must generalize well across the proteome. Existing benchmarks, however, fail to rigorously assess generalizability. Therefore, we develop DockGen, a new benchmark based on the ligand-binding domains of proteins, and we show that existing machine learning-based docking models have very weak generalization abilities. We carefully analyze the scaling laws of ML-based docking and show that, by scaling data and model size, as well as integrating synthetic data strategies, we are able to significantly increase the generalization capacity and set new state-of-the-art performance across benchmarks. Further, we propose Confidence Bootstrapping, a new training paradigm that solely relies on the interaction between diffusion and confidence models and exploits the multi-resolution generation process of diffusion models. We demonstrate that Confidence Bootstrapping significantly improves the ability of ML-based docking methods to dock to unseen protein classes, edging closer to accurate and generalizable blind docking methods.
****
Junyoung Seo,Wooseok Jang,Min-Seop Kwak,Hyeonsu Kim,Jaehoon Ko,Junho Kim,Jin-Hwa Kim,Jiyoung Lee,Seungryong Kim
Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation
Text-to-3D generation has shown rapid progress in recent days with the advent of score distillation sampling (SDS), a methodology of using pretrained text-to-2D diffusion models to optimize a neural radiance field (NeRF) in a zero-shot setting. However, the lack of 3D awareness in the 2D diffusion model often destabilizes previous methods from generating a plausible 3D scene. To address this issue, we propose 3DFuse, a novel framework that incorporates 3D awareness into the pretrained 2D diffusion model, enhancing the robustness and 3D consistency of score distillation-based methods. Specifically, we introduce a consistency injection module which constructs a 3D point cloud from the text prompt and utilizes its projected depth map at given view as a condition for the diffusion model. The 2D diffusion model, through its generative capability, robustly infers dense structure from the sparse point cloud depth map and generates a geometrically consistent and coherent 3D scene. We also introduce a new technique called semantic coding that reduces semantic ambiguity of the text prompt for improved results. Our method can be easily adapted to various text-to-3D baselines, and we experimentally demonstrate how our method notably enhances the 3D consistency of generated scenes in comparison to previous baselines, achieving state-of-the-art performance in geometric robustness and fidelity.
****
Sobhan Mohammadpour,Emma Frejinger,Pierre-Luc Bacon
Decoupling regularization from the action space
Regularized reinforcement learning (RL), particularly the entropy-regularized kind, has gained traction in optimal control and inverse RL. While standard unregularized RL methods remain unaffected by changes in the number of actions, we show that it can severely impact their regularized counterparts. This paper demonstrates the importance of decoupling the regularizer from the action space: that is, to maintain a consistent level of regularization regardless of how many actions are involved to avoid over-regularization. Whereas the problem can be avoided by introducing a task-specific temperature parameter, it is often undesirable and cannot solve the problem when action spaces are state-dependent. In the state-dependent action context, different states with varying action spaces are regularized inconsistently. We introduce two solutions: a static temperature selectio

n approach and a dynamic counterpart, universally applicable where this problem arises. Implementing these changes improves performance on the DeepMind control suite in static and dynamic temperature regimes and a biological design task.
****
Shahar Lutati,Eliya Nachmani,Lior Wolf
Separate and Diffuse: Using a Pretrained Diffusion Model for Better Source Separation
The problem of speech separation, also known as the cocktail party problem, refers to the task of isolating a single speech signal from a mixture of speech signals. Previous work on source separation derived an upper bound for the source separation task in the domain of human speech. This bound is derived for deterministic models. Recent advancements in generative models challenge this bound. We show how the upper bound can be generalized to the case of random generative models. Applying a diffusion model Vocoder that was pretrained to model single-speaker voices on the output of a deterministic separation model leads
to state-of-the-art separation results. It is shown that this requires one to combine
the output of the separation model with that of the diffusion model. In our method,
a linear combination is performed, in the frequency domain, using weights that are
inferred by a learned model. We show state-of-the-art results on 2, 3, 5, 10, and 20
speakers on multiple benchmarks. In particular, for two speakers, our method is able to surpass what was previously considered the upper performance bound.
****
Prakhar Kaushik,Aayush Mishra,Adam Kortylewski,Alan Yuille
Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation
We consider the problem of source-free unsupervised category-level 3D pose estimation from only RGB images to an non-annotated and unlabelled target domain without any access to source domain data or annotations during adaptation. Collecting and annotating real world 3D data and corresponding images is laborious, expensive yet unavoidable process since even 3D pose domain adaptation methods require 3D data in the target domain. We introduce a method which is capable of adapting to a nuisance ridden target domain without any 3D data or annotations. We represent object categories as simple cuboid meshes, and harness a generative model of neural feature activations modeled as a von Mises Fisher distribution at each mesh vertex learnt using differential rendering. We focus on individual mesh vertex features and iteratively update them based on their proximity to corresponding features in the target domain. Our key insight stems from the observation that specific object subparts remain stable across out-of-domain (OOD) scenarios, enabling strategic utilization of these invariant subcomponents for effective model updates. Our model is then trained in an EM fashion alternating between updating the vertex features and feature extractor. We show that our method simulates fine-tuning on a global-pseudo labelled dataset under mild assumptions which converges to the target domain asymptotically. Through extensive empirical validation, we demonstrate the potency of our simple approach in addressing the domain shift challenge and significantly enhancing pose estimation accuracy. By accentuating robust and less changed object subcomponents, our framework contributes to the evolution of UDA techniques in the context of 3D pose estimation using only images from the target domain.
****
Yair Ori Gat,Nitay Calderon,Amir Feder,Alexander Chapanin,Amit Sharma,Roi Reichart
Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals
Causal explanations of the predictions of NLP systems are essential to ensure safety and establish trust. Yet, existing methods often fall short of explaining m

odel predictions effectively or efficiently and are often model-specific. In thi
s paper, we address model-agnostic explanations, proposing two approaches for co
unterfactual (CF) approximation. The first approach is CF generation, where a la
rge language model (LLM) is prompted to change a specific text concept while kee
ping confounding concepts unchanged. While this approach is demonstrated to be v
ery effective, applying LLM at inference-time is costly. We hence present a seco
nd approach based on matching, and propose a method that is guided by an LLM at
training-time and learns a dedicated embedding space. This space is faithful to
a given causal graph and effectively serves to identify matches that approximate
 CFs. After showing theoretically that approximating CFs is required in order to
 construct faithful explanations, we benchmark our approaches and explain severa
l models, including LLMs with billions of parameters. Our empirical results demo
nstrate the excellent performance of CF generation models as model-agnostic expl
ainers. Moreover, our matching approach, which requires far less test-time resou
rces, also provides effective explanations, surpassing many baselines. We also f
ind that Top-K techniques universally improve every tested method. Finally, we s
howcase the potential of LLMs in constructing new benchmarks for model explanati
on and subsequently validate our conclusions. Our work illuminates new pathways
for efficient and accurate approaches to interpreting NLP systems.
****
Gregory Dexter,Borja Ocejo,Sathiya Keerthi,Aman Gupta,Ayan Acharya,Rajiv Khanna
A Precise Characterization of SGD Stability Using Loss Surface Geometry
Stochastic Gradient Descent (SGD) stands as a cornerstone optimization algorithm
 with proven real-world empirical successes but relatively limited theoretical u
nderstanding. Recent research has illuminated a key factor contributing to its p
ractical efficacy: the implicit regularization it instigates. Several studies ha
ve investigated the linear stability property of SGD in the vicinity of a statio
nary point as a predictive proxy for sharpness and generalization error in overp
arameterized neural networks (Wu et al., 2022; Jastrzebski et al., 2019; Cohen e
t al., 2021). In this paper, we delve deeper into the relationship between linea
r stability and sharpness. More specifically, we meticulously delineate the nece
ssary and sufficient conditions for linear stability, contingent on hyperparamet
ers of SGD and the sharpness at the optimum. Towards this end, we introduce a no
vel coherence measure of the loss Hessian that encapsulates pertinent geometric
properties of the loss function that are relevant to the linear stability of SGD
. It enables us to provide a simplified sufficient condition for identifying lin
ear instability at an optimum. Notably, compared to previous works, our analysis
 relies on significantly milder assumptions and is applicable for a broader clas
s of loss functions than known before, encompassing not only mean-squared error
but also cross-entropy loss.
****
Darshil Doshi,Aritra Das,Tianyu He,Andrey Gromov
To Grok or not to Grok: Disentangling Generalization and Memorization on Corrupt
ed Algorithmic Datasets
Robust generalization is a major challenge in deep learning, particularly when t
he number of trainable parameters is very large. In general, it is very difficul
t to know if the network has memorized a particular set of examples or understoo
d the underlying rule (or both). Motivated by this challenge, we study an interp
retable model where generalizing representations are understood analytically, an
d are easily distinguishable from the memorizing ones. Namely, we consider multi
-layer perceptron (MLP) and Transformer architectures trained on modular arithme
tic tasks, where ($\xi \cdot 100\\%$) of labels are corrupted (*i.e.* some resul
ts of the modular operations in the training set are incorrect). We show that (i
) it is possible for the network to memorize the corrupted labels *and* achieve
$100\\%$ generalization at the same time; (ii) the memorizing neurons can be ide
ntified and pruned, lowering the accuracy on corrupted data and improving the ac
curacy on uncorrupted data; (iii) regularization methods such as weight decay, d
ropout and BatchNorm force the network to ignore the corrupted data during optim
ization, and achieve $100\\%$ accuracy on the uncorrupted dataset; and (iv) the
effect of these regularization methods is ("mechanistically") interpretable: wei

ght decay and dropout force all the neurons to learn generalizing representations, while BatchNorm de-amplifies the output of memorizing neurons and amplifies the output of the generalizing ones. Finally, we show that in the presence of regularization, the training dynamics involves two consecutive stages: first, the network undergoes *grokking* dynamics reaching high train *and* test accuracy; second, it unlearns the memorizing representations, where the train accuracy suddenly jumps from $100\\%$ to $100 (1-\xi)\\%$.
****
Tianyu Li,Hyunyoung Jung,Matthew Gombolay,Yong Cho,Sehoon Ha
CrossLoco: Human Motion Driven Control of Legged Robots via Guided Unsupervised Reinforcement Learning
Human motion driven control (HMDC) is an effective approach for generating natural and compelling robot motions while preserving high-level semantics. However, establishing the correspondence between humans and robots with different body structures is not straightforward due to the mismatches in kinematics and dynamics properties, which causes intrinsic ambiguity to the problem. Many previous algorithms approach this motion retargeting problem with unsupervised learning, which requires the prerequisite skill sets. However, it will be extremely costly to learn all the skills without understanding the given human motions, particularly for high-dimensional robots. In this work, we introduce CrossLoco, a guided unsupervised reinforcement learning framework that simultaneously learns robot skills and their correspondence to human motions. Our key innovation is to introduce a cycle-consistency-based reward term designed to maximize the mutual information between human motions and robot states. We demonstrate that the proposed framework can generate compelling robot motions by translating diverse human motions, such as running, hopping, and dancing. We quantitatively compare our CrossLoco against the manually engineered and unsupervised baseline algorithms along with the ablated versions of our framework and demonstrate that our method translates human motions with better accuracy, diversity, and user preference. We also showcase its utility in other applications, such as synthesizing robot movements from language input and enabling interactive robot control.
****
Siyuan Qi,Shuo Chen,Yexin Li,Xiangyu Kong,Junqi Wang,Bangcheng Yang,Pring Wong,Yifan Zhong,Xiaoyuan Zhang,Zhaowei Zhang,Nian Liu,Yaodong Yang,Song-Chun Zhu
CivRealm: A Learning and Reasoning Odyssey in Civilization for Decision-Making Agents
The generalization of decision-making agents encompasses two fundamental elements: learning from past experiences and reasoning in novel contexts. However, the predominant emphasis in most interactive environments is on learning, often at the expense of complexity in reasoning. In this paper, we introduce CivRealm, an environment inspired by the Civilization game. Civilization's profound alignment with human history and society necessitates sophisticated learning, while its ever-changing situations demand strong reasoning to generalize. Particularly, CivRealm sets up an imperfect-information general-sum game with a changing number of players; it presents a plethora of complex features, challenging the agent to deal with open-ended stochastic environments that require diplomacy and negotiation skills. Within CivRealm, we provide interfaces for two typical agent types: tensor-based agents that focus on learning, and language-based agents that emphasize reasoning. To catalyze further research, we present initial results for both paradigms. The canonical RL-based agents exhibit reasonable performance in mini-games, whereas both RL- and LLM-based agents struggle to make substantial progress in the full game.
Overall, CivRealm stands as a unique learning and reasoning challenge for decision-making agents. The code is available at https://github.com/bigai-ai/civrealm.
****
Jiachen Sun,Haizhong Zheng,Qingzhao Zhang,Atul Prakash,Zhuoqing Mao,Chaowei Xiao
CALICO: Self-Supervised Camera-LiDAR Contrastive Pre-training for BEV Perception
Perception is crucial in the realm of autonomous driving systems, where bird's eye view (BEV)-based architectures have recently reached state-of-the-art performance. The desirability of self-supervised representation learning stems from the

expensive and laborious process of annotating 2D and 3D data. Although previou
s research has investigated pretraining methods for both LiDAR and camera-based
3D object detection, a unified pretraining framework for multimodal BEV percepti
on is missing. In this study, we introduce CALICO, a novel framework that applie
s contrastive objectives to both LiDAR and camera backbones. Specifically, CALIC
O incorporates two stages: point-region contrast (PRC) and region-aware distilla
tion (RAD). PRC better balances the region- and scene-level representation learn
ing on the LiDAR modality and offers significant performance improvement compare
d to existing methods. RAD effectively achieves contrastive distillation on our
self-trained teacher model. CALICO's efficacy is substantiated by extensive eval
uations on 3D object detection and BEV map segmentation tasks, where it delivers
significant performance improvements. Notably, CALICO outperforms the baseline
method by 10.5\% and 8.6\% on NDS and mAP. Moreover, CALICO boosts the robustnes
s of multimodal 3D object detection against adversarial attacks and corruption.
Additionally, our framework can be tailored to different backbones and heads, po
sitioning it as a promising approach for multimodal BEV perception.
****
Sangyu Han,Yearim Kim,Nojun Kwak
Respect the model: Fine-grained and Robust Explanation with Sharing Ratio Decomp
osition
The truthfulness of existing explanation methods in authentically elucidating th
e underlying model's decision-making process has been questioned. Existing metho
ds have deviated from faithfully representing the model, thus susceptible to adv
ersarial attacks.
To address this, we propose a novel eXplainable AI (XAI) method called SRD (Shar
ing Ratio Decomposition), which sincerely reflects the model's inference process
, resulting in significantly enhanced robustness in our explanations.
Different from the conventional emphasis on the neuronal level, we adopt a vecto
r perspective to consider the intricate nonlinear interactions between filters.
We also introduce an interesting observation termed Activation-Pattern-Only Pred
iction (APOP), letting us emphasize the importance of inactive neurons and redef
ine relevance encapsulating all relevant information including both active and i
nactive neurons.
Our method, SRD, allows for the recursive decomposition of a Pointwise Feature V
ector (PFV), providing a high-resolution Effective Receptive Field (ERF) at any
layer.
****
Chen Zhao,Tong Zhang,Mathieu Salzmann
3D-Aware Hypothesis & Verification for Generalizable Relative Object Pose Estima
tion
Prior methods that tackle the problem of generalizable object pose estimation hi
ghly rely on having dense views of the unseen object. By contrast, we address th
e scenario where only a single reference view of the object is available. Our go
al then is to estimate the relative object pose between this reference view and
a query image that depicts the object in a different pose. In this scenario, rob
ust generalization is imperative due to the presence of unseen objects during te
sting and the large-scale object pose variation between the reference and the qu
ery. To this end, we present a new hypothesis-and-verification framework, in whi
ch we generate and evaluate multiple pose hypotheses, ultimately selecting the m
ost reliable one as the relative object pose. To measure reliability, we introdu
ce a 3D-aware verification that explicitly applies 3D transformations to the 3D
object representations learned from the two input images. Our comprehensive expe
riments on the Objaverse, LINEMOD, and CO3D datasets evidence the superior accur
acy of our approach in relative pose estimation and its robustness in large-scal
e pose variations, when dealing with unseen objects.
****
Ruiquan Huang,Yuan Cheng,Jing Yang,Vincent Tan,Yingbin Liang
Provable Benefits of Multi-task RL under Non-Markovian Decision Making Processes
In multi-task reinforcement learning (RL) under Markov decision processes (MDPs)
, the presence of shared latent structures among multiple MDPs has been shown to

yield significant benefits to the sample efficiency compared to single-task RL. In this paper, we investigate whether such a benefit can extend to more general sequential decision making problems such as predictive state representations (PSRs). The main challenge here is that the large and complex model space makes it hard to identify what types of common latent structure of multi-task PSRs can reduce the model complexity and improve sample efficiency.
To this end, we posit a joint model class for tasks and use the notion of η-bracketing number to quantify its complexity; this number also serves as a general metric to capture the similarity of tasks and thus determines the benefit of multi-task over single-task RL. We first study upstream multi-task learning over PSRs, in which all tasks share the same observation and action spaces. We propose a provably efficient algorithm UMT-PSR for finding near-optimal policies for all PSRs, and demonstrate that the advantage of multi-task learning manifests if the joint model class of PSRs has a smaller η-bracketing number compared to that of individual single-task learning. We further investigate downstream learning, in which the agent needs to learn a new target task that shares some commonalities with the upstream tasks via a similarity constraint. By exploiting the learned PSRs from the upstream, we develop a sample-efficient algorithm that provably finds a near-optimal policy.
Upon specialization to some examples with small η-bracketing numbers, our results further highlight the benefit compared to directly learning a single-task PSR.
****
Rundi Wu,Ruoshi Liu,Carl Vondrick,Changxi Zheng
Sin3DM: Learning a Diffusion Model from a Single 3D Textured Shape
Synthesizing novel 3D models that resemble the input example as long been pursued by graphics artists and machine learning researchers. In this paper, we present Sin3DM, a diffusion model that learns the internal patch distribution from a single 3D textured shape
and generates high-quality variations with fine geometry and texture details. Training a diffusion model directly in 3D would induce large memory and computational cost. Therefore, we first compress the input into a lower-dimensional latent space and then train a diffusion model on it. Specifically, we encode the input 3D textured shape into triplane feature maps that represent the signed distance and texture fields of the input. The denoising network of our diffusion model has a limited receptive field to avoid overfitting, and uses triplane-aware 2D convolution blocks to improve the result quality. Aside from randomly generating new samples, our model also facilitates applications such as retargeting, outpainting and local editing. Through extensive qualitative and quantitative evaluation, we show that our method outperforms prior methods in generation quality of 3D shapes.
****
Stéphane d'Ascoli,Sören Becker,Philippe Schwaller,Alexander Mathis,Niki Kilbertus
ODEFormer: Symbolic Regression of Dynamical Systems with Transformers
We introduce ODEFormer, the first transformer able to infer multidimensional ordinary differential equation (ODE) systems in symbolic form from the observation of a single solution trajectory. We perform extensive evaluations on two datasets: (i) the existing ‘Strogatz’ dataset featuring two-dimensional systems; (ii) ODEBench, a collection of one- to four-dimensional systems that we carefully curated from the literature to provide a more holistic benchmark. ODEFormer consistently outperforms existing methods while displaying substantially improved robustness to noisy and irregularly sampled observations, as well as faster inference. We release our code, model and benchmark at https://github.com/sdascoli/odeformer.
****
Vivien Cabannes,Elvis Dohmatob,Alberto Bietti
Scaling Laws for Associative Memories
Learning arguably involves the discovery and memorization of abstract rules. The aim of this paper is to study associative memory mechanisms. Our model is based

on high-dimensional matrices consisting of outer products of embeddings, which relates to the inner layers of transformer language models. We derive precise scaling laws with respect to sample size and parameter size, and discuss the statistical efficiency of different estimators, including optimization-based algorithms. We provide extensive numerical experiments to validate and interpret theoretical results, including fine-grained visualizations of the stored memory associations.
****
Xiaotian Han,Jianfeng Chi,Yu Chen,Qifan Wang,Han Zhao,Na Zou,Xia Hu
FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods
This paper introduces the Fair Fairness Benchmark (FFB), a benchmarking framework for in-processing group fairness methods. Ensuring fairness in machine learning is important for ethical compliance. However, there exist challenges in comparing and developing fairness methods due to inconsistencies in experimental settings, lack of accessible algorithmic implementations, and limited extensibility of current fairness packages and tools. To address these issues, we introduce an open-source standardized benchmark for evaluating in-processing group fairness methods and provide a comprehensive analysis of state-of-the-art methods to ensure different notions of group fairness. This work offers the following key contributions: the provision of flexible, extensible, minimalistic, and research-oriented open-source code; the establishment of unified fairness method benchmarking pipelines; and extensive benchmarking, which yields key insights from 45,079 experiments, 14,428 GPU hours. We believe that our work will significantly facilitate the growth and development of the fairness research community. The benchmark is available at https://github.com/ahxt/fair_fairness_benchmark.
****
Josef Dai,Xuehai Pan,Ruiyang Sun,Jiaming Ji,Xinbo Xu,Mickel Liu,Yizhou Wang,Yaodong Yang
Safe RLHF: Safe Reinforcement Learning from Human Feedback
With the development of large language models (LLMs), striking a balance between the performance and safety of AI systems has never been more critical. However, the inherent tension between the objectives of helpfulness and harmlessness presents a significant challenge during LLM training. To address this issue, we propose Safe Reinforcement Learning from Human Feedback (Safe RLHF), a novel algorithm for human value alignment. Safe RLHF explicitly decouples human preferences regarding helpfulness and harmlessness, effectively avoiding the crowd workers' confusion about the tension and allowing us to train separate reward and cost models. We formalize the safety concern of LLMs as an optimization task of maximizing the reward function while satisfying specified cost constraints. Leveraging the Lagrangian method to solve this constrained problem, Safe RLHF dynamically adjusts the balance between the two objectives during fine-tuning. Through a three-round fine-tuning using Safe RLHF, we demonstrate a superior ability to mitigate harmful responses while enhancing model performance compared to existing value-aligned algorithms. Experimentally, we fine-tuned the Alpaca-7B using Safe RLHF and aligned it with collected human preferences, significantly improving its helpfulness and harmlessness according to human evaluations.

Warning: This paper contains example data that may be offensive or harmful.
****
Wei-Hong Li,Steven McDonagh,Ales Leonardis,Hakan Bilen
Multi-task Learning with 3D-Aware Regularization
Deep neural networks have become the standard solution for designing models that can perform multiple dense computer vision tasks such as depth estimation and semantic segmentation thanks to their ability to capture complex correlations in high dimensional feature space across tasks. However, the cross-task correlations that are learned in the unstructured feature space can be extremely noisy and susceptible to overfitting, consequently hurting performance. We propose to address this problem by introducing a structured 3D-aware regularizer which interfaces multiple tasks through the projection of features extracted from an image encoder to a shared 3D feature space and decodes them into their task output space

through differentiable rendering. We show that the proposed method is architectu
re agnostic and can be plugged into various prior multi-task backbones to improv
e their performance; as we evidence using standard benchmarks NYUv2 and PASCAL-C
ontext.
****
Xiong Xu,Kunzhe Huang,Yiming Li,Zhan Qin,Kui Ren
Towards Reliable and Efficient Backdoor Trigger Inversion via Decoupling Benign Features
Recent studies revealed that using third-party models may lead to backdoor threa
ts, where adversaries can maliciously manipulate model predictions based on back
doors implanted during model training. Arguably, backdoor trigger inversion (BTI
), which generates trigger patterns of given benign samples for a backdoored mod
el, is the most critical module for backdoor defenses used in these scenarios. W
ith BTI, defenders can remove backdoors by fine-tuning based on generated poison
ed samples with ground-truth labels or deactivate backdoors by removing trigger
patterns during the inference process. However, we find that existing BTI method
s suffer from relatively poor performance, $i.e.$, their generated triggers are
significantly different from the ones used by the adversaries even in the featur
e space. We argue that it is mostly because existing methods require to 'extract
' backdoor features at first, while this task is very difficult since defenders
have no information ($e.g.$, trigger pattern or target label) about poisoned sam
ples. In this paper, we explore BTI from another perspective where we decouple b
enign features instead of decoupling backdoor features directly. Specifically, o
ur method consists of two main steps, including \textbf{(1)} decoupling benign f
eatures and \textbf{(2)} trigger inversion by minimizing the differences between
 benign samples and their generated poisoned version in decoupled benign feature
s while maximizing the differences in remaining backdoor features. In particular
, our method is more efficient since it doesn't need to `scan' all classes to sp
eculate the target label, as required by existing BTI. We also exploit our BTI m
odule to further design backdoor-removal and pre-processing-based defenses. Exte
nsive experiments on benchmark datasets demonstrate that our defenses can reach
state-of-the-art performances.
****
Chawin Sitawarin,Jaewon Chang,David Huang,Wesson Altoyan,David Wagner
PubDef: Defending Against Transfer Attacks From Public Models
Adversarial attacks have been a looming and unaddressed threat in the industry.
However, through a decade-long history of the robustness evaluation literature,
we have learned that mounting a strong or optimal attack is challenging. It requ
ires both machine learning and domain expertise. In other words, the white-box t
hreat model, religiously assumed by a large majority of the past literature, is
unrealistic. In this paper, we propose a new practical threat model where the ad
versary relies on transfer attacks through publicly available surrogate models.
We argue that this setting will become the most prevalent for security-sensitive
 applications in the future. We evaluate the transfer attacks in this setting an
d propose a specialized defense method based on a game-theoretic perspective. Th
e defenses are evaluated under 24 public models and 11 attack algorithms across
three datasets (CIFAR-10, CIFAR-100, and ImageNet). Under this threat model, our
 defense, PubDef, outperforms the state-of-the-art white-box adversarial trainin
g by a large margin with almost no loss in the normal accuracy. For instance, on
 ImageNet, our defense achieves 62% accuracy under the strongest transfer attack
 vs only 36% of the best adversarially trained model. Its accuracy when not unde
r attack is only 2% lower than that of an undefended model (78% vs 80%). We rele
ase our code at https://github.com/wagner-group/pubdef.
****
Chenxi Sun,Hongyan Li,Yaliang Li,Shenda Hong
TEST: Text Prototype Aligned Embedding to Activate LLM's Ability for Time Series
This work summarizes two ways to accomplish Time-Series (TS) tasks in today's La
rge Language Model (LLM) context: LLM-for-TS (model-centric) designs and trains
a fundamental large model, or fine-tunes a pre-trained LLM for TS data; TS-for-L
LM (data-centric) converts TS into a model-friendly representation to enable the

pre-trained LLM to handle TS data. Given the lack of data, limited resources, semantic context requirements, and so on, this work focuses on TS-for-LLM, where we aim to activate LLM's ability for TS data by designing a TS embedding method suitable for LLM. The proposed method is named TEST. It first tokenizes TS, builds an encoder to embed TS via instance-wise, feature-wise, and text-prototype-aligned contrast, where the TS embedding space is aligned to LLM’s embedding layer space, then creates soft prompts to make LLM more open to that embeddings, and finally implements TS tasks using the frozen LLM. We also demonstrate the feasibility of TS-for-LLM through theory and experiments. Experiments are carried out on TS classification, forecasting, and representation tasks using eight frozen LLMs with various structures and sizes. The results show that the pre-trained LLM with TEST strategy can achieve better or comparable performance than today's SOTA TS models, and offers benefits for few-shot and generalization. By treating LLM as the pattern machine, TEST can endow LLM's ability to process TS data without compromising language ability. We hope that this study will serve as a foundation for future work to support TS+LLM progress.
****
Junbo Li,Zichen Miao,Qiang Qiu,Ruqi Zhang
Training Bayesian Neural Networks with Sparse Subspace Variational Inference
Bayesian neural networks (BNNs) offer uncertainty quantification but come with the downside of substantially increased training and inference costs. Sparse BNNs have been investigated for efficient inference, typically by either slowly introducing sparsity throughout the training or by post-training compression of dense BNNs. The dilemma of how to cut down massive training costs remains, particularly given the requirement to learn about the uncertainty. To solve this challenge, we introduce Sparse Subspace Variational Inference (SSVI), the first fully sparse BNN framework that maintains a consistently sparse Bayesian model throughout the training and inference phases. Starting from a randomly initialized low-dimensional sparse subspace, our approach alternately optimizes the sparse subspace basis selection and its associated parameters. While basis selection is characterized as a non-differentiable problem, we approximate the optimal solution with a removal-and-addition strategy, guided by novel criteria based on weight distribution statistics. Our extensive experiments show that SSVI sets new benchmarks in crafting sparse BNNs, achieving, for instance, a 10-20× compression in model size with under 3\% performance drop, and up to 20× FLOPs reduction during training. Remarkably, SSVI also demonstrates enhanced robustness to hyperparameters, reducing the need for intricate tuning in VI and occasionally even surpassing VI-trained dense BNNs.
****
Junmo Cho,Jaesik Yoon,Sungjin Ahn
Spatially-Aware Transformers for Embodied Agents
Episodic memory plays a crucial role in various cognitive processes, such as the ability to mentally recall past events. While cognitive science emphasizes the significance of spatial context in the formation and retrieval of episodic memory, the current primary approach to implementing episodic memory in AI systems is through transformers that store temporally ordered experiences, which overlooks the spatial dimension. As a result, it is unclear how the underlying structure could be extended to incorporate the spatial axis beyond temporal order alone and thereby what benefits can be obtained. To address this, this paper explores the use of Spatially-Aware Transformer models that incorporate spatial information. These models enable the creation of place-centric episodic memory that considers both temporal and spatial dimensions. Adopting this approach, we demonstrate that memory utilization efficiency can be improved, leading to enhanced accuracy in various place-centric downstream tasks. Additionally, we propose the Adaptive Memory Allocator, a memory management method based on reinforcement learning that aims to optimize efficiency of memory utilization. Our experiments demonstrate the advantages of our proposed model in various environments and across multiple downstream tasks, including prediction, generation, reasoning, and reinforcement learning. The source code for our models and experiments will be available at \href{https://github.com/spatially_aware_transformer}{https://github.com/spat

ially_aware_transformer}.
****
Ashmit Khandelwal,Aditya Agrawal,Aanisha Bhattacharyya,Yaman Kumar,Somesh Singh,Uttaran Bhattacharya,Ishita Dasgupta,Stefano Petrangeli,Rajiv Ratn Shah,Changyou Chen,Balaji Krishnamurthy
Large Content And Behavior Models To Understand, Simulate, And Optimize Content And Behavior
Shannon and Weaver's seminal information theory divides communication into three levels: technical, semantic, and effectiveness. While the technical level deals with the accurate reconstruction of transmitted symbols, the semantic and effectiveness levels deal with the inferred meaning and its effect on the receiver. Large Language Models (LLMs), with their wide generalizability, make some progress towards the second level. However, LLMs and other communication models are not conventionally designed for predicting and optimizing communication for desired receiver behaviors and intents. As a result, the effectiveness level remains largely untouched by modern communication systems. In this paper, we introduce the receivers' "behavior tokens," such as shares, likes, clicks, purchases, and retweets, in the LLM's training corpora to optimize content for the receivers and predict their behaviors. Other than showing similar performance to LLMs on content understanding tasks, our trained models show generalization capabilities on the behavior dimension for behavior simulation, content simulation, behavior understanding, and behavior domain adaptation. We show results on all these capabilities using a wide range of tasks on three corpora. We call these models Large Content and Behavior Models (LCBMs). Further, to spur more research on LCBMs, we release our new Content Behavior Corpus (CBC), a repository containing communicator, message, and corresponding receiver behavior (https://behavior-in-the-wild.github.io/LCBM).
****
Thomas TCK Zhang,Leonardo Felipe Toso,James Anderson,Nikolai Matni
Sample-Efficient Linear Representation Learning from Non-IID Non-Isotropic Data
A powerful concept behind much of the recent progress in machine learning is the extraction of common features across data from heterogeneous sources or tasks. Intuitively, using all of one's data to learn a common representation function benefits both computational effort and statistical generalization by leaving a smaller number of parameters to fine-tune on a given task. Toward theoretically grounding these merits, we propose a general setting of recovering linear operators M
from noisy vector measurements $y = Mx + w$, where the covariates x may be both non-i.i.d. and non-isotropic. We demonstrate that existing isotropy-agnostic meta-learning approaches incur biases on the representation update, which causes the scaling of the noise terms to lose favorable dependence on the number of source tasks. This in turn can cause the sample complexity of representation learning to be bottlenecked by the single-task data size. We introduce an adaptation, $\texttt{De-bias}$ & $\texttt{Feature-Whiten}$ ($\texttt{DFW}$), of the popular alternating minimization-descent (AMD) scheme proposed in Collins et al., (2021), and establish linear convergence to the optimal representation with noise level scaling down with the $\textit{total}$ source data size. This leads to generalization bounds on the same order as an oracle empirical risk minimizer. We verify the vital importance of $\texttt{DFW}$ on various numerical simulations. In particular, we show that vanilla alternating-minimization descent fails catastrophically even for iid, but mildly non-isotropic data.
Our analysis unifies and generalizes prior work, and provides a flexible framework for a wider range of applications, such as in controls and dynamical systems.
****
Yingtao Zhang,Haoli Bai,Haokun Lin,Jialin Zhao,Lu Hou,Carlo Vittorio Cannistraci
Plug-and-Play: An Efficient Post-training Pruning Method for Large Language Models
With the rapid growth of large language models (LLMs), there is increasing demand for memory and computation in LLMs. Recent efforts on post-training pruning of LLMs aim to reduce the model size and computation requirements, yet the perform

ance is still sub-optimal.
In this paper, we present a plug-and-play solution for post-training pruning of LLMs.
The proposed solution has two innovative components: 1) **Relative Importance and Activations (RIA)**, a new pruning metric that jointly considers the weight and activations efficiently on LLMs, and 2) **Channel Permutation**, a new approach to maximally preserves important weights under N:M sparsity.
The two proposed components can be readily combined to further enhance the N:M semi-structured pruning of LLMs.
Our empirical experiments show that RIA alone can already surpass all existing post-training pruning methods on prevalent LLMs, e.g., LLaMA ranging from 7B to 65B. Furthermore, N:M semi-structured pruning with channel permutation can even outperform the original LLaMA2-70B on zero-shot tasks, together with practical speed-up on specific hardware.
Our code is available at: https://github.com/biomedical-cybernetics/Relative-importance-and-activation-pruning
****
Gabriel Grand,Lionel Wong,Matthew Bowers,Theo X. Olausson,Muxin Liu,Joshua B. Tenenbaum,Jacob Andreas
LILO: Learning Interpretable Libraries by Compressing and Documenting Code
While large language models (LLMs) now excel at code generation, a key aspect of software development is the art of refactoring: consolidating code into libraries of reusable and readable programs. In this paper, we introduce LILO, a neurosymbolic framework that iteratively synthesizes, compresses, and documents code to build libraries tailored to particular problem domains. LILO combines LLM-guided program synthesis with recent algorithmic advances in automated refactoring from Stitch: a symbolic compression system that efficiently identifies optimal lambda abstractions across large code corpora. To make these abstractions interpretable, we introduce an auto-documentation (AutoDoc) procedure that infers natural language names and docstrings based on contextual examples of usage. In addition to improving human readability, we find that AutoDoc boosts performance by helping LILO's synthesizer to interpret and deploy learned abstractions. We evaluate LILO on three inductive program synthesis benchmarks for string editing, scene reasoning, and graphics composition. Compared to existing neural and symbolic methods—including the state-of-the-art library learning algorithm DreamCoder—LILO solves more complex tasks and learns richer libraries that are grounded in linguistic knowledge.
****
Rohin Manvi,Samar Khanna,Gengchen Mai,Marshall Burke,David B. Lobell,Stefano Ermon
GeoLLM: Extracting Geospatial Knowledge from Large Language Models
The application of machine learning (ML) in a range of geospatial tasks is increasingly common but often relies on globally available covariates such as satellite imagery that can either be expensive or lack predictive power.
Here we explore the question of whether the vast amounts of knowledge found in Internet language corpora, now compressed within large language models (LLMs), can be leveraged for geospatial prediction tasks.
We first demonstrate that LLMs embed remarkable spatial information about locations, but
naively querying LLMs using geographic coordinates alone is ineffective in predicting key indicators like population density.
We then present GeoLLM, a novel method that can effectively extract geospatial knowledge from LLMs with auxiliary map data from OpenStreetMap.
We demonstrate the utility of our approach across multiple tasks of central interest to the international community, including the measurement of population density and economic livelihoods.
Across these tasks, our method demonstrates a 70\% improvement in performance (measured using Pearson's r^2) relative to baselines that use nearest neighbors or use information directly from the prompt, and performance equal to or exceeding satellite-based benchmarks in the literature.

With GeoLLM, we observe that GPT-3.5 outperforms Llama 2 and RoBERTa by 19\% and 51\% respectively, suggesting that the performance of our method scales well with the size of the model and its pretraining dataset.
Our experiments reveal that LLMs are remarkably sample-efficient, rich in geospatial information, and robust across the globe.
Crucially, GeoLLM shows promise in mitigating the limitations of existing geospatial covariates and complementing them well.
Code is available on the project website: https://rohinmanvi.github.io/GeoLLM
****
Yichen Wu,Long-Kai Huang,Renzhen Wang,Deyu Meng,Ying Wei
Meta Continual Learning Revisited: Implicitly Enhancing Online Hessian Approximation via Variance Reduction
Regularization-based methods have so far been among the *de facto* choices for continual learning. Recent theoretical studies have revealed that these methods all boil down to relying on the Hessian matrix approximation of model weights.
However, these methods suffer from suboptimal trade-offs between knowledge transfer and forgetting due to fixed and unchanging Hessian estimations during training.
Another seemingly parallel strand of Meta-Continual Learning (Meta-CL) algorithms enforces alignment between gradients of previous tasks and that of the current task.
In this work we revisit Meta-CL and for the first time bridge it with regularization-based methods. Concretely, Meta-CL implicitly approximates Hessian in an online manner, which enjoys the benefits of timely adaptation but meantime suffers from high variance induced by random memory buffer sampling.
We are thus highly motivated to combine the best of both worlds, through the proposal of Variance Reduced Meta-CL (VR-MCL) to achieve both timely and accurate Hessian approximation.
Through comprehensive experiments across three datasets and various settings, we consistently observe that VR-MCL outperforms other SOTA methods, which further validates the effectiveness of VR-MCL.
****
Denizalp Goktas,David C. Parkes,Ian Gemp,Luke Marris,Georgios Piliouras,Romuald Elie,Guy Lever,Andrea Tacchetti
Generative Adversarial Equilibrium Solvers
We introduce the use of generative adversarial learning to compute equilibria in general game-theoretic settings, specifically the generalized Nash equilibrium (GNE) in pseudo-games, and its specific instantiation as the competitive equilibrium (CE) in Arrow-Debreu competitive economies. Pseudo-games are a generalization of games in which players' actions affect not only the payoffs of other players but also their feasible action spaces. Although the computation of GNE and CE is intractable in the worst-case, i.e., PPAD-hard, in practice, many applications only require solutions with high accuracy in expectation over a distribution of problem instances. We introduce Generative Adversarial Equilibrium Solvers (GAES): a family of generative adversarial neural networks that can learn GNE and CE from only a sample of problem instances. We provide computational and sample complexity bounds for Lipschitz-smooth function approximators in a large class of concave pseudo-games, and apply the framework to finding Nash equilibria in normal-form games, CE in Arrow-Debreu competitive economies, and GNE in an environmental economic model of the Kyoto mechanism.
****
Yin Fang,Xiaozhuan Liang,Ningyu Zhang,Kangwei Liu,Rui Huang,Zhuo Chen,Xiaohui Fan,Huajun Chen
Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models
Large Language Models (LLMs), with their remarkable task-handling capabilities and innovative outputs, have catalyzed significant advancements across a spectrum of fields. However, their proficiency within specialized domains such as biomolecular studies remains limited. To address this challenge, we introduce Mol-Instructions, a comprehensive instruction dataset designed for the biomolecular doma

in. Mol-Instructions encompasses three key components: molecule-oriented instructions, protein-oriented instructions, and biomolecular text instructions. Each component aims to improve the understanding and prediction capabilities of LLMs concerning biomolecular features and behaviors. Through extensive instruction tuning experiments on LLMs, we demonstrate the effectiveness of Mol-Instructions in enhancing large models' performance in the intricate realm of biomolecular studies, thus fostering progress in the biomolecular research community. Mol-Instructions is publicly available for ongoing research and will undergo regular updates to enhance its applicability (https://github.com/zjunlp/Mol-Instructions).
****
Hao Chen,Jindong Wang,Ankit Shah,Ran Tao,Hongxin Wei,Xing Xie,Masashi Sugiyama,Bhiksha Raj
Understanding and Mitigating the Label Noise in Pre-training on Downstream Tasks
Pre-training on large-scale datasets and then fine-tuning on downstream tasks have become a standard practice in deep learning. However, pre-training data often contain label noise that may adversely affect the generalization of the model. This paper aims to understand the nature of noise in pre-training datasets and to mitigate its impact on downstream tasks. More specifically, through extensive experiments of supervised pre-training models on synthetic noisy ImageNet-1K and YFCC15M datasets, we demonstrate that while slight noise in pre-training can benefit in-domain (ID) transfer performance, where the training and testing data share the same distribution, it always deteriorates out-of-domain (OOD) performance, where training and testing data distribution are different. We empirically verify that the reason behind is noise in pre-training shapes the feature space differently. We then propose a light-weight black-box tuning method (NMTune) to affine the feature space to mitigate the malignant effect of noise and improve generalization on both ID and OOD tasks, considering one may not be able to fully fine-tune or even access the pre-trained models. We conduct practical experiments on popular vision and language models that are pre-trained on noisy data for evaluation of our approach. Our analysis and results show the importance of this interesting and novel research direction, which we term Noisy Model Learning.
****
Soroush Abbasi Koohpayegani,Navaneet K L,Parsa Nooralinejad,Soheil Kolouri,Hamed Pirsiavash
NOLA: Networks as Linear Combination of Low Rank Random Basis
Large Language Models (LLMs) have recently gained popularity due to their impressive few-shot performance across various downstream tasks. However, fine-tuning all parameters and storing a unique model for each downstream task or domain becomes impractical because of the massive size of checkpoints (e.g., 350GB in GPT-3). Current literature, such as LoRA, showcases the potential of low-rank modifications to the original weights of an LLM, enabling efficient adaptation and storage for task-specific models. These methods can reduce the number of parameters needed to fine-tune an LLM by several orders of magnitude. Yet, these methods face two primary limitations: 1) the parameter reduction is lower-bounded by the rank one decomposition, and 2) the extent of reduction is heavily influenced by both the model architecture and the chosen rank.
For instance, in larger models, even a rank one decomposition might exceed the number of parameters truly needed for adaptation. In this paper, we introduce NOLA, which overcomes the rank one lower bound present in LoRA. It achieves this by re-parameterizing the low-rank matrices in LoRA using linear combinations of randomly generated matrices (basis) and optimizing the linear mixture coefficients only. This approach allows us to decouple the number of trainable parameters from both the choice of rank and the network architecture. We present adaptation results using GPT-2 and ViT in natural language and computer vision tasks. NOLA performs as well as, or better than models with equivalent parameter counts. Furthermore, we demonstrate that we can halve the parameters in larger models compared to LoRA with rank one, without sacrificing performance.
****
Yong Wu,Yanwei Fu,Shouyan Wang,Xinwei Sun
Doubly Robust Proximal Causal Learning for Continuous Treatments

Proximal causal learning is a powerful framework for identifying the causal effe
ct under the existence of unmeasured confounders. Within this framework, the dou
bly robust (DR) estimator was derived and has shown its effectiveness in estimat
ion, especially when the model assumption is violated. However, the current form
 of the DR estimator is restricted to binary treatments, while the treatments ca
n be continuous in many real-world applications. The primary obstacle to continu
ous treatments resides in the delta function present in the original DR estimato
r, making it infeasible in causal effect estimation and introducing a heavy comp
utational burden in nuisance function estimation. To address these challenges, w
e propose a kernel-based DR estimator that can well handle continuous treatments
 for proximal causal learning. Equipped with its smoothness, we show that its or
acle form is a consistent approximation of the influence function. Further, we p
ropose a new approach to efficiently solve the nuisance functions. We then provi
de a comprehensive convergence analysis in terms of the mean square error. We de
monstrate the utility of our estimator on synthetic datasets and real-world appl
ications.
****
Christian Gumbsch,Noor Sajid,Georg Martius,Martin V. Butz
Learning Hierarchical World Models with Adaptive Temporal Abstractions from Disc
rete Latent Dynamics
Hierarchical world models can significantly improve model-based reinforcement le
arning (MBRL) and planning by enabling reasoning across multiple time scales. No
netheless, the majority of state-of-the-art MBRL methods employ flat, non-hierar
chical models. We propose Temporal Hierarchies from Invariant Context Kernels (T
HICK), an algorithm that learns a world model hierarchy via discrete latent dyna
mics. The lower level of THICK updates parts of its latent state sparsely in tim
e, forming invariant contexts. The higher level exclusively predicts situations
involving context changes. Our experiments demonstrate that THICK learns categor
ical, interpretable, temporal abstractions on the high level, while maintaining
precise low-level predictions. Furthermore, we show that the emergent hierarchic
al predictive model seamlessly enhances the abilities of MBRL or planning method
s. We believe that THICK contributes to the further development of hierarchical
agents capable of more sophisticated planning and reasoning abilities.
****
Chendi Qian,Andrei Manolache,Kareem Ahmed,Zhe Zeng,Guy Van den Broeck,Mathias Ni
epert,Christopher Morris
Probabilistically Rewired Message-Passing Neural Networks
Message-passing graph neural networks (MPNNs) emerged as powerful tools for proc
essing graph-structured input. However, they operate on a fixed input graph stru
cture, ignoring potential noise and missing information. Furthermore, their loca
l aggregation mechanism can lead to problems such as over-squashing and limited
expressive power in capturing relevant graph structures. Existing solutions to t
hese challenges have primarily relied on heuristic methods, often disregarding t
he underlying data distribution. Hence, devising principled approaches for learn
ing to infer graph structures relevant to the given prediction task remains an o
pen challenge. In this work, leveraging recent progress in exact and differentia
ble k-subset sampling, we devise probabilistically rewired MPNNs (PR-MPNNs), whi
ch learn to add relevant edges while omitting less beneficial ones. For the firs
t time, our theoretical analysis explores how PR-MPNNs enhance expressive power,
 and we identify precise conditions under which they outperform purely randomize
d approaches. Empirically, we demonstrate that our approach effectively mitigate
s issues like over-squashing and under-reaching. In addition, on established rea
l-world datasets, our method exhibits competitive or superior predictive perform
ance compared to traditional MPNN models and recent graph transformer architectu
res.
****
Katherine Hermann,Hossein Mobahi,Thomas FEL,Michael Curtis Mozer
On the Foundations of Shortcut Learning
Deep-learning models can extract a rich assortment of features from data. Which
features a model uses depends not only on *predictivity*---how reliably a featur

e indicates training-set labels---but also on *availability*---how easily the feature can be extracted from inputs. The literature on shortcut learning has noted examples in which models privilege one feature over another, for example texture over shape and image backgrounds over foreground objects. Here, we test hypotheses about which input properties are more available to a model, and systematically study how predictivity and availability interact to shape models' feature use. We construct a minimal, explicit generative framework for synthesizing classification datasets with two latent features that vary in predictivity and in factors we hypothesize to relate to availability, and we quantify a model's shortcut bias---its over-reliance on the shortcut (more available, less predictive) feature at the expense of the core (less available, more predictive) feature. We find that linear models are relatively unbiased, but introducing a single hidden layer with ReLU or Tanh units yields a bias. Our empirical findings are consistent with a theoretical account based on Neural Tangent Kernels. Finally, we study how models used in practice trade off predictivity and availability in naturalistic datasets, discovering availability manipulations which increase models' degree of shortcut bias. Taken together, these findings suggest that the propensity to learn shortcut features is a fundamental characteristic of deep nonlinear architectures warranting systematic study given its role in shaping how models solve tasks.
****
Sheng Li,Chao Wu,Ao Li,Yanzhi Wang,Xulong Tang,Geng Yuan
Waxing-and-Waning: a Generic Similarity-based Framework for Efficient Self-Supervised Learning
Deep Neural Networks (DNNs), essential for diverse applications such as visual recognition and eldercare, often require a large amount of labeled data for training, making widespread deployment of DNNs a challenging task. Self-supervised learning (SSL) emerges as a promising approach, which leverages inherent patterns within data through diverse augmentations to train models without explicit labels. However, while SSL has shown notable advancements in accuracy, its high computation costs remain a daunting impediment, particularly for resource-constrained platforms. To address this problem, we introduce SimWnW, a similarity-based efficient self-supervised learning framework. By strategically removing less important regions in augmented images and feature maps, SimWnW not only reduces computation costs but also eliminates irrelevant features that might slow down the learning process, thereby accelerating model convergence. The experimental results show that SimWnW effectively reduces the amount of computation costs in self-supervised model training without compromising accuracy. Specifically, SimWnW yields up to 54\% and 51\% computation savings in training from scratch and transfer learning tasks, respectively.
****
Danny Halawi,Jean-Stanislas Denain,Jacob Steinhardt
Overthinking the Truth: Understanding how Language Models Process False Demonstrations
Modern language models can imitate complex patterns through few-shot learning, enabling them to complete challenging tasks without fine-tuning. However, imitation can also lead models to reproduce inaccuracies or harmful content if present in the context. We study harmful imitation through the lens of a model’s internal representations, and identify two related phenomena: overthinking and false induction heads. The first phenomenon, overthinking, appears when we decode predictions from intermediate layers, given correct vs. incorrect few-shot demonstrations. At early layers, both demonstrations induce similar model behavior, but the behavior diverges sharply at some “critical layer”, after which the accuracy given incorrect demonstrations progressively decreases. The second phenomenon, false induction heads, are a possible mechanistic cause of overthinking: these are heads in late layers that attend to and copy false information from previous demonstrations, and whose ablation reduces overthinking. Beyond scientific understanding, our results suggest that studying intermediate model computations could be a promising avenue for understanding and guarding against harmful model behaviors.

****
Yung-Sung Chuang,Yujia Xie,Hongyin Luo,Yoon Kim,James R. Glass,Pengcheng He
DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Model
s
Despite their impressive capabilities, large language models (LLMs) are prone to
 hallucinations, i.e., generating content that deviates from facts seen during p
retraining. We propose a simple decoding strategy for reducing hallucinations wi
th pretrained LLMs that does not require conditioning on retrieved external know
ledge nor additional fine-tuning. Our approach obtains the next-token distributi
on by contrasting the differences in logits obtained from projecting the later l
ayers versus earlier layers to the vocabulary space, exploiting the fact that fa
ctual knowledge in an LLMs has generally been shown to be localized to particula
r transformer layers. We find that this **D**ecoding by C**o**ntrasting **La**ye
rs (DoLa) approach is able to better surface factual knowledge and reduce the ge
neration of incorrect facts. DoLa consistently improves the truthfulness across
multiple choices tasks and open-ended generation tasks, for example improving th
e performance of LLaMA family models on TruthfulQA by 12-17% absolute points, de
monstrating its potential in making LLMs reliably generate truthful facts.
****
Zheng Ding,Mengqi Zhang,Jiajun Wu,Zhuowen Tu
Patched Denoising Diffusion Models For High-Resolution Image Synthesis
We propose an effective denoising diffusion model for generating high-resolution
 images (e.g., 1024$\times$512), trained on small-size image patches (e.g., 64$\
times$64). We name our algorithm Patch-DM, in which a new feature collage strate
gy is designed to avoid the boundary artifact when synthesizing large-size image
s. Feature collage systematically crops and combines partial features of the nei
ghboring patches to predict the features of a shifted image patch, allowing the
seamless generation of the entire image due to the overlap in the patch feature
space. Patch-DM produces high-quality image synthesis results on our newly colle
cted dataset of nature images (1024$\times$512), as well as on standard benchmar
ks of LHQ(1024$\times$ 1024), FFHQ(1024$\times$ 1024) and on other datasets with
 smaller sizes (256$\times$256), including LSUN-Bedroom, LSUN-Church, and FFHQ.
We compare our method with previous patch-based generation methods and achieve s
tate-of-the-art FID scores on all six datasets. Further, Patch-DM also reduces m
emory complexity compared to the classic diffusion models.
****
Ziheng Cheng,Xinmeng Huang,Pengfei Wu,Kun Yuan
Momentum Benefits Non-iid Federated Learning Simply and Provably
Federated learning is a powerful paradigm for large-scale machine learning, but
it
faces significant challenges due to unreliable network connections, slow commu-
nication, and substantial data heterogeneity across clients. FedAvg and SCAFFOLD
 are two prominent algorithms to address these challenges. In particular,
FedAvg employs multiple local updates before communicating with a central
server, while SCAFFOLD maintains a control variable on each client to compen-
sate for “client drift” in its local updates. Various methods have been proposed
to enhance the convergence of these two algorithms, but they either make imprac-
tical adjustments to algorithmic structure, or rely on the assumption of bounded
data heterogeneity. This paper explores the utilization of momentum to enhance
the performance of FedAvg and SCAFFOLD. When all clients participate in the
training process, we demonstrate that incorporating momentum allows FedAvg
to converge without relying on the assumption of bounded data heterogeneity even
using a constant local learning rate. This is novel and fairly suprising as exis
ting
analyses for FedAvg require bounded data heterogeneity even with diminishing
local learning rates. In partial client participation, we show that momentum en-
ables SCAFFOLD to converge provably faster without imposing any additional
assumptions. Furthermore, we use momentum to develop new variance-reduced
extensions of FedAvg and SCAFFOLD, which exhibit state-of-the-art conver-
gence rates. Our experimental results support all theoretical findings.

****
Krzysztof Kacprzyk,Tennison Liu,Mihaela van der Schaar
Towards Transparent Time Series Forecasting
Transparent machine learning (ML) models are essential for ensuring interpretability and trustworthiness in decision-making systems, particularly in high-stakes domains such as healthcare, finance, and criminal justice. While transparent machine learning models have been proposed for classification and regression, time series forecasting presents some unique challenges for ensuring transparency. In particular, currently used bottom-up approaches that focus on the values of the time series at specific time points (usually regularly spaced) do not provide a holistic understanding of the entire time series. This limits the applicability of ML in many critical areas. To open up these domains for ML, we propose a top-down framework of bi-level transparency, which involves understanding the higher-level trends and the lower-level properties of the predicted time series. Applying this framework, we develop TIMEVIEW, a transparent ML model for time series forecasting based on static features, complemented with an interactive visualization tool. Through a series of experiments, we demonstrate the efficacy and interpretability of our approach, paving the way for more transparent and reliable applications of ML in various domains.
****
Li Ren,Chen Chen,Liqiang Wang,Kien A. Hua
Learning Semantic Proxies from Visual Prompts for Parameter-Efficient Fine-Tuning in Deep Metric Learning
Deep Metric Learning (DML) has long attracted the attention of the machine learning community as a key objective. Existing solutions concentrate on fine-tuning the pre-trained models on conventional image datasets. As a result of the success of recent pre-trained models derived from larger-scale datasets, it is challenging to adapt the model to the DML tasks in the local data domain while retaining the previously gained knowledge. In this paper, we investigate parameter-efficient methods for fine-tuning the pre-trained model for DML tasks. In particular, we propose a novel and effective framework based on learning Visual Prompts (VPT) in the pre-trained Vision Transformers (ViT). Based on the conventional proxy-based DML paradigm, we augment the proxy by incorporating the semantic information from the input image and the ViT, in which we optimize the visual prompts for each class. We demonstrate that our new approximations with semantic information are superior to representative capabilities, thereby improving metric learning performance. We conduct extensive experiments to demonstrate that our proposed framework is superior and efficient by evaluating popular DML benchmarks. In particular, we demonstrate that our fine-tuning method achieves comparable or even better performance than recent state-of-the-art full fine-tuning works of DML while tuning only a small percentage of total parameters.
****
Chengzhi Cao,Yinghao Fu,Sheng Xu,Ruimao Zhang,Shuang Li
Enhancing Human-AI Collaboration Through Logic-Guided Reasoning
We present a systematic framework designed to enhance human-robot perception and collaboration through the integration of logical rules and Theory of Mind (ToM). Logical rules provide interpretable predictions and generalize well across diverse tasks, making them valuable for learning and decision-making. Leveraging the ToM for understanding others' mental states, our approach facilitates effective collaboration. In this paper, we employ logic rules derived from observational data to infer human goals and guide human-like agents. These rules are treated as latent variables, and a rule encoder is trained alongside a multi-agent system in the robot's mind. We assess the posterior distribution of latent rules using learned embeddings, representing entities and relations. Confidence scores for each rule indicate their consistency with observed data. Then, we employ a hierarchical reinforcement learning model with ToM to plan robot actions for assisting humans. Extensive experiments validate each component of our framework, and results on multiple benchmarks demonstrate that our model outperforms the majority of existing approaches.
****

Feng Lu,Lijun Zhang,Xiangyuan Lan,Shuting Dong,Yaowei Wang,Chun Yuan
Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition
Recent studies show that vision models pre-trained in generic visual learning tasks with large-scale data can provide useful feature representations for a wide range of visual perception problems. However, few attempts have been made to exploit pre-trained foundation models in visual place recognition (VPR). Due to the inherent difference in training objectives and data between the tasks of model pre-training and VPR, how to bridge the gap and fully unleash the capability of pre-trained models for VPR is still a key issue to address. To this end, we propose a novel method to realize seamless adaptation of pre-trained models for VPR. Specifically, to obtain both global and local features that focus on salient landmarks for discriminating places, we design a hybrid adaptation method to achieve both global and local adaptation efficiently, in which only lightweight adapters are tuned without adjusting the pre-trained model. Besides, to guide effective adaptation, we propose a mutual nearest neighbor local feature loss, which ensures proper dense local features are produced for local matching and avoids time-consuming spatial verification in re-ranking. Experimental results show that our method outperforms the state-of-the-art methods with less training data and training time, and uses about only 3% retrieval runtime of the two-stage VPR methods with RANSAC-based spatial verification. It ranks 1st on the MSLS challenge leaderboard (at the time of submission). The code is released at https://github.com/Lu-Feng/SelaVPR.
****
Zhiwei Tang,Dmitry Rybin,Tsung-Hui Chang
Zeroth-Order Optimization Meets Human Feedback: Provable Learning via Ranking Oracles
In this study, we delve into an emerging optimization challenge involving a black-box objective function that can only be gauged via a ranking oracle—a situation frequently encountered in real-world scenarios, especially when the function is evaluated by human judges. A prominent instance of such a situation is Reinforcement Learning with Human Feedback (RLHF), an approach recently employed to enhance the performance of Large Language Models (LLMs) using human guidance [Ouyang et al. 2022, Liu et al. 2023, OpenAI et al. 2022, Bai et al. 2022]. We introduce ZO-RankSGD, an innovative zeroth-order optimization algorithm designed to tackle this optimization problem, accompanied by theoretical assurances. Our algorithm utilizes a novel rank-based random estimator to determine the descent direction and guarantees convergence to a stationary point. Moreover, ZO-RankSGD is readily applicable to policy optimization problems in Reinforcement Learning (RL), particularly when only ranking oracles for the episode reward are available. Last but not least, we demonstrate the effectiveness of ZO-RankSGD in a novel application: improving the quality of images generated by a diffusion generative model with human ranking feedback. Throughout experiments, we found that ZO-RankSGD can significantly enhance the detail of generated images with only a few rounds of human feedback. Overall, our work advances the field of zeroth-order optimization by addressing the problem of optimizing functions with only ranking feedback, and offers a new and effective approach for aligning Artificial Intelligence (AI) with human intentions.
****
Ronak Mehta,Vincent Roulet,Krishna Pillutla,Zaid Harchaoui
Distributionally Robust Optimization with Bias and Variance Reduction
We consider the distributionally robust optimization (DRO) problem, wherein a learner optimizes the worst-case empirical risk achievable by reweighing the observed training examples. We present Prospect, a stochastic gradient-based algorithm that only requires tuning a single learning rate hyperparameter, and prove that it enjoys linear convergence for smooth regularized losses. This contrasts with previous algorithms that either require tuning multiple hyperparameters or potentially fail to converge due to biased gradient estimates or inadequate regularization. Empirically, we show that Prospect can converge 2-3x faster than baselines such as SGD and stochastic saddle-point methods on distribution shift and fairness benchmarks spanning tabular, vision, and language domains.

****
Philippe Chlenski,Ethan Turok,Antonio Khalil Moretti,Itsik Pe'er
Fast Hyperboloid Decision Tree Algorithms
Hyperbolic geometry is gaining traction in machine learning due to its capacity to effectively capture hierarchical structures in real-world data. Hyperbolic spaces, where neighborhoods grow exponentially, offer substantial advantages and have consistently delivered state-of-the-art results across diverse applications. However, hyperbolic classifiers often grapple with computational challenges. Methods reliant on Riemannian optimization frequently exhibit sluggishness, stemming from the increased computational demands of operations on Riemannian manifolds. In response to these challenges, we present HyperDT, a novel extension of decision tree algorithms into hyperbolic space. Crucially, HyperDT eliminates the need for computationally intensive Riemannian optimization, numerically unstable exponential and logarithmic maps, or pairwise comparisons between points by leveraging inner products to adapt Euclidean decision tree algorithms to hyperbolic space. Our approach is conceptually straightforward and maintains constant-time decision complexity while mitigating the scalability issues inherent in high-dimensional Euclidean spaces. Building upon HyperDT, we introduce HyperRF, a hyperbolic random forest model. Extensive benchmarking across diverse datasets underscores the superior performance of these models, providing a swift, precise, accurate, and user-friendly toolkit for hyperbolic data analysis.
****
Joonhun Lee,Myeongho Jeon,Myungjoo Kang,Kyunghyun Park
Feature-aligned N-BEATS with Sinkhorn divergence
We propose Feature-aligned N-BEATS as a domain-generalized time series forecasting model. It is a nontrivial extension of N-BEATS with doubly residual stacking principle (Oreshkin et al. [45]) into a representation learning framework. In particular, it revolves around marginal feature probability measures induced by the intricate composition of residual and feature extracting operators of N-BEATS in each stack and aligns them stack-wise via an approximate of an optimal transport distance referred to as the Sinkhorn divergence. The training loss consists of an empirical risk minimization from multiple source domains, i.e., forecasting loss, and an alignment loss calculated with the Sinkhorn divergence, which allows the model to learn invariant features stack-wise across multiple source data sequences while retaining N-BEATS’s interpretable design and forecasting power. Comprehensive experimental evaluations with ablation studies are provided and the corresponding results demonstrate the proposed model’s forecasting and generalization capabilities.
****
Mouxing Yang,Yunfan Li,Changqing Zhang,Peng Hu,Xi Peng
Test-time Adaption against Multi-modal Reliability Bias
Test-time adaptation (TTA) has emerged as a new paradigm for reconciling distribution shifts across domains without accessing source data.
However, existing TTA methods mainly concentrate on uni-modal tasks, overlooking the complexity in multi-modal scenarios.
In this paper, we delve into the multi-modal test-time adaptation and reveal a new challenge named reliability bias.
Different from the definition of traditional distribution shifts, reliability bias refers to the information discrepancies across different modalities derived from intra-modal distribution shifts.
To solve the challenge, we propose a novel method, dubbed REliable fusion and robust ADaptation (READ).
On the one hand, unlike the existing TTA paradigm that mainly repurposes the normalization layers, READ employs a new paradigm that modulates the attention between modalities in a self-adaptive way, supporting reliable fusion against reliability bias.
On the other hand, READ adopts a novel objective function for robust multi-modal adaptation, where the contributions of confident predictions could be amplified and the negative impacts of noisy predictions could be mitigated.
Moreover, we introduce two new benchmarks to facilitate comprehensive evaluation

s of multi-modal TTA under reliability bias.
Extensive experiments on the benchmarks verify the effectiveness of our method against multi-modal reliability bias.
The code and benchmarks are available on https://github.com/XLearning-SCU/2024-ICLR-READ.
****
Qihao Liu,Adam Kortylewski,Yutong Bai,Song Bai,Alan Yuille
Discovering Failure Modes of Text-guided Diffusion Models via Adversarial Search
Text-guided diffusion models (TDMs) are widely applied but can fail unexpectedly. Common failures include: _(i)_ natural-looking text prompts generating images with the wrong content, or _(ii)_ different random samples of the latent variables that generate vastly different, and even unrelated, outputs despite being conditioned on the same text prompt. In this work, we aim to study and understand the failure modes of TDMs in more detail. To achieve this, we propose SAGE, the first adversarial search method on TDMs that systematically explores the discrete prompt space and the high-dimensional latent space, to automatically discover undesirable behaviors and failure cases in image generation. We use image classifiers as surrogate loss functions during searching, and employ human inspections to validate the identified failures. For the first time, our method enables efficient exploration of both the discrete and intricate human language space and the challenging latent space, overcoming the gradient vanishing problem. Then, we demonstrate the effectiveness of SAGE on five widely used generative models and reveal four typical failure modes that have not been systematically studied before: (1) We find a variety of natural text prompts that generate images failing to capture the semantics of input texts. We further discuss the underlying causes and potential solutions based on the results. (2) We find regions in the latent space that lead to distorted images independent of the text prompt, suggesting that parts of the latent space are not well-structured. (3) We also find latent samples that result in natural-looking images unrelated to the text prompt, implying a possible misalignment between the latent and prompt spaces. (4) By appending a single adversarial token embedding to any input prompts, we can generate a variety of specified target objects, with minimal impact on CLIP scores, demonstrating the fragility of language representations.
****
Saurabh Garg,Mehrdad Farajtabar,Hadi Pouransari,Raviteja Vemulapalli,Sachin Mehta,Oncel Tuzel,Vaishaal Shankar,Fartash Faghri
TiC-CLIP: Continual Training of CLIP Models
Keeping large foundation models up to date on latest data is inherently expensive. To avoid the prohibitive costs of constantly retraining, it is imperative to continually train these models. This problem is exacerbated by the lack of any large scale continual learning benchmarks or baselines. We introduce the first set of web-scale Time-Continual (TiC) benchmarks for training vision-language models: TiC-DataComp, TiC-YFCC, and TiC-Redcaps. TiC-DataComp, our largest dataset, contains over 12.7B timestamped image-text pairs spanning 9 years (2014-2022). We first use our benchmarks to curate various dynamic evaluations to measure temporal robustness of existing models. We show OpenAI's CLIP (trained on data up to 2020) loses $\approx 8\%$ zero-shot accuracy on our curated retrieval task from 2021-2022 compared with more recently trained models in OpenCLIP repository. We then study how to efficiently train models on time-continuous data. We demonstrate that a simple rehearsal-based approach that continues training from the last checkpoint and replays old data reduces compute by $2.5\times$ when compared to the standard practice of retraining from scratch. Code is available at https://github.com/apple/ml-tic-clip.
****
Xi Weng,Yunhao Ni,Tengwei Song,Jie Luo,Rao Muhammad Anwer,Salman Khan,Fahad Khan,Lei Huang
Modulate Your Spectrum in Self-Supervised Learning
Whitening loss offers a theoretical guarantee against feature collapse in self-supervised learning (SSL) with joint embedding architectures. Typically, it involves a hard whitening approach, transforming the embedding and applying loss to t

he whitened output. In this work, we introduce Spectral Transformation (ST), a framework to modulate the spectrum of embedding and to seek for functions beyond whitening that can avoid dimensional collapse. We show that whitening is a special instance of ST by definition, and our empirical investigations unveil other ST instances capable of preventing collapse. Additionally, we propose a novel ST instance named IterNorm with trace loss (INTL). Theoretical analysis confirms INTL's efficacy in preventing collapse and modulating the spectrum of embedding toward equal-eigenvalues during optimization. Our experiments on ImageNet classification and COCO object detection demonstrate INTL's potential in learning superior representations. The code is available at https://github.com/winci-ai/INTL.
****
Aoran Wang,Jun Pang
Structural Inference with Dynamics Encoding and Partial Correlation Coefficients
This paper introduces a novel approach to structural inference, combining a variational dynamics encoder with partial correlation coefficients.
In contrast to prior methods, our approach leverages variational inference to encode node dynamics within latent variables, and structural reconstruction relies on the calculation of partial correlation coefficients derived from these latent variables.
This unique design endows our method with scalability and extends its applicability to both one-dimensional and multi-dimensional feature spaces.
Furthermore, by reorganizing latent variables according to temporal steps, our approach can effectively reconstruct directed graph structures.
We validate our method through extensive experimentation on twenty datasets from a benchmark dataset and biological networks.
Our results showcase the superior scalability, accuracy, and versatility of our proposed approach compared to existing methods.
Moreover, experiments conducted on noisy data affirm the robustness of our method.
****
Seong Jin Cho,Gwangsu Kim,Junghyun Lee,Jinwoo Shin,Chang D. Yoo
Querying Easily Flip-flopped Samples for Deep Active Learning
Active learning, a paradigm within machine learning, aims to select and query unlabeled data to enhance model performance strategically. A crucial selection strategy leverages the model's predictive uncertainty, reflecting the informativeness of a data point. While the sample's distance to the decision boundary intuitively measures predictive uncertainty, its computation becomes intractable for complex decision boundaries formed in multiclass classification tasks. This paper introduces the *least disagree metric* (LDM), the smallest probability of predicted label disagreement. We propose an asymptotically consistent estimator for LDM under mild assumptions. The estimator boasts computational efficiency and straightforward implementation for deep learning models using parameter perturbation. The LDM-based active learning algorithm queries unlabeled data with the smallest LDM, achieving state-of-the-art *overall* performance across various datasets and deep architectures, as demonstrated by the experimental results.
****
Philip Amortila,Dylan J Foster,Nan Jiang,Ayush Sekhari,Tengyang Xie
Harnessing Density Ratios for Online Reinforcement Learning
The theories of offline and online reinforcement learning, despite having evolved in parallel, have begun to show signs of the possibility for a unification, with algorithms and analysis techniques for one setting often having natural counterparts in the other. However, the notion of *density ratio modeling*, an emerging paradigm in offline RL, has been largely absent from online RL, perhaps for good reason: the very existence and boundedness of density ratios relies on access to an exploratory dataset with good coverage, but the core challenge in online RL is to collect such a dataset without having one to start.

In this work we show---perhaps surprisingly---that density ratio-based algorithms have online counterparts. Assuming only the existence of an exploratory distribution with good coverage, a structural condition known as *coverability* (Xie

et al., 2023), we give a new algorithm (GLOW) that uses density ratio realizabil
ity and value function realizability to perform sample-efficient online explorat
ion. GLOW addresses unbounded density ratios via careful use of truncation, and
combines this with optimism to guide exploration. GLOW is computationally ineffi
cient; we complement it with a more efficient counterpart, HyGLOW, for the Hybri
d RL setting (Song et al., 2023) wherein online RL is augmented with additional
offline data. HyGLOW is derived as a special case of a more general meta-algorit
hm that provides a provable black-box reduction from hybrid RL to offline RL, wh
ich may be of independent interest.
****
Sumeet Batra,Bryon Tjanaka,Matthew Christopher Fontaine,Aleksei Petrenko,Stefano
s Nikolaidis,Gaurav S. Sukhatme
Proximal Policy Gradient Arborescence for Quality Diversity Reinforcement Learni
ng
Training generally capable agents that thoroughly explore their environment and
learn new and diverse skills is a long-term goal of robot learning. Quality Dive
rsity
Reinforcement Learning (QD-RL) is an emerging research area that blends the
best aspects of both fields – Quality Diversity (QD) provides a principled form
of exploration and produces collections of behaviorally diverse agents, while
Reinforcement Learning (RL) provides a powerful performance improvement
operator enabling generalization across tasks and dynamic environments. Existing
QD-RL approaches have been constrained to sample efficient, deterministic off-
policy RL algorithms and/or evolution strategies and struggle with highly stocha
stic
environments. In this work, we, for the first time, adapt on-policy RL, specific
ally
Proximal Policy Optimization (PPO), to the Differentiable Quality Diversity (DQD
)
framework and propose several changes that enable efficient optimization and
discovery of novel skills on high-dimensional, stochastic robotics tasks. Our ne
w
algorithm, Proximal Policy Gradient Arborescence (PPGA), achieves state-of-
the-art results, including a 4x improvement in best reward over baselines on the
challenging humanoid domain.
****
Panagiotis Eustratiadis,■ukasz Dudziak,Da Li,Timothy Hospedales
Neural Fine-Tuning Search for Few-Shot Learning
In few-shot recognition, a classifier that has been trained on one set of classe
s is required to rapidly adapt and generalize to a disjoint, novel set of classe
s. To that end, recent studies have shown the efficacy of fine-tuning with caref
ully-crafted adaptation architectures. However this raises the question of: How
can one design the optimal adaptation strategy? In this paper, we study this que
stion through the lens of neural architecture search (NAS). Given a pre-trained
neural network, our algorithm discovers the optimal arrangement of adapters, whi
ch layers to keep frozen, and which to fine-tune. We demonstrate the generality
of our NAS method by applying it to both residual networks and vision transforme
rs and report state-of-the-art performance on Meta-Dataset and Meta-Album.
****
Justin Dumouchelle,Esther Julien,Jannis Kurtz,Elias Boutros Khalil
Neur2RO: Neural Two-Stage Robust Optimization
Robust optimization provides a mathematical framework for modeling and solving d
ecision-making problems under worst-case uncertainty. This work addresses two-s
tage robust optimization (2RO) problems (also called *adjustable robust optimiza
tion*), wherein first-stage and second-stage decisions are made before and after
 uncertainty is realized, respectively. This results in a nested min-max-min op
timization problem which is extremely challenging computationally, especially wh
en the decisions are discrete. We propose Neur2RO, an efficient machine learnin
g-driven instantiation of column-and-constraint generation (CCG), a classical it
erative algorithm for 2RO. Specifically, we learn to estimate the value functio

n of the second-stage problem via a novel neural network architecture that is ea
sy to optimize over by design. Embedding our neural network into CCG yields high
-quality solutions quickly as evidenced by experiments on two 2RO benchmarks, kn
apsack and capital budgeting. For knapsack, Neur2RO finds solutions that are wit
hin roughly 2% of the best-known values in a few seconds compared to the three
 hours of the state-of-the-art exact branch-and-price algorithm; for larger and
more complex instances, Neur2RO finds even better solutions. For capital budgeti
ng, Neur2RO outperforms three variants of the k-adaptability algorithm, partic
ularly on the largest instances, with a 10 to 100-fold reduction in solution
 time. Our code and data are available at https://github.com/khalil-research/Neu
r2RO.
****
Mitja Nikolaus
Emergent Communication with Conversational Repair
Research on conversation has put emphasis on the importance of a multi-level com
munication system, in which the interlocutors aim to establish and maintain comm
on ground. In natural conversations, repair mechanisms such as clarification req
uests are frequently used to improve mutual understanding.
Here we explore the effects of conversational repair on languages emerging in si
gnaling games. We extend the basic Lewis signaling game setup with a feedback ch
annel that allows for the transmission of messages backwards from the receiver t
o the sender. Further, we add noise to the communication channel so that repair
mechanisms become necessary for optimal performance.

We find that languages emerging in setups with feedback channel are less composi
tional.
However, the models still achieve a substantially higher generalization performa
nce in conditions with noise, putting to question the role of compositionality f
or generalization.
These findings generalize also to a more realistic case involving a guessing gam
e with naturalistic images.

More broadly speaking, this study provides an important step towards the creatio
n of signaling games that more closely resemble the conditions under which human
 languages emerged.
****
James Harrison,John Willes,Jasper Snoek
Variational Bayesian Last Layers
We introduce a deterministic variational formulation for training Bayesian last
layer neural networks. This yields a sampling-free, single-pass model and loss t
hat effectively improves uncertainty estimation. Our variational Bayesian last l
ayer (VBLL) can be trained and evaluated with only quadratic complexity in last
layer width, and is thus (nearly) computationally free to add to standard archit
ectures. We experimentally investigate VBLLs, and show that they improve predict
ive accuracy, calibration, and out of distribution detection over baselines acro
ss both regression and classification. Finally, we investigate combining VBLL la
yers with variational Bayesian feature learning, yielding a lower variance colla
psed variational inference method for Bayesian neural networks.
****
Zhibin Gou,Zhihong Shao,Yeyun Gong,yelong shen,Yujiu Yang,Nan Duan,Weizhu Chen
CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing
Recent developments in large language models (LLMs) have been impressive. Howeve
r, these models sometimes show inconsistencies and problematic behavior, such as
 hallucinating facts, generating flawed code, or creating offensive and toxic co
ntent. Unlike these models, humans typically utilize external tools to cross-che
ck and refine their initial content, like using a search engine for fact-checkin
g, or a code interpreter for debugging. Inspired by this observation, we introdu
ce a framework called CRITIC that allows LLMs, which are essentially “black boxe
s” to validate and progressively amend their own outputs in a manner similar to
human interaction with tools. More specifically, starting with an initial output

, CRITIC interacts with appropriate tools to evaluate certain aspects of the tex t, and then revises the output based on the feedback obtained during this valida tion process. Comprehensive evaluations involving free-form question answering, mathematical program synthesis, and toxicity reduction demonstrate that CRITIC c onsistently enhances the performance of LLMs. Meanwhile, our research highlights the crucial importance of external feedback in promoting the ongoing self-impro vement of LLMs.
****
Khai Nguyen,Nhat Ho
Sliced Wasserstein Estimation with Control Variates
The sliced Wasserstein (SW) distances between two probability measures are defin ed as the expectation of the Wasserstein distance between two one-dimensional pr ojections of the two measures. The randomness comes from a projecting direction that is used to project the two input measures to one dimension. Due to the intr actability of the expectation, Monte Carlo integration is performed to estimate the value of the SW distance. Despite having various variants, there has been no prior work that improves the Monte Carlo estimation scheme for the SW distance in terms of controlling its variance. To bridge the literature on variance reduc tion and the literature on the SW distance, we propose computationally efficient control variates to reduce the variance of the empirical estimation of the SW d istance. The key idea is to first find Gaussian approximations of projected one-dimensional measures, then we utilize the closed-form of the Wasserstein-2 dista nce between two Gaussian distributions to design the control variates. In partic ular, we propose using a lower bound and an upper bound of the Wasserstein-2 dis tance between two fitted Gaussians as two computationally efficient control vari ates. We empirically show that the proposed control variate estimators can help to reduce the variance considerably when comparing measures over images and poin t-clouds. Finally, we demonstrate the favorable performance of the proposed cont rol variate estimators in gradient flows to interpolate between two point-clouds and in deep generative modeling on standard image datasets, such as CIFAR10 and CelebA.
****
Xuandong Zhao,Prabhanjan Vijendra Ananth,Lei Li,Yu-Xiang Wang
Provable Robust Watermarking for AI-Generated Text
We study the problem of watermarking large language models (LLMs) generated text — one of the most promising approaches for addressing the safety challenges of LLM usage. In this paper, we propose a rigorous theoretical framework to quantif y the effectiveness and robustness of LLM watermarks. We propose a robust and hi gh-quality watermark method, Unigram-Watermark, by extending an existing approac h with a simplified fixed grouping strategy. We prove that our watermark method enjoys guaranteed generation quality, correctness in watermark detection, and is robust against text editing and paraphrasing. Experiments on three varying LLMs and two datasets verify that our Unigram-Watermark achieves superior detection accuracy and comparable generation quality in perplexity, thus promoting the res ponsible use of LLMs.
****
Shaokun Zhang,Xiaobo Xia,Zhaoqing Wang,Ling-Hao Chen,Jiale Liu,Qingyun Wu,Tongli ang Liu
IDEAL: Influence-Driven Selective Annotations Empower In-Context Learners in Lar ge Language Models
In-context learning is a promising paradigm that utilizes in-context examples as prompts for the predictions of large language models. These prompts are crucial for achieving strong performance. However, since the prompts need to be sampled from a large volume of annotated examples, finding the right prompt may result in high annotation costs. To address this challenge, this paper introduces an in fluence-driven selective annotation method that aims to minimize annotation cost s while improving the quality of in-context examples. The essence of our method is to select a pivotal subset from a large-scale unlabeled data pool to annotate for the subsequent sampling of prompts. Specifically, a directed graph is first constructed to represent unlabeled data. Afterward, the influence of candidate

unlabeled subsets is quantified with a diffusion process. A simple yet effective greedy algorithm for unlabeled data selection is lastly introduced. It iteratively selects the data if it provides a maximum marginal gain with respect to quantified influence. Compared with previous efforts on selective annotations, our influence-driven method works in an end-to-end manner, avoids an intractable explicit balance between data diversity and representativeness, and enjoys theoretical support. Experiments confirm the superiority of the proposed method on various benchmarks, achieving better performance under lower time consumption during subset selection.
****
Guan-Horng Liu,Yaron Lipman,Maximilian Nickel,Brian Karrer,Evangelos Theodorou,Ricky T. Q. Chen
Generalized Schrödinger Bridge Matching
Modern distribution matching algorithms for training diffusion or flow models directly prescribe the time evolution of the marginal distributions between two boundary distributions. In this work, we consider a generalized distribution matching setup, where these marginals are only implicitly described as a solution to some task-specific objective function. The problem setup, known as the Generalized Schrödinger Bridge (GSB), appears prevalently in many scientific areas both within and without machine learning. We propose Generalized Schödinger Bridge Matching (GSBM), a new matching algorithm inspired by recent advances, generalizing them beyond kinetic energy minimization and to account for nonlinear state costs. We show that such a generalization can be cast as solving conditional stochastic optimal control, for which efficient variational approximations can be used, and further debiased with the aid of path integral theory. Compared to prior methods for solving GSB problems, our GSBM algorithm always preserves a feasible transport map between the boundary distributions throughout training, thereby enabling stable convergence and significantly improved scalability. We empirically validate our claims on an extensive suite of experimental setups, including crowd navigation, opinion depolarization, LiDAR manifolds, and image domain transfer. Our work brings new algorithmic opportunities for training diffusion models enhanced with task-specific optimality structures.
****
Mengyuan Chen,Junyu Gao,Changsheng Xu
R-EDL: Relaxing Nonessential Settings of Evidential Deep Learning
A newly-arising uncertainty estimation method named Evidential Deep Learning (EDL), which can obtain reliable predictive uncertainty in a single forward pass, has garnered increasing interest. Guided by the subjective logic theory, EDL obtains Dirichlet concentration parameters from deep neural networks, thus constructing a Dirichlet probability density function (PDF) to model the distribution of class probabilities. Despite its great success, we argue that EDL keeps nonessential settings in both stages of model construction and optimization.
In this work, our analysis indicates that (1) in the construction of the Dirichlet PDF, a commonly ignored parameter termed prior weight governs the balance between leveraging the proportion of evidence and its magnitude in deriving predictive scores, and (2) in model optimization, a variance-minimized regularization term adopted by traditional EDL encourages the Dirichlet PDF to approach a Dirac delta function, potentially exacerbating overconfidence. Therefore, we propose the R-EDL (Relaxed-EDL) method by relaxing these nonessential settings. Specifically, R-EDL treats the prior weight as an adjustable hyper-parameter instead of a fixed scalar, and directly optimizes the expectation of the Dirichlet PDF provided to deprecate the variance-minimized regularization term. Extensive experiments and SOTA performances demonstrate the effectiveness of our method. Source codes are provided in Appendix E.
****
Sina Alemohammad,Josue Casco-Rodriguez,Lorenzo Luzi,Ahmed Imtiaz Humayun,Hossein Babaei,Daniel LeJeune,Ali Siahkoohi,Richard Baraniuk
Self-Consuming Generative Models Go MAD
Seismic advances in generative AI algorithms for imagery, text, and other data types have led to the temptation to use AI-synthesized data to train next-generat

ion models. Repeating this process creates an autophagous ("self-consuming") loo p whose properties are poorly understood. We conduct a thorough analytical and empirical analysis using state-of-the-art generative image models of three famil ies of autophagous loops that differ in how fixed or fresh real training data is available through the generations of training and whether the samples from prev ious-generation models have been biased to trade off data quality versus diversi ty. Our primary conclusion across all scenarios is that *without enough fresh r eal data in each generation of an autophagous loop, future generative models are doomed to have their quality (precision) or diversity (recall) progressively de crease.* We term this condition Model Autophagy Disorder (MAD), by analogy to ma d cow disease, and show that appreciable MADness arises in just a few generation s.

****

Abudukelimu Wuerkaixi,Sen Cui,Jingfeng Zhang,Kunda Yan,Bo Han,Gang Niu,Lei Fang, Changshui Zhang,Masashi Sugiyama

Accurate Forgetting for Heterogeneous Federated Continual Learning

Recent years have witnessed a burgeoning interest in federated learning (FL). Ho wever, the contexts in which clients engage in sequential learning remain under- explored. Bridging FL and continual learning (CL) gives rise to a challenging p ractical problem: federated continual learning (FCL). Existing research in FCL p rimarily focuses on mitigating the catastrophic forgetting issue of continual le arning while collaborating with other clients. We argue that forgetting phenomen a are not invariably detrimental. In this paper, we consider a more practical an d challenging FCL setting characterized by potentially unrelated or even antagon istic data/tasks across different clients. In the FL scenario, statistical heter ogeneity and data noise among clients may exhibit spurious correlations which re sult in biased feature learning. While existing CL strategies focus on the compl ete utilization of previous knowledge, we found that forgetting biased informati on was beneficial in our study. Therefore, we propose a new concept accurate for getting (AF) and develop a novel generative-replay method AF-FCL that selectivel y utilizes previous knowledge in federated networks. We employ a probabilistic f ramework based on a normalizing flow model to quantify the credibility of previo us knowledge. Comprehensive experiments affirm the superiority of our method ove r baselines.

****

Francis Engelmann,Fabian Manhardt,Michael Niemeyer,Keisuke Tateno,Federico Tomba ri

OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Ren dered Novel Views

Large visual-language models (VLMs), like CLIP, enable open-set image segmentati on to segment arbitrary concepts from an image in a zero-shot manner. This goes beyond the traditional closed-set assumption, i.e., where models can only segmen t classes from a pre-defined training set. More recently, first works on open-se t segmentation in 3D scenes have appeared in the literature. These methods are h eavily influenced by closed-set 3D convolutional approaches that process point c louds or polygon meshes. However, these 3D scene representations do not align we ll with the image-based nature of the visual-language models. Indeed, point clou d and 3D meshes typically have a lower resolution than images and the reconstruc ted 3D scene geometry might not project well to the underlying 2D image sequence s used to compute pixel-aligned CLIP features. To address these challenges, we p ropose OpenNeRF which naturally operates on posed images and directly encodes th e VLM features within the NeRF. This is similar in spirit to LERF, however our w ork shows that using pixel-wise VLM features (instead of global CLIP features) r esults in an overall less complex architecture without the need for additional D INO regularization. Our OpenNeRF further leverages NeRF’s ability to render nove l views and extract open-set VLM features from areas that are not well observed in the initial posed images. For 3D point cloud segmentation on the Replica data set, OpenNeRF outperforms recent open-vocabulary methods such as LERF and OpenSc ene by at least +4.9 mIoU.

****

Hsiang Hsu,Guihong Li,Shaohan Hu,Chun-Fu Chen
Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Est imation
Predictive multiplicity refers to the phenomenon in which classification tasks m ay admit multiple competing models that achieve almost-equally-optimal performan ce, yet generate conflicting outputs for individual samples.
This presents significant concerns, as it can potentially result in systemic exc lusion, inexplicable discrimination, and unfairness in practical applications.
Measuring and mitigating predictive multiplicity, however, is computationally ch allenging due to the need to explore all such almost-equally-optimal models, kno wn as the Rashomon set, in potentially huge hypothesis spaces.
To address this challenge, we propose a novel framework that utilizes dropout te chniques for exploring models in the Rashomon set.
We provide rigorous theoretical derivations to connect the dropout parameters to properties of the Rashomon set, and empirically evaluate our framework through extensive experimentation.
Numerical results show that our technique consistently outperforms baselines in terms of the effectiveness of predictive multiplicity metric estimation, with ru ntime speedup up to $20\times \sim 5000\times$.
With efficient Rashomon set exploration and metric estimation, mitigation of pre dictive multiplicity is then achieved through dropout ensemble and model selecti on.
****
Shubham Ugare,Tarun Suresh,Debangshu Banerjee,Gagandeep Singh,Sasa Misailovic
Incremental Randomized Smoothing Certification
Randomized smoothing-based certification is an effective approach for obtaining robustness certificates of deep neural networks (DNNs) against adversarial attac ks. This method constructs a smoothed DNN model and certifies its robustness thr ough statistical sampling, but it is computationally expensive, especially when certifying with a large number of samples. Furthermore, when the smoothed model is modified (e.g., quantized or pruned), certification guarantees may not hold f or the modified DNN, and recertifying from scratch can be prohibitively expensiv e.

We present the first approach for incremental robustness certification for rando mized smoothing, IRS. We show how to reuse the certification guarantees for the original smoothed model to certify an approximated model with very few samples. IRS significantly reduces the computational cost of certifying modified DNNs whi le maintaining strong robustness guarantees. We experimentally demonstrate the e ffectiveness of our approach, showing up to 4.1x certification speedup over the certification that applies randomized smoothing of the approximate model from sc ratch.
****
Zhang Zihan,Jason D. Lee,Yuxin Chen,Simon Shaolei Du
Horizon-Free Regret for Linear Markov Decision Processes
A recent line of works showed regret bounds in reinforcement learning (RL) can b e (nearly) independent of planning horizon, a.k.a. the horizon-free bounds. Howe ver, these regret bounds only apply to settings where a polynomial dependency on the size of transition model is allowed, such as tabular Markov Decision Proces s (MDP) and linear mixture MDP. We give the first horizon-free bound for the pop ular linear MDP setting where the size of the transition model can be exponentia lly large or even uncountable. In contrast to prior works which explicitly esti mate the transition model and compute the inhomogeneous value functions at diffe rent time steps, we directly estimate the value functions and confidence sets.
We obtain the horizon-free bound by: (1) maintaining multiple weighted least squ are estimators for the value functions; and (2) a structural lemma which shows t he maximal total variation of the inhomogeneous value functions is bounded by a polynomial factor of the feature dimension.
****
Elias Abad Rocamora,Fanghui Liu,Grigorios Chrysos,Pablo M. Olmos,Volkan Cevher

Efficient local linearity regularization to overcome catastrophic overfitting
Catastrophic overfitting (CO) in single-step adversarial training (AT) results in abrupt drops in the adversarial test accuracy (even down to 0%). For models trained with multi-step AT, it has been observed that the loss function behaves locally linearly with respect to the input, this is however lost in single-step AT. To address CO in single-step AT, several methods have been proposed to enforce local linearity of the loss via regularization. However, these regularization terms considerably slow down training due to *Double Backpropagation*. Instead, in this work, we introduce a regularization term, called ELLE, to mitigate CO *effectively* and *efficiently* in classical AT evaluations, as well as some more difficult regimes, e.g., large adversarial perturbations and long training schedules. Our regularization term can be theoretically linked to curvature of the loss function and is computationally cheaper than previous methods by avoiding *Double Backpropagation*. Our thorough experimental validation demonstrates that our work does not suffer from CO, even in challenging settings where previous works suffer from it. We also notice that adapting our regularization parameter during training (ELLE-A) greatly improves the performance, specially in large ϵ setups. Our implementation is available in https://github.com/LIONS-EPFL/ELLE.
****
Hongcheng Guo,Jian Yang,Jiaheng Liu,Liqun Yang,Linzheng Chai,Jiaqi Bai,Junran Peng,Xiaorong Hu,Chao Chen,Dongfeng Zhang,xu Shi,Tieqiao Zheng,liangfan zheng,Bo Zhang,Ke Xu,Zhoujun Li
OWL: A Large Language Model for IT Operations
With the rapid advancement of IT operations, managing and analyzing large data volumes efficiently for practical applications has become increasingly critical. Natural Language Processing (NLP) techniques have demonstrated remarkable capabilities in various tasks, including named entity recognition, machine translation, and dialogue systems. Recently, Large Language Models (LLMs) have achieved significant improvements across various domain-specific areas. However, there is a noticeable gap in the development of specialized Large Language Models (LLMs) tailored for IT operations. In this paper, we introduce the OWL, a large language model trained on our constructed Owl-Instruct with a wide range of IT-related information. Specifically, limited by the maximum input length, we propose the \textbf{H}omogeneous \textbf{M}arkov \textbf{C}ontext \textbf{E}xtension method (HMCE). The mixture-of-adapter strategy is leveraged to improve the parameter-efficient tuning across different domains or tasks.
Further, we evaluate the performance of OWL on the Owl-Bench established by us and open IT-related benchmarks. OWL demonstrates superior performance results on IT tasks, which outperforms existing models by significant margins. Moreover, we hope that the findings of our work will provide more insights to revolutionize the techniques of IT operations with specialized LLMs.
****
Linwei Chen,Lin Gu,Ying Fu
When Semantic Segmentation Meets Frequency Aliasing
Despite recent advancements in semantic segmentation, where and what pixels are hard to segment remains largely unexplored.
Existing research only separates an image into easy and hard regions and empirically observes the latter are associated with object boundaries.
In this paper, we conduct a comprehensive analysis of hard pixel errors, categorizing them into three types: false responses, merging mistakes, and displacements.
Our findings reveal a quantitative association between hard pixels and aliasing,

which is distortion caused by the overlapping of frequency components in the Fourier domain during downsampling.
To identify the frequencies responsible for aliasing, we propose using the equivalent sampling rate to calculate the Nyquist frequency, which marks the threshold for aliasing.
Then, we introduce the aliasing score as a metric to quantify the extent of alia

sing.
While positively correlated with the proposed aliasing score, three types of har d pixels exhibit different patterns.
Here, we propose two novel de-aliasing filter (DAF) and frequency mixing (FreqMi x) modules to alleviate aliasing degradation by accurately removing or adjusting frequencies higher than the Nyquist frequency.
The DAF precisely removes the frequencies responsible for aliasing before downsa mpling,
while the FreqMix dynamically selects high-frequency components within the encod er block.
Experimental results demonstrate consistent improvements in semantic segmentatio n and low-light instance segmentation tasks.
The code is at: \url{https://github.com/Linwei-Chen/Seg-Aliasing}.
****
Dongqi Fu,Zhigang Hua,Yan Xie,Jin Fang,Si Zhang,Kaan Sancak,Hao Wu,Andrey Malevi ch,Jingrui He,Bo Long
VCR-Graphormer: A Mini-batch Graph Transformer via Virtual Connections
Graph transformer has been proven as an effective graph learning method for its adoption of attention mechanism that is capable of capturing expressive represen tations from complex topological and feature information of graphs. Graph transf ormer conventionally performs dense attention (or global attention) for every pa ir of nodes to learn node representation vectors, resulting in quadratic computa tional costs that are unaffordable for large-scale graph data. Therefore, mini-b atch training for graph transformers is a promising direction, but limited sampl es in each mini-batch can not support effective dense attention to encode inform ative representations. Facing this bottleneck, (1) we start by assigning each no de a token list that is sampled by personalized PageRank (PPR) and then apply st andard multi-head self-attention only on this list to compute its node represent ations. This PPR tokenization method decouples model training from complex graph topological information and makes heavy feature engineering offline and indepen dent, such that mini-batch training of graph transformers is possible by loading each node's token list in batches. We further prove this PPR tokenization is vi able as a graph convolution network with a fixed polynomial filter and jumping k nowledge. However, only using personalized PageRank may limit information carrie d by a token list, which could not support different graph inductive biases for model training. To this end, (2) we rewire graphs by introducing multiple types of virtual connections through structure- and content-based super nodes that ena ble PPR tokenization to encode local and global contexts, long-range interaction , and heterophilous information into each node's token list, and then formalize our $\underline{\textbf{V}}$irtual $\underline{\textbf{C}}$onnection $\underline {\textbf{R}}$anking based $\underline{\textbf{Graph}}$ Trans$\underline{\textbf{ former}}$ (VCR-Graphormer). Overall, VCR-Graphormer needs $O(m+klogk)$ complexit y for graph tokenization as compared to $O(n^{3})$ of previous works. The [code] (https://github.com/DongqiFu/VCR-Graphormer) is provided.
****
Enric Boix-Adserà,Omid Saremi,Emmanuel Abbe,Samy Bengio,Etai Littwin,Joshua M. S usskind
When can transformers reason with abstract symbols?
We investigate the capability of Transformer large language models (LLMs) to gen eralize on unseen symbols when trained on tasks that rely on abstract symbols (e.g., variables in programming and mathematics). Such a 'variable-binding' capa bility has long been studied in the neuroscience literature as one of the most basic 'reasoning' capabilities. For (i) binary classification tasks, we prove th at Transformers can generalize to unseen symbols but require astonishingly large training data. For (ii) tasks with labels dependent on input symbols, we show a n ''inverse scaling law'': Transformers fail to generalize to unseen symbols as their embedding dimension increases. For both cases (i) and (ii), we propose a T ransformer modification, adding two trainable parameters per head that can reduc e the amount of data needed.
****

Sohan Patnaik,Heril Changwal,Milan Aggarwal,Sumit Bhatia,Yaman Kumar,Balaji Kris hnamurthy
CABINET: Content Relevance-based Noise Reduction for Table Question Answering
Table understanding capability of Large Language Models (LLMs) has been extensiv ely studied through the task of question-answering (QA) over tables. Typically, only a small part of the whole table is relevant to derive the answer for a give n question. The irrelevant parts act as noise and are distracting information, r esulting in sub-optimal performance due to the vulnerability of LLMs to noise. T o mitigate this, we propose CABINET (Content RelevAnce-Based NoIse ReductioN for TablE QuesTion-Answering) – a framework to enable LLMs to focus on relevant tab ular data by suppressing extraneous information. CABINET comprises an Unsupervis ed Relevance Scorer (URS), trained differentially with the QA LLM, that weighs t he table content based on its relevance to the input question before feeding it to the question answering LLM (QA LLM). To further aid the relevance scorer, CAB INET employs a weakly supervised module that generates a parsing statement descr ibing the criteria of rows and columns relevant to the question and highlights t he content of corresponding table cells. CABINET significantly outperforms vario us tabular LLM baselines, as well as GPT3-based in-context learning methods, is more robust to noise, maintains outperformance on tables of varying sizes, and e stablishes new SoTA performance on WikiTQ, FeTaQA, and WikiSQL datasets. We rele ase our code and datasets here.
****
Saeed Saadatnejad,Yang Gao,Kaouther Messaoud,Alexandre Alahi
Social-Transmotion: Promptable Human Trajectory Prediction
Accurate human trajectory prediction is crucial for applications such as autonom ous vehicles, robotics, and surveillance systems. Yet, existing models often fai l to fully leverage the non-verbal social cues human subconsciously communicate when navigating the space.
To address this, we introduce \textit{Social-Transmotion}, a generic model that exploits the power of transformers to handle diverse and numerous visual cues, c apturing the multi-modal nature of human behavior. We translate the idea of a pr ompt from Natural Language Processing (NLP) to the task of human trajectory pred iction, where a prompt can be a sequence of x-y coordinates on the ground, bound ing boxes or body poses. This, in turn, augments trajectory data, leading to enh anced human trajectory prediction.
Our model exhibits flexibility and adaptability by capturing spatiotemporal inte ractions between pedestrians based on the available visual cues, whether they ar e poses, bounding boxes, or a combination thereof.
By the masking technique, we ensure our model's effectiveness even when certain visual cues are unavailable, although performance is further boosted with the pr esence of comprehensive visual data.
We delve into the merits of using 2d versus 3d poses, and a limited set of poses . Additionally, we investigate the spatial and temporal attention map to identif y which keypoints and frames of poses are vital for optimizing human trajectory prediction.
Our approach is validated on multiple datasets, including JTA, JRDB, Pedestrians and Cyclists in Road Traffic, and ETH-UCY.
****
Yan Liu,Yu Liu,Xiaokang Chen,Pin-Yu Chen,Daoguang Zan,Min-Yen Kan,Tsung-Yi Ho
The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Langua ge Models
Pre-trained Language models (PLMs) have been acknowledged to contain harmful inf ormation, such as social biases, which may cause negative social impacts or even bring catastrophic results in application. Previous works on this problem mainl y focused on using black-box methods such as probing to detect and quantify soci al biases in PLMs by observing model outputs. As a result, previous debiasing me thods mainly finetune or even pre-train PLMs on newly constructed anti-stereotyp ical datasets, which are high-cost. In this work, we try to unveil the mystery o f social bias inside language models by introducing the concept of {\sc Social B ias Neurons}. Specifically, we propose {\sc Integrated Gap Gradients (IG^2)} t

o accurately pinpoint units (i.e., neurons) in a language model that can be attr
ibuted to undesirable behavior, such as social bias. By formalizing undesirable
 behavior as a distributional property of language, we employ sentiment-bearing
prompts to elicit classes of sensitive words (demographics) correlated with such
 sentiments. Our IG^2 thus attributes the uneven distribution for different de
mographics to specific Social Bias Neurons, which track the trail of unwanted be
havior inside PLM units to achieve interoperability. Moreover, derived from our
interpretable technique, {\sc Bias Neuron Suppression (BNS)} is further proposed
 to mitigate social biases. By studying BERT, RoBERTa, and their attributable di
fferences from debiased FairBERTa, IG^2 allows us to locate and suppress ident
ified neurons, and further mitigate undesired behaviors. As measured by prior me
trics from StereoSet, our model achieves a higher degree of fairness while maint
aining language modeling ability with low cost\footnote{This work contains examp
les that potentially implicate stereotypes, associations, and other harms that c
ould be offensive to individuals in certain social groups.}.
****
Yibing Liu,Chris XING TIAN,Haoliang Li,Lei Ma,Shiqi Wang
Neuron Activation Coverage: Rethinking Out-of-distribution Detection and General
ization
The out-of-distribution (OOD) problem generally arises when neural networks enco
unter data that significantly deviates from the training data distribution, i.e.
, in-distribution (InD). In this paper, we study the OOD problem from a neuron a
ctivation view. We first formulate neuron activation states by considering both
the neuron output and its influence on model decisions. Then, to characterize th
e relationship between neurons and OOD issues, we introduce the *neuron activati
on coverage* (NAC) -- a simple measure for neuron behaviors under InD data. Leve
raging our NAC, we show that 1) InD and OOD inputs can be largely separated base
d on the neuron behavior, which significantly eases the OOD detection problem an
d beats the 21 previous methods over three benchmarks (CIFAR-10, CIFAR-100, and
ImageNet-1K). 2) a positive correlation between NAC and model generalization abi
lity consistently holds across architectures and datasets, which enables a NAC-b
ased criterion for evaluating model robustness. Compared to prevalent InD valida
tion criteria, we show that NAC not only can select more robust models, but also
 has a stronger correlation with OOD test performance.
****
Jasper Dekoninck,Marc Fischer,Luca Beurer-Kellner,Martin Vechev
Controlled Text Generation via Language Model Arithmetic
As Large Language Models (LLMs) are deployed more widely, customization with res
pect to vocabulary, style, and character becomes more important. In this work, w
e introduce model arithmetic, a novel inference framework for composing and bias
ing LLMs without the need for model (re)training or highly specific datasets. In
 addition, the framework allows for more precise control of generated text than
direct prompting and prior controlled text generation (CTG) techniques. Using mo
del arithmetic, we can express prior CTG techniques as simple formulas and natur
ally extend them to new and more effective formulations. Further, we show that s
peculative sampling, a technique for efficient LLM sampling, extends to our sett
ing. This enables highly efficient text generation with multiple composed models
 with only marginal overhead over a single model. Our empirical evaluation demon
strates that model arithmetic allows fine-grained control of generated text whil
e outperforming state-of-the-art on the task of toxicity reduction. We release a
n open source easy-to-use implementation of our framework at https://github.com/
eth-sri/language-model-arithmetic.
****
Ge Yan,Hongxu Chen,Kaisen Pan,Junchi Yan
Rethinking the symmetry-preserving circuits for constrained variational quantum
algorithms
With the arrival of the Noisy Intermediate-Scale Quantum (NISQ) era, Variational
 Quantum Algorithms (VQAs) have emerged as popular approaches to obtain possible
 quantum advantage in the relatively near future. In particular, how to effectiv
ely incorporate the common symmetries in physical systems as hard constraints in

VQAs remains a critical and open question. In this paper, we revisit the Hamming Weight (HW) preserving ansatz and establish the links from ansatz to various symmetries and constraints, which both enlarges the usage of HW preserving ansatz and provides a coherent solution for constrained VQAs. Meanwhile, we utilize the quantum optimal control theory and quantum overparameterization theory to analyze the capability and expressivity of HW preserving ansatz and verify these theoretical results on unitary approximation problem. We conduct detailed numerical experiments on two well-studied symmetry-preserving problems, namely ground state energy estimation and feature selection in machine learning. The superior performance demonstrates the efficiency and supremacy of the proposed HW preserving ansatz on constrained VQAs.
****
Aoqi Zuo,Yiqing Li,Susan Wei,Mingming Gong
Interventional Fairness on Partially Known Causal Graphs: A Constrained Optimization Approach
Fair machine learning aims to prevent discrimination against individuals or sub-populations based on sensitive attributes such as gender and race. In recent years, causal inference methods have been increasingly used in fair machine learning to measure unfairness by causal effects. However, current methods assume that the true causal graph is given, which is often not true in real-world applications. To address this limitation, this paper proposes a framework for achieving causal fairness based on the notion of interventions when the true causal graph is partially known. The proposed approach involves modeling fair prediction using a Partially Directed Acyclic Graph (PDAG), specifically, a class of causal DAGs that can be learned from observational data combined with domain knowledge. The PDAG is used to measure causal fairness, and a constrained optimization problem is formulated to balance between fairness and accuracy. Results on both simulated and real-world datasets demonstrate the effectiveness of this method.
****
Shaofei Shen,Chenhao Zhang,Yawen Zhao,Alina Bialkowski,Weitong Tony Chen,Miao Xu
Label-Agnostic Forgetting: A Supervision-Free Unlearning in Deep Models
Machine unlearning aims to remove information derived from forgotten data while preserving that of the remaining dataset in a well-trained model. With the increasing emphasis on data privacy, several approaches to machine unlearning have emerged. However, these methods typically rely on complete supervision throughout the unlearning process. Unfortunately, obtaining such supervision, whether for the forgetting or remaining data, can be impractical due to the substantial cost associated with annotating real-world datasets. This challenge prompts us to propose a supervision-free unlearning approach that operates without the need for labels during the unlearning process. Specifically, we introduce a variational approach to approximate the distribution of representations for the remaining data. Leveraging this approximation, we adapt the original model to eliminate information from the forgotten data at the representation level. To further address the issue of lacking supervision information, which hinders alignment with ground truth, we introduce a contrastive loss to facilitate the matching of representations between the remaining data and those of the original model, thus preserving predictive performance. Experimental results across various unlearning tasks demonstrate the effectiveness of our proposed method, Label-Agnostic Forgetting (LAF) without using any labels, which achieves comparable performance to state-of-the-art methods that rely on full supervision information. Furthermore, our approach excels in semi-supervised scenarios, leveraging limited supervision information to outperform fully supervised baselines. This work not only showcases the viability of supervision-free unlearning in deep models but also opens up a new possibility for future research in unlearning at the representation level.
****
Ganlin Yang,Guoqiang Wei,Zhizheng Zhang,Yan Lu,Dong Liu
Mask-Based Modeling for Neural Radiance Fields
Most Neural Radiance Fields (NeRFs) exhibit limited generalization capabilities, which restrict their applicability in representing multiple scenes using a single model. To address this problem, existing generalizable NeRF methods simply con

dition the model on image features. These methods still struggle to learn precis e global representations over diverse scenes since they lack an effective mechan ism for interacting among different points and views. In this work, we unveil th at 3D implicit representation learning can be significantly improved by mask-bas ed modeling. Specifically, we propose **m**asked **r**ay and **v**iew **m**odeli ng for generalizable **NeRF** (**MRVM-NeRF**), which is a self-supervised pretra ining target to predict complete scene representations from partially masked fea tures along each ray. With this pretraining target, MRVM-NeRF enables better use of correlations across different rays and views as the geometry priors, which t hereby strengthens the capability of capturing intricate details within the scen es and boosts the generalization capability across different scenes. Extensive e xperiments demonstrate the effectiveness of our proposed MRVM-NeRF on both synth etic and real-world datasets, qualitatively and quantitatively. Besides, we also conduct experiments to show the compatibility of our proposed method with vario us backbones and its superiority under few-shot cases.
****
Avni Kothari,Bogdan Kulynych,Tsui-Wei Weng,Berk Ustun
Prediction without Preclusion: Recourse Verification with Reachable Sets
Machine learning models are often used to decide who receives a loan, a job inte rview, or a public benefit. Standard methods to learn such models use features a bout people but overlook their actionability. As a result, models can assign pre dictions that are fixed – meaning that consumers who are denied loans, interview s, or benefits are precluded from access to credit, employment, or assistance. I n this work, we present a task called recourse verification to flag models that assign fixed predictions under a rich class of real-world actionability constrai nts. We develop methods to check if a model can provide recourse using reachable sets. We demonstrate how our tools can verify recourse in real-world lending da tasets. Our results highlight how models can inadvertently assign fixed predicti ons that permanently bar access, and underscore the need to account for actionab ility in model development.
****
Junchi Yu,Ran He,Zhitao Ying
THOUGHT PROPAGATION: AN ANALOGICAL APPROACH TO COMPLEX REASONING WITH LARGE LANG UAGE MODELS
Large Language Models (LLMs) have achieved remarkable success in reasoning tasks with the development of prompting methods.
However, existing prompting approaches cannot reuse insights of solving similar problems and suffer from accumulated errors in multi-step reasoning, since they prompt LLMs to reason \textit{from scratch}.
To address these issues, we propose \textbf{\textit{Thought Propagation} (TP)}, which explores the analogous problems and leverages their solutions to enhance t he complex reasoning ability of LLMs.
These analogous problems are related to the input one, with reusable solutions a nd problem-solving strategies.
Thus, it is promising to propagate insights of solving previous analogous proble ms to inspire new problem-solving.
To achieve this, TP first prompts LLMs to propose and solve a set of analogous p roblems that are related to the input one.
Then, TP reuses the results of analogous problems to directly yield a new soluti on or derive a knowledge-intensive plan for execution to amend the initial solut ion obtained from scratch.
TP is compatible with existing prompting approaches, allowing plug-and-play gene ralization and enhancement in a wide range of tasks without much labor in task-s pecific prompt engineering.
Experiments across three challenging tasks demonstrate TP enjoys a substantial i mprovement over the baselines by an average of 12\% absolute increase in finding the optimal solutions in Shortest-path Reasoning, 13\% improvement of human pre ference in Creative Writing, and 15\% enhancement in the task completion rate of LLM-Agent Planning.
****

Faisal Hamman,Sanghamitra Dutta
Demystifying Local & Global Fairness Trade-offs in Federated Learning Using Part
ial Information Decomposition
This work presents an information-theoretic perspective to group fairness trade-
offs in federated learning (FL) with respect to sensitive attributes, such as ge
nder, race, etc. Existing works often focus on either $\textit{global fairness}$
 (overall disparity of the model across all clients) or $\textit{local fairness}
$ (disparity of the model at each client), without always considering their trad
e-offs. There is a lack of understanding regarding the interplay between global
and local fairness in FL, particularly under data heterogeneity, and if and when
 one implies the other. To address this gap, we leverage a body of work in infor
mation theory called partial information decomposition (PID), which first identi
fies three sources of unfairness in FL, namely, $\textit{Unique Disparity}$, $\t
extit{Redundant Disparity}$, and $\textit{Masked Disparity}$. We demonstrate ho
w these three disparities contribute to global and local fairness using canonica
l examples. This decomposition helps us derive fundamental limits on the trade-o
ff between global and local fairness, highlighting where they agree or disagree.
  We introduce the $\textit{Accuracy and Global-Local Fairness Optimality Proble
m}$ (AGLFOP), a convex optimization that defines the theoretical limits of accur
acy and fairness trade-offs, identifying the best possible performance any FL st
rategy can attain given a dataset and client distribution. We also present exper
imental results on synthetic datasets and the ADULT dataset to support our theor
etical findings.
****
Yucen Lily Li,Tim G. J. Rudner,Andrew Gordon Wilson
A Study of Bayesian Neural Network Surrogates for Bayesian Optimization
Bayesian optimization is a highly efficient approach to optimizing objective fun
ctions which are expensive to query. These objectives are typically represented
by Gaussian process (GP) surrogate models which are easy to optimize and support
 exact inference. While standard GP surrogates have been well-established in Bay
esian optimization, Bayesian neural networks (BNNs) have recently become practic
al function approximators, with many benefits over standard GPs such as the abil
ity to naturally handle non-stationarity and learn representations for high-dime
nsional data. In this paper, we study BNNs as alternatives to standard GP surrog
ates for optimization. We consider a variety of approximate inference procedures
 for finite-width BNNs, including high-quality Hamiltonian Monte Carlo, low-cost
 stochastic MCMC, and heuristics such as deep ensembles. We also consider infini
te-width BNNs, linearized Laplace approximations, and partially stochastic model
s such as deep kernel learning. We evaluate this collection of surrogate models
on diverse problems with varying dimensionality, number of objectives, non-stati
onarity, and discrete and continuous inputs. We find: (i) the ranking of methods
 is highly problem dependent, suggesting the need for tailored inductive biases;
 (ii) HMC is the most successful approximate inference procedure for fully stoch
astic BNNs; (iii) full stochasticity may be unnecessary as deep kernel learning
is relatively competitive; (iv) deep ensembles perform relatively poorly; (v) in
finite-width BNNs are particularly promising, especially in high dimensions.
****
Zixiang Chen,Yihe Deng,Yuanzhi Li,Quanquan Gu
Understanding Transferable Representation Learning and Zero-shot Transfer in CLI
P
Multi-modal learning has become increasingly popular due to its ability to lever
age information from different data sources (e.g., text and images) to improve t
he model performance. Recently, CLIP has emerged as an effective approach that e
mploys vision-language contrastive pretraining to learn joint image and text rep
resentations and exhibits remarkable performance in zero-shot learning and text-
guided natural image generation. Despite the substantial practical success of CL
IP, its theoretical understanding remains elusive. In this paper, we formally st
udy transferrable representation learning underlying CLIP and demonstrate how fe
atures from different modalities get aligned. We also analyze its zero-shot tran
sfer performance on the downstream tasks. In addition, we conduct empirical eval

uations on real data to back
up our theory. Inspired by our analysis, we propose a new CLIP-type approach, which achieves better performance than CLIP and other state-of-the-art methods on benchmark datasets.
****
Benjamin S. H. Lyo,Cristina Savin
Complex priors and flexible inference in recurrent circuits with dendritic nonlinearities
Despite many successful examples in which probabilistic inference can account for perception, we have little understanding of how the brain represents and uses structured priors that capture the complexity of natural input statistics. Here we construct a recurrent circuit model that can implicitly represent priors over latent variables, and combine them with sensory and contextual sources of information to encode task-specific posteriors. Inspired by the recent success of diffusion models as means of learning and using priors over images, our model uses dendritic nonlinearities optimized for denoising, and stochastic somatic integration with the degree of noise modulated by an oscillating global signal. Combining these elements into a recurrent network yields a dynamical system that samples from the prior at a rate prescribed by the period of the global oscillator. Additional inputs reflecting sensory or top-down contextual information alter these dynamics to generate samples from the corresponding posterior, with different input gating patterns selecting different inference tasks. We demonstrate that this architecture can sample from low dimensional nonlinear manifolds and multimodal posteriors. Overall, the model provides a new framework for circuit-level representation of probabilistic information, in a format that facilitates flexible inference.
****
Yifei Wang,Jizhe Zhang,Yisen Wang
Do Generated Data Always Help Contrastive Learning?
Contrastive Learning (CL) has emerged as one of the most successful paradigms for unsupervised visual representation learning, yet it often depends on intensive manual data augmentations. With the rise of generative models, especially diffusion models, the ability to generate realistic images close to the real data distribution has been well recognized. These generated high-equality images have been successfully applied to enhance contrastive representation learning, a technique termed ``data inflation''. However, we find that the generated data (even from a good diffusion model like DDPM) may sometimes even harm contrastive learning. We investigate the causes behind this failure from the perspective of both data inflation and data augmentation. For the first time, we reveal the complementary roles that stronger data inflation should be accompanied by weaker augmentations, and vice versa. We also provide rigorous theoretical explanations for these phenomena via deriving its generalization bounds under data inflation. Drawing from these insights, we propose **Adaptive Inflation (AdaInf)**, a purely data-centric strategy without introducing any extra computation cost. On benchmark datasets, AdaInf can bring significant improvements for various contrastive learning methods. Notably, without using external data, AdaInf obtains 94.70% linear accuracy on CIFAR-10 with SimCLR, setting a new record that surpasses many sophisticated methods. Code is available at https://github.com/PKU-ML/adainf.
****
Erdun Gao,Howard Bondell,Wei Huang,Mingming Gong
A Variational Framework for Estimating Continuous Treatment Effects with Measurement Error
Estimating treatment effects has numerous real-world applications in various fields, such as epidemiology and political science. While much attention has been devoted to addressing the challenge using fully observational data, there has been comparatively limited exploration of this issue in cases when the treatment is not directly observed. In this paper, we tackle this problem by developing a general variational framework, which is flexible to integrate with advanced neural network-based approaches, to identify the average dose-response function (ADRF) with the continuously valued error-contaminated treatment. Our approach begins

with the formulation of a probabilistic data generation model, treating the unob served treatment as a latent variable. In this model, we leverage a learnable de nsity estimation neural network to derive its prior distribution conditioned on covariates. This module also doubles as a generalized propensity score estimator , effectively mitigating selection bias arising from observed confounding variab les. Subsequently, we calculate the posterior distribution of the treatment, tak ing into account the observed measurement and outcome. To mitigate the impact of treatment error, we introduce a re-parametrized treatment value, replacing the error-affected one, to make more accurate predictions regarding the outcome. To demonstrate the adaptability of our framework, we incorporate two state-of-the-a rt ADRF estimation methods and rigorously assess its efficacy through extensive simulations and experiments using semi-synthetic data.
****
Yue Wu,Xuan Tang,Tom Mitchell,Yuanzhi Li
SmartPlay : A Benchmark for LLMs as Intelligent Agents
Recent large language models (LLMs) have demonstrated great potential toward int elligent agents and next-gen automation, but there currently lacks a systematic benchmark for evaluating LLMs' abilities as agents. We introduce SmartPlay: both a challenging benchmark and a methodology for evaluating LLMs as agents. SmartP lay consists of 6 different games, including Rock-Paper-Scissors, Tower of Hanoi , Minecraft. Each game features a unique setting, providing up to 20 evaluation settings and infinite environment variations. Each game in SmartPlay uniquely ch allenges a subset of 9 important capabilities of an intelligent LLM agent, inclu ding reasoning with object dependencies, planning ahead, spatial reasoning, lear ning from history, and understanding randomness. The distinction between the set of capabilities each game test allows us to analyze each capability separately. SmartPlay serves not only as a rigorous testing ground for evaluating the overal l performance of LLM agents but also as a road-map for identifying gaps in curre nt methodologies.
We release our benchmark at https://github.com/microsoft/SmartPlay
****
Tsu-Jui Fu,Wenze Hu,Xianzhi Du,William Yang Wang,Yinfei Yang,Zhe Gan
Guiding Instruction-based Image Editing via Multimodal Large Language Models
Instruction-based image editing improves the controllability and flexibility of image manipulation via natural commands without elaborate descriptions or region al masks. However, human instructions are sometimes too brief for current method s to capture and follow. Multimodal large language models (MLLMs) show promising capabilities in cross-modal understanding and visual-aware response generation via LMs. We investigate how MLLMs facilitate edit instructions and present MLLM-Guided Image Editing (MGIE). MGIE learns to derive expressive instructions and p rovides explicit guidance. The editing model jointly captures this visual imagin ation and performs manipulation through end-to-end training. We evaluate various aspects of Photoshop-style modification, global photo optimization, and local e diting. Extensive experimental results demonstrate that expressive instructions are crucial to instruction-based image editing, and our MGIE can lead to a notab le improvement in automatic metrics and human evaluation while maintaining compe titive inference efficiency.
****
Binwu Wang,Pengkun Wang,Wei Xu,Xu Wang,Yudong Zhang,Kun Wang,Yang Wang
Kill Two Birds with One Stone: Rethinking Data Augmentation for Deep Long-tailed Learning
Real-world tasks are universally associated with training samples that exhibit a long-tailed class distribution, and traditional deep learning models are not su itable for fitting this distribution, thus resulting in a biased trained model. To surmount this dilemma, massive deep long-tailed learning studies have been pr oposed to achieve inter-class fairness models by designing sophisticated samplin g strategies or improving existing model structures and loss functions. Habitual ly, these studies tend to apply data augmentation strategies to improve the gene ralization performance of their models. However, this augmentation strategy appl ied to balanced distributions may not be the best option for long-tailed distrib

utions. For a profound understanding of data augmentation, we first theoreticall y analyze the gains of traditional augmentation strategies in long-tailed learni ng, and observe that augmentation methods cause the long-tailed distribution to be imbalanced again, resulting in an intertwined imbalance: inherent data-wise i mbalance and extrinsic augmentation-wise imbalance, i.e., two 'birds' co-exist i n long-tailed learning. Motivated by this observation, we propose an adaptive Dy namic Optional Data Augmentation (DODA) to address this intertwined imbalance, i .e., one 'stone' simultaneously 'kills' two 'birds', which allows each class to choose appropriate augmentation methods by maintaining a corresponding augmentat ion probability distribution for each class during training. Extensive experimen ts across mainstream long-tailed recognition benchmarks (e.g., CIFAR-100-LT, Ima geNet-LT, and iNaturalist 2018) prove the effectiveness and flexibility of the D ODA in overcoming the intertwined imbalance.
****
Juno Kim,Jaehyuk Kwon,Mincheol Cho,Hyunjong Lee,Joong-Ho Won
t^3-Variational Autoencoder: Learning Heavy-tailed Data with Student's t and P ower Divergence
The variational autoencoder (VAE) typically employs a standard normal prior as a regularizer for the probabilistic latent encoder. However, the Gaussian tail of ten decays too quickly to effectively accommodate the encoded points, failing to preserve crucial structures hidden in the data. In this paper, we explore the u se of heavy-tailed models to combat over-regularization. Drawing upon insights f rom information geometry, we propose t^3VAE, a modified VAE framework that inc orporates Student's t-distributions for the prior, encoder, and decoder. This re sults in a joint model distribution of a power form which we argue can better fi t real-world datasets. We derive a new objective by reformulating the evidence l ower bound as joint optimization of KL divergence between two statistical manifo lds and replacing with γ-power divergence, a natural alternative for powe r families. t^3VAE demonstrates superior generation of low-density regions whe n trained on heavy-tailed synthetic data. Furthermore, we show that t^3VAE sig nificantly outperforms other models on CelebA and imbalanced CIFAR-100 datasets.
****
Tommaso Salvatori,Yuhang Song,Yordan Yordanov,Beren Millidge,Lei Sha,Cornelius E mde,Zhenghua Xu,Rafal Bogacz,Thomas Lukasiewicz
A Stable, Fast, and Fully Automatic Learning Algorithm for Predictive Coding Net works
Predictive coding networks are neuroscience-inspired models with roots in both B ayesian statistics and neuroscience. Training such models, however, is quite ine fficient and unstable. In this work, we show how by simply changing the temporal scheduling of the update rule for the synaptic weights leads to an algorithm th at is much more efficient and stable than the original one, and has theoretical guarantees in terms of convergence. The proposed algorithm, that we call increme ntal predictive coding (iPC) is also more biologically plausible than the origin al one, as it it fully automatic. In an extensive set of experiments, we show th at iPC constantly performs better than the original formulation on a large numbe r of benchmarks for image classification, as well as for the training of both co nditional and masked language models, in terms of test accuracy, efficiency, and convergence with respect to a large set of hyperparameters.
****
Josue Ortega Caro,Antonio Henrique de Oliveira Fonseca,Syed A Rizvi,Matteo Rosat i,Christopher Averill,James L Cross,Prateek Mittal,Emanuele Zappala,Rahul Madhav Dhodapkar,Chadi Abdallah,David van Dijk
BrainLM: A foundation model for brain activity recordings
We introduce the Brain Language Model (BrainLM), a foundation model for brain ac tivity dynamics trained on 6,700 hours of fMRI recordings. Utilizing self-superv ised masked-prediction training, BrainLM demonstrates proficiency in both fine-t uning and zero-shot inference tasks. Fine-tuning allows for the accurate predict ion of clinical variables like age, anxiety, and PTSD as well as forecasting of future brain states. Critically, the model generalizes well to entirely new exte rnal cohorts not seen during training. In zero-shot inference mode, BrainLM can

identify intrinsic functional networks directly from raw fMRI data without any network-based supervision during training. The model also generates interpretable latent representations that reveal relationships between brain activity patterns and cognitive states. Overall, BrainLM offers a versatile and interpretable framework for elucidating the complex spatiotemporal dynamics of human brain activity. It serves as a powerful "lens" through which massive repositories of fMRI data can be analyzed in new ways, enabling more effective interpretation and utilization at scale. The work demonstrates the potential of foundation models to advance computational neuroscience research.
****
Yanai Elazar,Akshita Bhagia,Ian Helgi Magnusson,Abhilasha Ravichander,Dustin Schwenk,Alane Suhr,Evan Pete Walsh,Dirk Groeneveld,Luca Soldaini,Sameer Singh,Hannaneh Hajishirzi,Noah A. Smith,Jesse Dodge
What's In My Big Data?
Large text corpora are the backbone of language models.
However, we have a limited understanding of the content of these corpora, including general statistics, quality, social factors, and inclusion of evaluation data (contamination).
In this work, we propose What's In My Big Data? (WIMBD), a platform and a set of sixteen analyses that allow us to reveal and compare the contents of large text corpora. WIMBD builds on two basic capabilities---count and search---*at scale*, which allows us to analyze more than 35 terabytes on a standard compute node.
We apply WIMBD to ten different corpora used to train popular language models, including *C4*, *The Pile*, and *RedPajama*.
Our analysis uncovers several surprising and previously undocumented findings about these corpora, including the high prevalence of duplicate, synthetic, and low-quality content, personally identifiable information, toxic language, and benchmark contamination.
For instance, we find that about 50% of the documents in *RedPajama* and *LAION-2B-en* are duplicates. In addition, several datasets used for benchmarking models trained on such corpora are contaminated with respect to important benchmarks, including the Winograd Schema Challenge and parts of GLUE and SuperGLUE.
We open-source WIMBD's code and artifacts to provide a standard set of evaluations for new text-based corpora and to encourage more analyses and transparency around them.
****
Lin-Han Jia,Lan-Zhe Guo,Zhi Zhou,Yu-Feng Li
Realistic Evaluation of Semi-supervised Learning Algorithms in Open Environments
Semi-supervised learning (SSL) is a powerful paradigm for leveraging unlabeled data and has been proven to be successful across various tasks. Conventional SSL studies typically assume close environment scenarios where labeled and unlabeled examples are independently sampled from the same distribution. However, real-world tasks often involve open environment scenarios where the data distribution, label space, and feature space could differ between labeled and unlabeled data. This inconsistency introduces robustness challenges for SSL algorithms. In this paper, we first propose several robustness metrics for SSL based on the Robustness Analysis Curve (RAC), secondly, we establish a theoretical framework for studying the generalization performance and robustness of SSL algorithms in open environments, thirdly, we re-implement widely adopted SSL algorithms within a unified SSL toolkit and evaluate their performance on proposed open environment SSL benchmarks, including both image, text, and tabular datasets. By investigating the empirical and theoretical results, insightful discussions on enhancing the robustness of SSL algorithms in open environments are presented. The re-implementation and benchmark datasets are all publicly available. More details can be found at https://ygzwqzd.github.io/Robust-SSL-Benchmark.
****
Xingyu Liu,Deepak Pathak,Ding Zhao
Meta-Evolve: Continuous Robot Evolution for One-to-many Policy Transfer
We investigate the problem of transferring an expert policy from a source robot to multiple different robots. To solve this problem, we propose a method named *

Meta-Evolve* that uses continuous robot evolution to efficiently transfer the policy to each target robot through a set of tree-structured evolutionary robot sequences. The robot evolution tree allows the robot evolution paths to be shared, so our approach can significantly outperform naive one-to-one policy transfer. We present a heuristic approach to determine an optimized robot evolution tree. Experiments have shown that our method is able to improve the efficiency of one-to-three transfer of manipulation policy by up to 3.2$\times$ and one-to-six transfer of agile locomotion policy by 2.4$\times$ in terms of simulation cost over the baseline of launching multiple independent one-to-one policy transfers. Supplementary videos available at the project website: https://sites.google.com/view/meta-evolve.
****
Bobby He,Thomas Hofmann
Simplifying Transformer Blocks
A simple design recipe for deep Transformers is to compose identical building blocks. But standard transformer blocks are far from simple, interweaving attention and MLP sub-blocks with skip connections \& normalisation layers in precise arrangements. This complexity leads to brittle architectures, where seemingly minor changes can significantly reduce training speed, or render models untrainable.

In this work, we ask to what extent the standard transformer block can be simplified? Combining signal propagation theory and empirical observations, we motivate modifications that allow many block components to be removed with no loss of training speed, including skip connections, projection or value parameters, sequential sub-blocks and normalisation layers. In experiments on both autoregressive decoder-only and BERT encoder-only models, our simplified transformers match the per-iteration training speed and performance of standard transformers, while enjoying 15\% faster training throughput, and using 15\% fewer parameters.
****
Yangming Li,Mihaela van der Schaar
On Error Propagation of Diffusion Models
Although diffusion models (DMs) have shown promising performances in a number of tasks (e.g., speech synthesis and image generation), they might suffer from error propagation because of their sequential structure. However, this is not certain because some sequential models, such as Conditional Random Field (CRF), are free from this problem. To address this issue, we develop a theoretical framework to mathematically formulate error propagation in the architecture of DMs, The framework contains three elements, including modular error, cumulative error, and propagation equation. The modular and cumulative errors are related by the equation, which interprets that DMs are indeed affected by error propagation. Our theoretical study also suggests that the cumulative error is closely related to the generation quality of DMs. Based on this finding, we apply the cumulative error as a regularization term to reduce error propagation. Because the term is computationally intractable, we derive its upper bound and design a bootstrap algorithm to efficiently estimate the bound for optimization. We have conducted extensive experiments on multiple image datasets, showing that our proposed regularization reduces error propagation, significantly improves vanilla DMs, and outperforms previous baselines.
****
Tatjana Chavdarova,Tong Yang,Matteo Pagliardini,Michael Jordan
A Primal-Dual Approach to Solving Variational Inequalities with General Constraints
Yang et al. (2023) recently showed how to use first-order gradient methods to solve general variational inequalities (VIs) under a limiting assumption that analytic solutions of specific subproblems are available. In this paper, we circumvent this assumption via a warm-starting technique where we solve subproblems approximately and initialize variables with the approximate solution found at the previous iteration.
We prove the convergence of this method and show that the gap function of the last iterate of the method decreases at a rate of $\mathcal{O}(\frac{1}{\sqrt{K}})

$ when the operator is L-Lipschitz and monotone.
In numerical experiments, we show that this technique can converge much faster than its exact counterpart.
Furthermore, for the cases when the inequality constraints are simple, we introduce an alternative variant of ACVI and establish its convergence under the same conditions.
Finally, we relax the smoothness assumptions in Yang et al., yielding, to our knowledge, the first convergence result for VIs with general constraints that does not rely on the assumption that the operator is L-Lipschitz.
****
Yi Li,Honghao Lin,David Woodruff
Optimal Sketching for Residual Error Estimation for Matrix and Vector Norms
We study the problem of residual error estimation for matrix and vector norms using a linear sketch. Such estimates can be used, for example, to quickly assess how useful a more expensive low-rank approximation computation will be. The matrix case concerns the Frobenius norm and the task is to approximate the k-residual $\|A - A_k\|_F$ of the input matrix A within a $(1+\epsilon)$-factor, where A_k is the optimal rank-k approximation. We provide a tight bound of $\Theta(k^2/\epsilon^4)$ on the size of bilinear sketches, which have the form of a matrix product SAT. This improves the previous $O(k^2/\epsilon^6)$ upper bound in (Andoni et al. SODA 2013) and gives the first non-trivial lower bound, to the best of our knowledge.
In our algorithm, our sketching matrices S and T can both be sparse matrices, allowing for a very fast update time.
We demonstrate that this gives a substantial advantage empirically, for roughly the same sketch size and accuracy as in previous work.

For the vector case, we consider the ℓ_p-norm for $p>2$, where the task is to approximate the k-residual $\|x - x_k\|_p$ up to a constant factor, where x_k is the optimal k-sparse approximation to x. Such vector norms are frequently studied in the data stream literature and are useful for finding frequent items or so-called heavy hitters. We establish an upper bound of $O(k^{2/p}n^{1-2/p}\operatorname{poly}(\log n))$ for constant ϵ on the dimension of a linear sketch for this problem. Our algorithm can be extended to the ℓ_p sparse recovery problem with the same sketching dimension, which seems to be the first such bound for $p > 2$. We also show an $\Omega(k^{2/p}n^{1-2/p})$ lower bound for the sparse recovery problem, which is tight up to a $\mathrm{poly}(\log n)$ factor.
****
Ibraheem Muhammad Moosa,Rui Zhang,Wenpeng Yin
MT-Ranker: Reference-free machine translation evaluation by inter-system ranking
Traditionally, Machine Translation (MT) Evaluation has been treated as a regression problem -- producing an absolute translation-quality score. This approach has two limitations: i) the scores lack interpretability, and human annotators struggle with giving consistent scores; ii) most scoring methods are based on (reference, translation) pairs, limiting their applicability in real-world scenarios where references are absent. In practice, we often care about whether a new MT system is better or worse than some competitors. In addition, reference-free MT evaluation is increasingly practical and necessary. Unfortunately, these two practical considerations have yet to be jointly explored. In this work, we formulate the reference-free MT evaluation into a pairwise ranking problem. Given the source sentence and a pair of translations, our system predicts which translation is better. In addition to proposing this new formulation, we further show that this new paradigm can demonstrate superior correlation with human judgments by merely using indirect supervision from natural language inference and weak supervision from our synthetic data. In the context of reference-free evaluation, MT-Ranker, trained without any human annotations, achieves state-of-the-art results on the WMT Shared Metrics Task benchmarks DARR20, MQM20, and MQM21. On a more challenging benchmark, ACES, which contains fine-grained evaluation criteria such as addition, omission, and mistranslation errors, MT-Ranker marks state-of-the-art

against reference-free as well as reference-based baselines.
****
Asaf Yehudai,Boaz Carmeli,Yosi Mass,Ofir Arviv,Nathaniel Mills,Eyal Shnarch,Lesh
em Choshen
Achieving Human Parity in Content-Grounded Datasets Generation
The lack of high-quality data for content-grounded generation tasks has been ide
ntified as a major obstacle to advancing these tasks. To address this gap, we pr
opose a novel method for automatically generating high-quality content-grounded
data. It consists of three stages: (a) Content Preparation, (b) Generation: crea
ting task-specific examples from the content (e.g., question-answer pairs or sum
maries). (c) Filtering mechanism aiming to ensure the quality and faithfulness o
f the generated data. We showcase this methodology by generating large-scale dat
a for synthetic Long-form question-answering (LFQA) and summarization. In a huma
n evaluation, our generated data was found to be natural and of high quality. Fu
rthermore, we compare models trained on our data with models trained on human-wr
itten data – ELI5 and ASQA for LFQA and CNN-DailyMail for Summarization. We show
 that our models are on par with or outperforming models trained on human-genera
ted data and consistently outperforming them in faithfulness. Finally, we applie
d our method to create LFQA data within the medical domain and compared a model
trained on it with models trained on other domains.
****
Guanting Chen,Xiaocheng Li,Chunlin Sun,Hanzhao Wang
Learning to Make Adherence-aware Advice
As artificial intelligence (AI) systems play an increasingly prominent role in h
uman decision-making, challenges surface in the realm of human-AI interactions.
One challenge arises from the suboptimal AI policies due to the inadequate consi
deration of humans disregarding AI recommendations, as well as the need for AI t
o provide advice selectively when it is most pertinent. This paper presents a se
quential decision-making model that (i) takes into account the human's adherence
 level (the probability that the human follows/rejects machine advice) and (ii)
incorporates a defer option so that the machine can temporarily refrain from mak
ing advice. We provide learning algorithms that learn the optimal advice policy
and make advice only at critical time stamps. Compared to problem-agnostic reinf
orcement learning algorithms, our specialized learning algorithms not only enjoy
 better theoretical convergence properties but also show strong empirical perfor
mance.
****
Cheng Tan,Yijie Zhang,Zhangyang Gao,Bozhen Hu,Siyuan Li,Zicheng Liu,Stan Z. Li
RDesign: Hierarchical Data-efficient Representation Learning for Tertiary Struct
ure-based RNA Design
While artificial intelligence has made remarkable strides in revealing the relat
ionship between biological macromolecules' primary sequence and tertiary structu
re, designing RNA sequences based on specified tertiary structures remains chall
enging. Though existing approaches in protein design have thoroughly explored st
ructure-to-sequence dependencies in proteins, RNA design still confronts difficu
lties due to structural complexity and data scarcity. Moreover, direct transplan
tation of protein design methodologies into RNA design fails to achieve satisfac
tory outcomes although sharing similar structural components. In this study, we
aim to systematically construct a data-driven RNA design pipeline. We crafted a
large, well-curated benchmark dataset and designed a comprehensive structural mo
deling approach to represent the complex RNA tertiary structure. More importantl
y, we proposed a hierarchical data-efficient representation learning framework t
hat learns structural representations through contrastive learning at both clust
er-level and sample-level to fully leverage the limited data. By constraining da
ta representations within a limited hyperspherical space, the intrinsic relation
ships between data points could be explicitly imposed. Moreover, we incorporated
 extracted secondary structures with base pairs as prior knowledge to facilitate
 the RNA design process. Extensive experiments demonstrate the effectiveness of
our proposed method, providing a reliable baseline for future RNA design tasks.
The source code and benchmark dataset are available at https://github.com/A4Bio/

RDesign.
****
Kai Shen,Zeqian Ju,Xu Tan,Eric Liu,Yichong Leng,Lei He,Tao Qin,sheng zhao,Jiang Bian
NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers
Scaling text-to-speech (TTS) to large-scale, multi-speaker, and in-the-wild datasets is important to capture the diversity in human speech such as speaker identities, prosodies, and styles (e.g., singing). Current large TTS systems usually quantize speech into discrete tokens and use language models to generate these tokens one by one, which suffer from unstable prosody, word skipping/repeating issue, and poor voice quality. In this paper, we develop NaturalSpeech 2, a TTS system that leverages a neural audio codec with residual vector quantizers to get the quantized latent vectors and uses a diffusion model to generate these latent vectors conditioned on text input. To enhance the zero-shot capability that is important to achieve diverse speech synthesis, we design a speech prompting mechanism to facilitate in-context learning in the diffusion model and the duration/pitch predictor. We scale NaturalSpeech 2 to large-scale datasets with 44K hours of speech and singing data and evaluate its voice quality on unseen speakers. NaturalSpeech 2 outperforms previous TTS systems by a large margin in terms of prosody/timbre similarity, robustness, and voice quality in a zero-shot setting, and performs novel zero-shot singing synthesis with only a speech prompt. Audio samples are available at https://naturalspeech2.github.io/.
****
David Valensi,Esther Derman,Shie Mannor,Gal Dalal
Tree Search-Based Policy Optimization under Stochastic Execution Delay
The standard formulation of Markov decision processes (MDPs) assumes that the agent's decisions are executed immediately.
However, in numerous realistic applications such as robotics or healthcare, actions are performed with a delay whose value can even be stochastic. In this work, we introduce stochastic delayed execution MDPs, a new formalism addressing random delays without resorting to state augmentation. We show that given observed delay values, it is sufficient to perform a policy search in the class of Markov policies in order to reach optimal performance, thus extending the deterministic fixed delay case. Armed with this insight, we devise DEZ, a model-based algorithm that optimizes over the class of Markov policies. DEZ leverages Monte-Carlo tree search similar to its non-delayed variant EfficientZero to accurately infer future states from the action queue. Thus, it handles delayed execution while preserving the sample efficiency of EfficientZero. Through empirical analysis, we stress that none of the prior benchmarks consistently outperforms others across different delays. We demonstrate that our algorithm surpasses all benchmark methods in Atari games when dealing with constant or stochastic delays. The code is available at \url{https://github.com/davidva1/Delayed-EZ}.
****
Xiong Zhou,Xianming Liu,Feilong Zhang,Gang Wu,Deming Zhai,Junjun Jiang,Xiangyang Ji
Zero-Mean Regularized Spectral Contrastive Learning: Implicitly Mitigating Wrong Connections in Positive-Pair Graphs
Contrastive learning has emerged as a popular paradigm of self-supervised learning that learns representations by encouraging representations of positive pairs to be similar while representations of negative pairs to be far apart. The spectral contrastive loss, in synergy with the notion of positive-pair graphs, offers valuable theoretical insights into the empirical successes of contrastive learning. In this paper, we propose incorporating an additive factor into the term of spectral contrastive loss involving negative pairs. This simple modification can be equivalently viewed as introducing a regularization term that enforces the mean of representations to be zero, which thus is referred to as *zero-mean regularization*. It intuitively relaxes the orthogonality of representations between negative pairs and implicitly alleviates the adverse effect of wrong connections in the positive-pair graph, leading to better performance and robustness. To c

larify this, we thoroughly investigate the role of zero-mean regularized spectra
l contrastive loss in both unsupervised and supervised scenarios with respect to
 theoretical analysis and quantitative evaluation. These results highlight the p
otential of zero-mean regularized spectral contrastive learning to be a promisin
g approach in various tasks.
****
Xiaoxin He,Xavier Bresson,Thomas Laurent,Adam Perold,Yann LeCun,Bryan Hooi
Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Grap
h Representation Learning
Representation learning on text-attributed graphs (TAGs) has become a critical r
esearch problem in recent years. A typical example of a TAG is a paper citation
graph, where the text of each paper serves as node attributes. Initial graph neu
ral network (GNN) pipelines handled these text attributes by transforming them i
nto shallow or hand-crafted features, such as skip-gram or bag-of-words features
. Recent efforts have focused on enhancing these pipelines with language models
(LMs), which typically demand intricate designs and substantial computational re
sources. With the advent of powerful large language models (LLMs) such as GPT or
 Llama2, which demonstrate an ability to reason and to utilize general knowledge
, there is a growing need for techniques which combine the textual modelling abi
lities of LLMs with the structural learning capabilities of GNNs. Hence, in this
 work, we focus on leveraging LLMs to capture textual information as features, w
hich can be used to boost GNN performance on downstream tasks. A key innovation
is our use of \emph{explanations as features}: we prompt an LLM to perform zero-
shot classification, request textual explanations for its decision-making proces
s, and design an \emph{LLM-to-LM interpreter} to translate these explanations in
to informative features for downstream GNNs. Our experiments demonstrate that ou
r method achieves state-of-the-art results on well-established TAG datasets, inc
luding \texttt{Cora}, \texttt{PubMed}, \texttt{ogbn-arxiv}, as well as our newly
 introduced dataset, \texttt{tape-arxiv23}. Furthermore, our method significantl
y speeds up training, achieving a 2.88 times improvement over the closest baseli
ne on \texttt{ogbn-arxiv}. Lastly, we believe the versatility of the proposed me
thod extends beyond TAGs and holds the potential to enhance other tasks involvin
g graph-text data~\footnote{Our codes and datasets are available at: \url{https:
//github.com/XiaoxinHe/TAPE}}.
****
Nanda H Krishna,Colin Bredenberg,Daniel Levenstein,Blake Aaron Richards,Guillaum
e Lajoie
Sufficient conditions for offline reactivation in recurrent neural networks
During periods of quiescence, such as sleep, neural activity in many brain circu
its resembles that observed during periods of task engagement. However, the prec
ise conditions under which task-optimized networks can autonomously reactivate t
he same network states responsible for online behavior is poorly understood. In
this study, we develop a mathematical framework that outlines sufficient conditi
ons for the emergence of neural reactivation in circuits that encode features of
 smoothly varying stimuli. We demonstrate mathematically that noisy recurrent ne
tworks optimized to track environmental state variables using change-based senso
ry information naturally develop denoising dynamics, which, in the absence of in
put, cause the network to revisit state configurations observed during periods o
f online activity. We validate our findings using numerical experiments on two c
anonical neuroscience tasks: spatial position estimation based on self-motion cu
es, and head direction estimation based on angular velocity cues. Overall, our w
ork provides theoretical support for modeling offline reactivation as an emergen
t consequence of task optimization in noisy neural circuits.
****
Zuxin Liu,Jesse Zhang,Kavosh Asadi,Yao Liu,Ding Zhao,Shoham Sabach,Rasool Fakoor
TAIL: Task-specific Adapters for Imitation Learning with Large Pretrained Models
The full potential of large pretrained models remains largely untapped in contro
l domains like robotics. This is mainly because of the scarcity of data and the
computational challenges associated with training or fine-tuning these large mod
els for such applications. Prior work mainly emphasizes either effective \emph{p

retraining} of large models for decision-making or single-task adaptation. But r
eal-world problems will require data-efficient, \emph{continual adaptation} for
new control tasks. Recognizing these constraints, we introduce TAIL (Task-specif
ic Adapters for Imitation Learning), a framework for efficient adaptation to new
 control tasks. Inspired by recent advancements in parameter-efficient fine-tuni
ng in language domains, we explore efficient fine-tuning techniques---e.g., Bott
leneck Adapters, P-Tuning, and Low-Rank Adaptation (LoRA)---in TAIL to adapt lar
ge pretrained models for new tasks with limited demonstration data. Our extensiv
e experiments comparing prevalent parameter-efficient fine-tuning techniques and
 adaptation baselines suggest that TAIL with LoRA can achieve the best post-adap
tation performance with only 1\% of the trainable parameters of full fine-tuning
, while avoiding catastrophic forgetting and preserving adaptation plasticity in
 continual learning settings.
****
Behzad Shayegh,Yanshuai Cao,Xiaodan Zhu,Jackie CK Cheung,Lili Mou
Ensemble Distillation for Unsupervised Constituency Parsing
We investigate the unsupervised constituency parsing task, which organizes words
 and phrases of a sentence into a hierarchical structure without using linguisti
cally annotated data. We observe that existing unsupervised parsers capture diff
erent aspects of parsing structures, which can be leveraged to enhance unsupervi
sed parsing performance.
To this end, we propose a notion of "tree averaging," based on which we further
propose a novel ensemble method for unsupervised parsing.
To improve inference efficiency, we further distill the ensemble knowledge into
a student model; such an ensemble-then-distill process is an effective approach
to mitigate the over-smoothing problem existing in common multi-teacher distilli
ng methods.
Experiments show that our method surpasses all previous approaches, consistently
 demonstrating its effectiveness and robustness across various runs, with differ
ent ensemble components, and under domain-shift conditions.
****
Ioannis Mavrothalassitis,Stratis Skoulakis,Leello Tadesse Dadi,Volkan Cevher
Efficient Continual Finite-Sum Minimization
Given a sequence of functions $f_1,\ldots,f_n$ with $f_i:\mathcal{D}\mapsto \mat
hbb{R}$, finite-sum minimization seeks a point ${x}^\star \in \mathcal{D}$ minim
izing $\sum_{j=1}^nf_j(x)/n$. In this work, we propose a key twist into the fini
te-sum minimization, dubbed as *continual finite-sum minimization*, that asks fo
r a sequence of points $x_1^\star, \ldots, x_n^\star \in D$ such that each ${x}^
\star_i \in D$ minimizes the prefix-sum $\sum_{j=1}^if_j(x)/i$. Assuming that ea
ch prefix-sum is strongly convex, we develop a first-order continual stochastic
variance reduction gradient method (CSVRG) producing an ϵ-op
timal sequence with $\tilde{\mathcal{O}}(n/\epsilon^{1/3} + 1/\sqrt{\epsilon})$
overall *first-order oracles* (FO). An FO corresponds to the computation of a si
ngle gradient $\nabla f_j(x)$ at a given $x \in \mathcal{D}$ for some $j \in [n]
$. Our approach significantly improves upon the $\mathcal{O}(n/\epsilon)$ FOs th
at $\mathrm{StochasticGradientDescent}$ requires and the $\mathcal{O}(n^2 \log (
1/\epsilon))$ FOs that state-of-the-art variance reduction methods such as $\mat
hrm{Katyusha}$ require. We also prove that there is no natural first-order metho
d with $\mathcal{O}\left(n/\epsilon^\alpha\right)$ gradient complexity for $\alp
ha < 1/4$, establishing that the first-order complexity of our method is nearly
tight.
****
Yifeng Fan,Yongqiang Li,Bo Chen
Weaker MVI Condition: Extragradient Methods with Multi-Step Exploration
This paper proposes a new framework of algorithms that is extended from the cele
brated extragradient algorithm. The min-max problem has attracted increasing att
ention because of its applications in machine learning tasks such as generative
adversarial networks (GANs) training. While there has been exhaustive research o
n convex-concave setting, problem of nonconvex-nonconcave setting faces many cha
llenges, such as convergence to limit cycles. Given that general min-max optimiz

ation has been found to be intractable, recent research efforts have shifted tow ards tackling structured problems. One of these follows the weak Minty variation al inequality (weak MVI), which is motivated by relaxing Minty variational inequ ality (mvi) without compromising convergence guarantee of extragradient algorith m. Existing extragradient-type algorithms involve one exploration step and one u pdate step per iteration. We analyze the algorithms with multiple exploration st eps and show that current assumption can be further relaxed when more exploratio n is introduced. Furthermore, we design an adaptive algorithm that explores unti l the optimal improvement is achieved. This process exploits information from th e whole trajectory and effectively tackles cyclic behaviors.
****
Yifei Zhou,Ayush Sekhari,Yuda Song,Wen Sun
Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees
Hybrid RL is the setting where an RL agent has access to both offline data and o nline data by interacting with the real-world environment. In this work, we prop ose a new hybrid RL algorithm that combines an on-policy actor-critic method wit h offline data. On-policy methods such as policy gradient and natural policy gra dient (NPG) have shown to be more robust to model misspecification, though somet imes it may not be as sample efficient as methods that rely on off-policy learni ng. On the other hand, offline methods that depend on off-policy training often require strong assumptions in theory and are less stable to train in practice. O ur new approach integrates a procedure of off-policy training on the offline dat a into an on-policy NPG framework. We show that our approach, in theory, can obt ain a *best-of-both-worlds* type of result --- it achieves the state-of-art theo retical guarantees of offline RL when offline RL-specific assumptions hold, whil e at the same time maintaining the theoretical guarantees of on-policy NPG regar dless of the offline RL assumptions' validity. Experimentally, in challenging ri ch-observation environments, we show that our approach outperforms a state-of-th e-art hybrid RL baseline which only relies on off-policy policy optimization, de monstrating the empirical benefit of combining on-policy and off-policy learning .
****
Jungtaek Kim,Jeongbeen Yoon,Minsu Cho
Generalized Neural Sorting Networks with Error-Free Differentiable Swap Function s
Sorting is a fundamental operation of all computer systems, having been a long-s tanding significant research topic. Beyond the problem formulation of traditiona l sorting algorithms, we consider sorting problems for more abstract yet express ive inputs, e.g., multi-digit images and image fragments, through a neural sorti ng network. To learn a mapping from a high-dimensional input to an ordinal varia ble, the differentiability of sorting networks needs to be guaranteed. In this p aper we define a softening error by a differentiable swap function, and develop an error-free swap function that holds a non-decreasing condition and differenti ability. Furthermore, a permutation-equivariant Transformer network with multi-h ead attention is adopted to capture dependency between given inputs and also lev erage its model capacity with self-attention. Experiments on diverse sorting ben chmarks show that our methods perform better than or comparable to baseline meth ods.
****
Hannah Kniesel,Leon Sick,Tristan Payer,Tim Bergner,Kavitha Shaga Devan,Clarissa Read,Paul Walther,Timo Ropinski,Pedro Hermosilla
Weakly Supervised Virus Capsid Detection with Image-Level Annotations in Electro n Microscopy Images
Current state-of-the-art methods for object detection rely on annotated bounding boxes of large data sets for training. However, obtaining such annotations is e xpensive and can require up to hundreds of hours of manual labor. This poses a c hallenge, especially since such annotations can only be provided by experts, as they require knowledge about the scientific domain. To tackle this challenge, we propose a domain-specific weakly supervised object detection algorithm that onl y relies on image-level annotations, which are significantly easier to acquire.

Our method distills the knowledge of a pre-trained model, on the task of predic
ting the presence or absence of a virus in an image, to obtain a set of pseudo-l
abels that can be used to later train a state-of-the-art object detection model.
 To do so, we use an optimization approach with a shrinking receptive field to e
xtract virus particles directly without specific network architectures. Through
a set of extensive studies, we show how the proposed pseudo-labels are easier to
 obtain, and, more importantly, are able to outperform other existing weak label
ing methods, and even ground truth labels, in cases where the time to obtain the
 annotation is limited.
****
Mahsa Keramati,Lili Meng,R. David Evans
ConR: Contrastive Regularizer for Deep Imbalanced Regression
Imbalanced distributions are ubiquitous in real-world data. They create constrai
nts on Deep Neural Networks to represent the minority labels and avoid bias towa
rds majority labels. The extensive body of imbalanced approaches address categor
ical label spaces but fail to effectively extend to regression problems where th
e label space is continuous. Local and global correlations among continuous labe
ls provide valuable insights towards effectively modelling relationships in feat
ure space. In this work, we propose ConR, a contrastive regularizer that models
global and local label similarities in feature space and prevents the features o
f minority samples from being collapsed into their majority neighbours. ConR dis
cerns the disagreements between the label space and feature space, and imposes
a penalty on these disagreements. ConR minds the continuous nature of label spac
e with two main strategies in a contrastive manner: incorrect proximities are pe
nalized proportionate to the label similarities and the correct ones are encoura
ged to model local similarities. ConR consolidates essential considerations into
 a generic, easy-to-integrate, and efficient method that effectively addresses d
eep imbalanced regression. Moreover, ConR is orthogonal to existing approaches a
nd smoothly extends to uni- and multi-dimensional label spaces. Our comprehensiv
e experiments show that ConR significantly boosts the performance of all the sta
te-of-the-art methods on four large-scale deep imbalanced regression benchmarks.
****
Melanie Sclar,Yejin Choi,Yulia Tsvetkov,Alane Suhr
Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design o
r: How I learned to start worrying about prompt formatting
As large language models (LLMs) are adopted as a fundamental component of langua
ge technologies, it is crucial to accurately characterize their performance. Bec
ause choices in prompt design can strongly influence model behavior, this design
 process is critical in effectively using any modern pre-trained generative lang
uage model. In this work, we focus on LLM sensitivity to a quintessential class
of meaning-preserving design choices: prompt formatting. We find that several wi
dely used open-source LLMs are extremely sensitive to subtle changes in prompt f
ormatting in few-shot settings, with performance differences of up to 76 accurac
y points when evaluated using LLaMA-2-13B. Sensitivity remains even when increas
ing model size, the number of few-shot examples, or performing instruction tunin
g. Our analysis suggests that work evaluating LLMs with prompting-based methods
would benefit from reporting a range of performance across plausible prompt form
ats, instead of the currently-standard practice of reporting performance on a si
ngle format. We also show that format performance only weakly correlates between
 models, which puts into question the methodological validity of comparing model
s with an arbitrarily chosen, fixed prompt format. To facilitate systematic anal
ysis we propose FormatSpread, an algorithm that rapidly evaluates a sampled set
of plausible prompt formats for a given task, and reports the interval of expect
ed performance without accessing model weights. Furthermore, we present a suite
of analyses that characterize the nature of this sensitivity, including explorin
g the influence of particular atomic perturbations and the internal representati
on of particular formats.
****
Yitian Zhang,Yue Bai,Huan Wang,Yizhou Wang,Yun Fu
Don't Judge by the Look: A Motion Coherent Augmentation for Video Recognition

Current training pipelines in object recognition neglect Hue Jittering when doin
g data augmentation as it not only brings appearance changes that are detrimenta
l to classification, but also the implementation is inefficient in practice. In
this study, we investigate the effect of hue variance in the context of video re
cognition and find this variance to be beneficial since static appearances are l
ess important in videos that contain motion information. Based on this observati
on, we propose a data augmentation method for video recognition, named Motion Co
herent Augmentation (MCA), that introduces appearance variation in videos and im
plicitly encourages the model to prioritize motion patterns, rather than static
appearances. Concretely, we propose an operation SwapMix to efficiently modify t
he appearance of video samples, and introduce Variation Alignment (VA) to resolv
e the distribution shift caused by SwapMix, enforcing the model to learn appeara
nce invariant representations. Comprehensive empirical evaluation across various
 architectures and different datasets solidly validates the effectiveness and ge
neralization ability of MCA, and the application of VA in other augmentation met
hods. Code is available at https://github.com/BeSpontaneous/MCA-pytorch.
****
Yunhui Jang,Dongwoo Kim,Sungsoo Ahn
Graph Generation with K^2-trees
Generating graphs from a target distribution is a significant challenge across m
any domains, including drug discovery and social network analysis. In this work,
 we introduce a novel graph generation method leveraging K^2 representation, o
riginally designed for lossless graph compression. The K^2 representation enab
les compact generation while concurrently capturing an inherent hierarchical str
ucture of a graph. In addition, we make contributions by (1) presenting a sequen
tial K^2 representation that incorporates pruning, flattening, and tokenizatio
n processes and (2) introducing a Transformer-based architecture designed to gen
erate the sequence by incorporating a specialized tree positional encoding schem
e. Finally, we extensively evaluate our algorithm on four general and two molecu
lar graph datasets to confirm its superiority for graph generation.
****
Joongkyu Lee,Min-hwan Oh
Demystifying Linear MDPs and Novel Dynamics Aggregation Framework
In this work, we prove that, in linear MDPs, the feature dimension d is lower
bounded by S/U in order to aptly represent transition probabilities, where S
 is the size of the state space and U is the maximum size of directly reachabl
e states.
Hence, d can still scale with S depending on the direct reachability of the
environment. To address this limitation of linear MDPs, we propose a novel stru
ctural aggregation framework based on dynamics, named as the *dynamics aggregati
on*.
    For this newly proposed framework,
    we design a provably efficient hierarchical reinforcement learning algorithm
 in linear function approximation that leverages aggregated sub-structures. Our
proposed algorithm exhibits statistical efficiency, achieving a regret of $\tild
e{O} \big(d_{\psi}^{3/2} H^{3/2}\sqrt{ NT} \big)$, where d_{ψ} represents
the feature dimension of *aggregated subMDPs* and N signifies the number of ag
gregated subMDPs.
    We establish that the condition $d_{\psi}^3 N \ll d^{3}$ is readily met in m
ost real-world environments with hierarchical structures, enabling a substantial
 improvement in the regret bound compared to LSVI-UCB, which enjoys a regret of
$\tilde{O}(d^{3/2} H^{3/2} \sqrt{ T})$.
    To the best of our knowledge, this work presents the first HRL algorithm wit
h linear function approximation that offers provable guarantees.
****
Kai Xu,Rongyu Chen,Gianni Franchi,Angela Yao
Scaling for Training Time and Post-hoc Out-of-distribution Detection Enhancement
Activation shaping has proven highly effective for identifying out-of-distributi
on (OOD) samples post-hoc. Activation shaping prunes and scales network activati
ons before estimating the OOD energy score; such an extremely simple approach ac

hieves state-of-the-art OOD detection with minimal in-distribution (ID) accuracy drops. This paper analyzes the working mechanism behind activation shaping. We directly show that the benefits for OOD detection derive only from scaling, while pruning is detrimental. Based on our analysis, we propose SCALE, an even simpler yet more effective post-hoc network enhancement method for OOD detection. SCALE attains state-of-the-art OOD detection performance without any compromises on ID accuracy. Furthermore, we integrate scaling concepts into learning and propose Intermediate Tensor SHaping (ISH) for training-time OOD detection enhancement. ISH achieves significant AUROC improvements for both near- and far-OOD, highlighting the importance of activation distributions in emphasizing ID data characteristics. Our code and models are available at https://github.com/kai422/SCALE.
****
Zhe Wu,Haofei Lu,Junliang Xing,You Wu,Renye Yan,Yaozhong Gan,Yuanchun Shi
PAE: Reinforcement Learning from External Knowledge for Efficient Exploration
Human intelligence is adept at absorbing valuable insights from external knowledge.
This capability is equally crucial for artificial intelligence.
In contrast, classical reinforcement learning agents lack such capabilities and often resort to extensive trial and error to explore the environment.
This paper introduces $\textbf{PAE}$: $\textbf{P}$lanner-$\textbf{A}$ctor-$\textbf{E}$valuator, a novel framework for teaching agents to $\textit{learn to absorb external knowledge}$.
PAE integrates the Planner's knowledge-state alignment mechanism, the Actor's mutual information skill control, and the Evaluator's adaptive intrinsic exploration reward to achieve 1) effective cross-modal information fusion, 2) enhanced linkage between knowledge and state, and 3) hierarchical mastery of complex tasks.
Comprehensive experiments across
11 challenging tasks from the BabyAI and MiniHack environment suites demonstrate PAE's superior exploration efficiency with good interpretability.
****
Kaustubh Sridhar,Souradeep Dutta,Dinesh Jayaraman,James Weimer,Insup Lee
Memory-Consistent Neural Networks for Imitation Learning
Imitation learning considerably simplifies policy synthesis compared to alternative approaches by exploiting access to expert demonstrations. For such imitation policies, errors away from the training samples are particularly critical. Even rare slip-ups in the policy action outputs can compound quickly over time, since they lead to unfamiliar future states where the policy is still more likely to err, eventually causing task failures. We revisit simple supervised "behavior cloning" for conveniently training the policy from nothing more than pre-recorded demonstrations, but carefully design the model class to counter the compounding error phenomenon. Our "memory-consistent neural network" (MCNN) outputs are hard-constrained to stay within clearly specified permissible regions anchored to prototypical "memory" training samples. We provide a guaranteed upper bound for the sub-optimality gap induced by MCNN policies. Using MCNNs on 10 imitation learning tasks, with MLP, Transformer, and Diffusion backbones, spanning dexterous robotic manipulation and driving, proprioceptive inputs and visual inputs, and varying sizes and types of demonstration data, we find large and consistent gains in performance, validating that MCNNs are better-suited than vanilla deep neural networks for imitation learning applications. Website: https://sites.google.com/view/mcnn-imitation
****
Yue Huang,Jiawen Shi,Yuan Li,Chenrui Fan,Siyuan Wu,Qihui Zhang,Yixin Liu,Pan Zhou,Yao Wan,Neil Zhenqiang Gong,Lichao Sun
MetaTool Benchmark for Large Language Models: Deciding Whether to Use Tools and Which to Use
Large language models (LLMs) have garnered significant attention due to their impressive natural language processing (NLP) capabilities. Recently, many studies have focused on the tool utilization ability of LLMs. They primarily investigated how LLMs effectively collaborate with given specific tools. However, in scenarios where LLMs serve as intelligent agents, as seen in applications like AutoGPT

and MetaGPT, LLMs are expected to engage in intricate decision-making processes that involve deciding whether to employ a tool and selecting the most suitable tool(s) from a collection of available tools to fulfill user requests. Therefore, in this paper, we introduce MetaTool, a benchmark designed to evaluate whether LLMs have tool usage awareness and can correctly choose tools. Specifically, we create a dataset called ToolE within the benchmark. This dataset contains various types of user queries in the form of prompts that trigger LLMs to use tools, including both single-tool and multi-tool scenarios. Subsequently, we set the tasks for both tool usage awareness and tool selection. We define four subtasks from different perspectives in tool selection, including tool selection with similar choices, tool selection in specific scenarios, tool selection with possible reliability issues, and multi-tool selection. We conduct experiments involving eight popular LLMs and find that the majority of them still struggle to effectively select tools, highlighting the existing gaps between LLMs and genuine intelligent agents. However, through the error analysis, we found there is still significant room for improvement. Finally, we conclude with insights for tool developers -- we strongly recommend that tool developers choose an appropriate rewrite model for generating new descriptions based on the downstream LLM the tool will apply to.
****
Weibang Jiang,Liming Zhao,Bao-liang Lu
Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI
The current electroencephalogram (EEG) based deep learning models are typically designed for specific datasets and applications in brain-computer interaction (BCI), limiting the scale of the models and thus diminishing their perceptual capabilities and generalizability. Recently, Large Language Models (LLMs) have achieved unprecedented success in text processing, prompting us to explore the capabilities of Large EEG Models (LEMs). We hope that LEMs can break through the limitations of different task types of EEG datasets, and obtain universal perceptual capabilities of EEG signals through unsupervised pre-training. Then the models can be fine-tuned for different downstream tasks. However, compared to text data, the volume of EEG datasets is generally small and the format varies widely. For example, there can be mismatched numbers of electrodes, unequal length data samples, varied task designs, and low signal-to-noise ratio. To overcome these challenges, we propose a unified foundation model for EEG called Large Brain Model (LaBraM). LaBraM enables cross-dataset learning by segmenting the EEG signals into EEG channel patches. Vector-quantized neural spectrum prediction is used to train a semantically rich neural tokenizer that encodes continuous raw EEG channel patches into compact neural codes. We then pre-train neural Transformers by predicting the original neural codes for the masked EEG channel patches. The LaBraMs were pre-trained on about 2,500 hours of various types of EEG signals from around 20 datasets and validated on multiple different types of downstream tasks. Experiments on abnormal detection, event type classification, emotion recognition, and gait prediction show that our LaBraM outperforms all compared SOTA methods in their respective fields. Our code is available at https://github.com/935963004/LaBraM.
****
Yuan Yuan,Chenyang Shao,Jingtao Ding,Depeng Jin,Yong Li
Spatio-Temporal Few-Shot Learning via Diffusive Neural Network Generation
Spatio-temporal modeling is foundational for smart city applications, yet it is often hindered by data scarcity in many cities and regions. To bridge this gap, we propose a novel generative pre-training framework, GPD, for spatio-temporal few-shot learning with urban knowledge transfer. Unlike conventional approaches that heavily rely on common feature extraction or intricate few-shot learning designs, our solution takes a novel approach by performing generative pre-training on a collection of neural network parameters optimized with data from source cities. We recast spatio-temporal few-shot learning as pre-training a generative diffusion model, which generates tailored neural networks guided by prompts, allowing for adaptability to diverse data distributions and city-specific characteri

stics. GPD employs a Transformer-based denoising diffusion model, which is model-agnostic to integrate with powerful spatio-temporal neural networks. By addressing challenges arising from data gaps and the complexity of generalizing knowledge across cities, our framework consistently outperforms state-of-the-art baselines on multiple real-world datasets for tasks such as traffic speed prediction and crowd flow prediction. The implementation of our approach is available: https://github.com/tsinghua-fib-lab/GPD.
****
Albert Bou,Matteo Bettini,Sebastian Dittert,Vikash Kumar,Shagun Sodhani,Xiaomeng Yang,Gianni De Fabritiis,Vincent Moens
TorchRL: A data-driven decision-making library for PyTorch
PyTorch has ascended as a premier machine learning framework, yet it lacks a native and comprehensive library for decision and control tasks suitable for large development teams dealing with complex real-world data and environments. To address this issue, we propose TorchRL, a generalistic control library for PyTorch that provides well-integrated, yet standalone components. We introduce a new and flexible PyTorch primitive, the TensorDict, which facilitates streamlined algorithm development across the many branches of Reinforcement Learning (RL) and control. We provide a detailed description of the building blocks and an extensive overview of the library across domains and tasks. Finally, we experimentally demonstrate its reliability and flexibility, and show comparative benchmarks to demonstrate its computational efficiency. TorchRL fosters long-term support and is publicly available on GitHub for greater reproducibility and collaboration within the research community. The code is open-sourced on GitHub.
****
Ivan Lee,Nan Jiang,Taylor Berg-Kirkpatrick
Is attention required for ICL? Exploring the Relationship Between Model Architecture and In-Context Learning Ability
What is the relationship between model architecture and the ability to perform in-context learning? In this empirical study, we take the first steps toward answering this question. We evaluate thirteen model architectures capable of causal language modeling across a suite of synthetic in-context learning tasks. These selected architectures represent a broad range of paradigms, including recurrent and convolution-based neural networks, transformers, state-space model inspired, and other emerging attention alternatives. We discover that all the considered architectures can perform in-context learning under a wider range of conditions than previously documented. Additionally, we observe stark differences in statistical efficiency and consistency by varying context length and task difficulty. We also measure each architecture's predisposition towards in-context learning when presented with alternative routes for task resolution. Finally, and somewhat surprisingly, we find that several attention alternatives are more robust in-context learners than transformers. Given that such approaches have constant-sized memory footprints at inference time, this result opens the possibility of scaling up in-context learning to accommodate vastly larger numbers of in-context examples.
****
Xiaoming Zhao,R Alex Colburn,Fangchang Ma,Miguel Ángel Bautista,Joshua M. Susskind,Alex Schwing
Pseudo-Generalized Dynamic View Synthesis from a Video
Rendering scenes observed in a monocular video from novel viewpoints is a challenging problem. For static scenes the community has studied both scene-specific optimization techniques, which optimize on every test scene, and generalized techniques, which only run a deep net forward pass on a test scene. In contrast, for dynamic scenes, scene-specific optimization techniques exist, but, to our best knowledge, there is currently no generalized method for dynamic novel view synthesis from a given monocular video. To explore whether generalized dynamic novel view synthesis from monocular videos is possible today, we establish an analysis framework based on existing techniques and work toward the generalized approach. We find a pseudo-generalized process without scene-specific \emph{appearance} optimization is possible, but geometrically and temporally consistent depth esti

mates are needed. Despite no scene-specific appearance optimization, the pseudo-generalized approach improves upon some scene-specific methods.For more information see project page at https://xiaoming-zhao.github.io/projects/pgdvs.
****
Karim Ahmed Abdel Sadek,Marek Elias
Algorithms for Caching and MTS with reduced number of predictions
ML-augmented algorithms utilize predictions to achieve performance beyond their worst-case bounds. Producing these predictions might be a costly operation – this motivated Im et al. [2022] to introduce the study of algorithms which use predictions parsimoniously. We design parsimonious algorithms for caching and MTS with action predictions, proposed by Antoniadis et al. [2023], focusing on the parameters of consistency (performance with perfect predictions) and smoothness (dependence of their performance on prediction error). Our algorithm for caching is 1-consistent, robust, and its smoothness deteriorates with decreasing number of available predictions. We propose an algorithm for general MTS whose consistency and smoothness both scale linearly with the decreasing number of predictions. Without restriction on the number of available predictions, both algorithms match the earlier guarantees achieved by Antoniadis et al. [2023].
****
Keivan Rezaei,Mehrdad Saberi,Mazda Moayeri,Soheil Feizi
PRIME: Prioritizing Interpretability in Failure Mode Extraction
In this work, we study the challenge of providing human-understandable descriptions for failure modes in trained image classification models.
Existing works address this problem by first identifying clusters (or directions) of incorrectly classified samples in a latent space and then aiming to provide human-understandable text descriptions for them.
We observe that in some cases, describing text does not match well
with identified failure modes, partially owing to the fact that shared interpretable attributes of failure modes may not be captured using clustering in the feature space.
To improve on these shortcomings, we propose a novel approach that prioritizes interpretability in this problem: we start by obtaining human-understandable concepts (tags) of images in the dataset and
then analyze the model's behavior based on the presence or absence of combinations of these tags.
Our method also ensures that the tags describing a failure mode form a minimal set,
avoiding redundant and noisy descriptions.
Through several experiments on different datasets, we show that our method successfully identifies failure modes and generates high-quality text descriptions associated with them.
These results highlight the importance of prioritizing interpretability in understanding model failures.
****
CHEN CHEN,Ruizhe Li,Yuchen Hu,Sabato Marco Siniscalchi,Pin-Yu Chen,Ensiong Chng,Chao-Han Huck Yang
It's Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition
Recent studies have successfully shown that large language models (LLMs) can be successfully used for generative error correction (GER) on top of the automatic speech recognition (ASR) output. Specifically, an LLM is utilized to carry out a direct mapping from the N-best hypotheses list generated by an ASR system to the predicted output transcription. However, despite its effectiveness, GER introduces extra data uncertainty since the LLM is trained without taking into account acoustic information available in the speech signal. In this work, we aim to overcome such a limitation by infusing acoustic information before generating the predicted transcription through a novel late fusion solution termed Uncertainty-Aware Dynamic Fusion (UADF). UADF is a multimodal fusion approach implemented into an auto-regressive decoding process and works in two stages: (i) It first analyzes and calibrates the token-level LLM decision, and (ii) it then dynamically

assimilates the information from the acoustic modality. Experimental evidence co llected from various ASR tasks shows that UADF surpasses existing fusion mechani sms in several ways. It yields significant improvements in word error rate (WER) while mitigating data uncertainty issues in LLM and addressing the poor general ization relied with sole modality during fusion. We also demonstrate that UADF s eamlessly adapts to audio-visual speech recognition.
****
Shengchao Hu,Li Shen,Ya Zhang,Dacheng Tao
Learning Multi-Agent Communication from Graph Modeling Perspective
In numerous artificial intelligence applications, the collaborative efforts of m ultiple intelligent agents are imperative for the successful attainment of targe t objectives. To enhance coordination among these agents, a distributed communic ation framework is often employed. However, information sharing among all agents proves to be resource-intensive, while the adoption of a manually pre-defined c ommunication architecture imposes limitations on inter-agent communication, ther eby constraining the potential for collaborative efforts. In this study, we intr oduce a novel approach wherein we conceptualize the communication architecture a mong agents as a learnable graph. We formulate this problem as the task of deter mining the communication graph while enabling the architecture parameters to upd ate normally, thus necessitating a bi-level optimization process. Utilizing cont inuous relaxation of the graph representation and incorporating attention units, our proposed approach, CommFormer, efficiently optimizes the communication grap h and concurrently refines architectural parameters through gradient descent in an end-to-end manner. Extensive experiments on a variety of cooperative tasks su bstantiate the robustness of our model across diverse cooperative scenarios, whe re agents are able to develop more coordinated and sophisticated strategies rega rdless of changes in the number of agents.
****
Samyadeep Basu,Nanxuan Zhao,Vlad I Morariu,Soheil Feizi,Varun Manjunatha
Localizing and Editing Knowledge In Text-to-Image Generative Models
Text-to-Image Diffusion Models such as Stable-Diffusion and Imagen have achieved unprecedented quality of photorealism with state-of-the-art FID scores on MS-CO CO and other generation benchmarks. Given a caption, image generation requires f ine-grained knowledge about attributes such as object structure, style, and view point amongst others. Where does this information reside in text-to-image genera tive models? In our paper, we tackle this question and understand how knowledge corresponding to distinct visual attributes is stored in large-scale text-to-ima ge diffusion models. We adapt Causal Mediation Analysis for text-to-image models and trace knowledge about distinct visual attributes to various (causal) compon ents in the (i) UNet and (ii) text-encoder of the diffusion model.
In particular, we show that unlike large-language models, knowledge about differ ent attributes is not localized in isolated components, but is instead distribut ed amongst a set of components in the conditional UNet. These sets of components are often distinct for different visual attributes (e.g., style} / objects).
Remarkably, we find that the text-encoder in public text-to-image models such as Stable-Diffusion contains {\it only} one causal state across different visual a ttributes, and this is the first self-attention layer corresponding to the last subject token of the attribute in the caption. This is in stark contrast to the causal states in other language models which are often the mid-MLP layers. Base d on this observation of only one causal state in the text-encoder, we introduce a fast, data-free model editing method DiffQuickFix which can effectively edit concepts (remove or update knowledge) in text-to-image models. DiffQuickFix can edit (ablate) concepts in under a second with a closed-form update, providing a significant 1000x speedup and comparable editing performance to existing fine-tu ning based editing methods.
****
Bin Zhu,Bin Lin,Munan Ning,Yang Yan,Jiaxi Cui,WANG HongFa,Yatian Pang,Wenhao Jia ng,Junwu Zhang,Zongwei Li,Cai Wan Zhang,Zhifeng Li,Wei Liu,Li Yuan
LanguageBind: Extending Video-Language Pretraining to N-modality by Language-bas ed Semantic Alignment

The video-language (VL) pretraining has achieved remarkable improvement in multiple downstream tasks. However, the current VL pretraining framework is hard to extend to multiple modalities (N modalities, N ≥ 3) beyond vision and language. We thus propose LanguageBind, taking the language as the bind across different modalities because the language modality is well-explored and contains rich semantics. Specifically, we freeze the language encoder acquired by VL pretraining and then train encoders for other modalities with contrastive learning. As a result, all modalities are mapped to a shared feature space, implementing multi-modal semantic alignment. While LanguageBind ensures that we can extend VL modalities to N modalities, we also need a high-quality dataset with alignment data pairs centered on language. We thus propose VIDAL-10M with 10 Million data with Video, Infrared, Depth, Audio and their corresponding Language. In our VIDAL-10M, all videos are from short video platforms with complete semantics rather than truncated segments from long videos, and all the video, depth, infrared, and audio modalities are aligned to their textual descriptions. LanguageBind has achieved superior performance on a wide range of 15 benchmarks covering video, audio, depth, and infrared. Moreover, multiple experiments have provided evidence for the effectiveness of LanguageBind in achieving indirect alignment and complementarity among diverse modalities.
****
Feiyang Kang,Hoang Anh Just,Yifan Sun,Himanshu Jahagirdar,Yuanzhi Zhang,Rongxing Du,Anit Kumar Sahu,Ruoxi Jia
Get more for less: Principled Data Selection for Warming Up Fine-Tuning in LLMs
This work focuses on leveraging and selecting from vast, unlabeled, open data to \emph{pre-fine-tune} a pre-trained language model. The goal is to minimize the need for costly domain-specific data for subsequent fine-tuning while achieving desired performance levels. While many data selection algorithms have been designed for small-scale applications, rendering them unsuitable for our context, some emerging methods do cater to language data scales. However, they often prioritize data that aligns with the target distribution. While this strategy may be effective when training a model from scratch, it can yield limited results when the model has already been pre-trained on a different distribution. Differing from prior work, our key idea is to select data that nudges the pre-training distribution closer to the target distribution. We show the optimality of this approach for fine-tuning tasks under certain conditions. We demonstrate the efficacy of our methodology across a diverse array of tasks, showing that it consistently surpasses other selection methods. Moreover, our proposed method is significantly faster than existing techniques, scaling to millions of samples within a single GPU hour. Our code is open-sourced \footnote{Code repository: \url{https://anonymous.4open.science/r/DV4LLM-D761/}}. While fine-tuning offers significant potential for enhancing performance across diverse tasks, its associated costs often limit its widespread adoption; with this work, we hope to lay the groundwork for cost-effective fine-tuning, making its benefits more accessible.
****
Jiahao Nie,Zhiwei He,Xudong Lv,Xueyi Zhou,Dong-Kyu Chae,Fei Xie
Towards Category Unification of 3D Single Object Tracking on Point Clouds
Category-specific models are provenly valuable methods in 3D single object tracking (SOT) regardless of Siamese or motion-centric paradigms. However, such over-specialized model designs incur redundant parameters, thus limiting the broader applicability of 3D SOT task. This paper first introduces unified models that can simultaneously track objects across all categories using a single network with shared model parameters. Specifically, we propose to explicitly encode distinct attributes associated to different object categories, enabling the model to adapt to cross-category data. We find that the attribute variances of point cloud objects primarily occur from the varying size and shape (e.g., large and square vehicles v.s. small and slender humans). Based on this observation, we design a novel point set representation learning network inheriting transformer architecture, termed AdaFormer, which adaptively encodes the dynamically varying shape and size information from cross-category data in a unified manner. We further incorporate the size and shape prior derived from the known template targets into the

model’s inputs and learning objective, facilitating the learning of unified representation. Equipped with such designs, we construct two category-unified models SiamCUT and MoCUT. Extensive experiments demonstrate that SiamCUT and MoCUT exhibit strong generalization and training stability. Furthermore, our category-unified models outperform the category-specific counterparts by a significant margin (e.g., on KITTI dataset, $\sim$12\% and $\sim$3\% performance gains on the Siamese and motion paradigms).
****
Hancheng Min,Enrique Mallada,Rene Vidal
Early Neuron Alignment in Two-layer ReLU Networks with Small Initialization
This paper studies the problem of training a two-layer ReLU network for binary classification using gradient flow with small initialization. We consider a training dataset with well-separated input vectors: Any pair of input data with the same label are positively correlated, and any pair with different labels are negatively correlated. Our analysis shows that, during the early phase of training, neurons in the first layer try to align with either the positive data or the negative data, depending on its corresponding weight on the second layer. A careful analysis of the neurons' directional dynamics allows us to provide an $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ upper bound on the time it takes for all neurons to achieve good alignment with the input data, where n is the number of data points and μ measures how well the data are separated. After the early alignment phase, the loss converges to zero at a $\mathcal{O}(\frac{1}{t})$ rate, and the weight matrix on the first layer is approximately low-rank. Numerical experiments on the MNIST dataset illustrate our theoretical findings.
****
Minh Hoang,Carl Kingsford
Efficient Heterogeneous Meta-Learning via Channel Shuffling Modulation
We tackle the problem of meta-learning across heterogenous tasks. This problem seeks to extract and generalize transferable meta-knowledge through streaming task sets from a multi-modal task distribution. The extracted meta-knowledge can be used to create predictors for new tasks using a small number of labeled samples. Most meta-learning methods assume a homogeneous task distribution, thus limiting their generalization capacity when handling multi-modal task distributions. Recent work has shown that the generalization of meta-learning depends on the similarity of tasks in the training distribution, and this has led to many clustering approaches that aim to detect homogeneous clusters of tasks. However, these methods suffer from a significant increase in parameter complexity. To overcome this weakness, we propose a new heterogeneous meta-learning strategy that efficiently captures the multi-modality of the task distribution via modulating the routing between convolution channels in the network, instead of directly modulating the network weights. This new mechanism can be cast as a permutation learning problem. We further introduce a novel neural permutation layer based on the classical Benes routing network, which has sub-quadratic parameter complexity in the total number of channels, as compared to the quadratic complexity of the state-of-the-art Gumbel-Sinkhorn layer. We demonstrate our approach on various multi-modal meta-learning benchmarks, showing that our framework outperforms previous methods in both generalization accuracy and convergence speed.
****
Zihao Wang,Eshaan Nichani,Jason D. Lee
Learning Hierarchical Polynomials with Three-Layer Neural Networks
We study the problem of learning hierarchical polynomials over the standard Gaussian distribution with three-layer neural networks. We specifically consider target functions of the form $h = g \circ p$ where $p : \mathbb{R}^d \rightarrow \mathbb{R}$ is a degree k polynomial and $g: \mathbb{R} \rightarrow \mathbb{R}$ is a degree q polynomial. This function class generalizes the single-index model, which corresponds to $k=1$, and is a natural class of functions possessing an underlying hierarchical structure. Our main result shows that for a large subclass of degree k polynomials p, a three-layer neural network trained via layerwise gradient descent on the square loss learns the target h up to vanishing test error in $\widetilde O(d^k)$ samples and polynomial time. This is a strict

improvement over kernel methods, which require $\widetilde \Theta(d^{kq})$ samples, as well as existing guarantees for two-layer networks, which require the target function to be low-rank. Our result also generalizes prior works on three-layer neural networks, which were restricted to the case of p being a quadratic. When p is indeed a quadratic, we achieve the information-theoretically optimal sample complexity $\widetilde O(d^2)$, which is an improvement over prior work (Nichani et al., 2023) requiring a sample size of $\widetilde\Theta(d^4)$. Our proof proceeds by showing that during the initial stage of training the network performs feature learning to recover the feature p with $\widetilde O(d^k)$ samples. This work demonstrates the ability of three-layer neural networks to learn complex features and as a result, learn a broad class of hierarchical functions.
****
Juncheng Liu,Bryan Hooi,Kenji Kawaguchi,Yiwei Wang,Chaosheng Dong,Xiaokui Xiao
Scalable and Effective Implicit Graph Neural Networks on Large Graphs
Graph Neural Networks (GNNs) have become the de facto standard for modeling graph-structured data in various applications. Among them, implicit GNNs have shown a superior ability to effectively capture long-range dependencies in underlying graphs. However, implicit GNNs tend to be computationally expensive and have high memory usage, due to 1) their use of full-batch training; and 2) they require a large number of iterations to solve a fixed-point equation. These compromise the scalability and efficiency of implicit GNNs especially on large graphs. In this paper, we aim to answer the question: how can we efficiently train implicit GNNs to provide effective predictions on large graphs? We propose a new scalable and effective implicit GNN (SEIGNN) with a mini-batch training method and a stochastic solver, which can be trained efficiently on large graphs. Specifically, SEIGNN can more effectively incorporate global and long-range information by introducing coarse-level nodes in the mini-batch training method. It also achieves reduced training time by obtaining unbiased approximate solutions with fewer iterations in the proposed solver. Comprehensive experiments on various large graphs demonstrate that SEIGNN outperforms baselines and achieves higher accuracy with less training time compared with existing implicit GNNs.
****
Chenxiang Ma,Jibin Wu,Chenyang Si,KC Tan
Scaling Supervised Local Learning with Augmented Auxiliary Networks
Deep neural networks are typically trained using global error signals that backpropagate (BP) end-to-end, which is not only biologically implausible but also suffers from the update locking problem and requires huge memory consumption. Local learning, which updates each layer independently with a gradient-isolated auxiliary network, offers a promising alternative to address the above problems. However, existing local learning methods are confronted with a large accuracy gap with the BP counterpart, particularly for large-scale networks. This is due to the weak coupling between local layers and their subsequent network layers, as there is no gradient communication across layers. To tackle this issue, we put forward an augmented local learning method, dubbed AugLocal. AugLocal constructs each hidden layer’s auxiliary network by uniformly selecting a small subset of layers from its subsequent network layers to enhance their synergy. We also propose to linearly reduce the depth of auxiliary networks as the hidden layer goes deeper, ensuring sufficient network capacity while reducing the computational cost of auxiliary networks. Our extensive experiments on four image classification datasets (i.e., CIFAR-10, SVHN, STL-10, and ImageNet) demonstrate that AugLocal can effectively scale up to tens of local layers with a comparable accuracy to BP-trained networks while reducing GPU memory usage by around 40%. The proposed AugLocal method, therefore, opens up a myriad of opportunities for training high-performance deep neural networks on resource-constrained platforms. Code is available at \url{https://github.com/ChenxiangMA/AugLocal}.
****
Kezhi Kong,Jiani Zhang,Zhengyuan Shen,Balasubramaniam Srinivasan,Chuan Lei,Christos Faloutsos,Huzefa Rangwala,George Karypis
OpenTab: Advancing Large Language Models as Open-domain Table Reasoners

Large Language Models (LLMs) trained on large volumes of data excel at various natural language tasks, but they cannot handle tasks requiring knowledge that has not been trained on previously. One solution is to use a retriever that fetches relevant information to expand LLM's knowledge scope. However, existing textual-oriented retrieval-based LLMs are not ideal on structured table data due to diversified data modalities and large table sizes. In this work, we propose OpenTab, an open-domain table reasoning framework powered by LLMs. Overall, OpenTab leverages table retriever to fetch relevant tables and then generates SQL programs to parse the retrieved tables efficiently. Utilizing the intermediate data derived from the SQL executions, it conducts grounded inference to produce accurate response. Extensive experimental evaluation shows that OpenTab significantly outperforms baselines in both open- and closed-domain settings, achieving up to 21.5% higher accuracy. We further run ablation studies to validate the efficacy of our proposed designs of the system.
****
Wentao Wu,Aleksei Timofeev,Chen Chen,Bowen Zhang,Kun Duan,Shuangning Liu,Yantao Zheng,Jonathon Shlens,Xianzhi Du,Yinfei Yang
MOFI: Learning Image Representations from Noisy Entity Annotated Images
We present MOFI, Manifold OF Images, a new vision foundation model designed to learn image representations from noisy entity annotated images. MOFI differs from previous work in two key aspects: 1. pre-training data, and 2. training recipe. Regarding data, we introduce a new approach to automatically assign entity labels to images from noisy image-text pairs. Our approach involves employing a named entity recognition model to extract entities from the alt-text, and then using a CLIP model to select the correct entities as labels of the paired image. It's a simple, cost-effective method that can scale to handle billions of web-mined image-text pairs. Through this method, we have created Image-to-Entities (I2E), a new dataset with 1 billion images and 2 million distinct entities, covering rich visual concepts in the wild. Building upon the I2E dataset, we study different training recipes like supervised pre-training, contrastive pre-training, and multi-task learning. For constrastive pre-training, we treat entity names as free-form text, and further enrich them with entity descriptions. Experiments show that supervised pre-training with large-scale fine-grained entity labels is highly effective for image retrieval tasks, and multi-task training further improves the performance. The final MOFI model achieves 86.66\% mAP on the challenging GPR1200 dataset, surpassing the previous state-of-the-art performance of 72.19% from OpenAI's CLIP model. Further experiments on zero-shot and linear probe image classification also show that MOFI outperforms a CLIP model trained on the original image-text data, demonstrating the effectiveness of the I2E dataset in learning strong image representations. We release our code and model weights at https://github.com/apple/ml-mofi.
****
Kaizhi Yang,Xiaoshuai Zhang,Zhiao Huang,Xuejin Chen,Zexiang Xu,Hao Su
MovingParts: Motion-based 3D Part Discovery in Dynamic Radiance Field
We present MovingParts, a NeRF-based method for dynamic scene reconstruction and part discovery. We consider motion as an important cue for identifying parts, that all particles on the same part share the common motion pattern. From the perspective of fluid simulation, existing deformation-based methods for dynamic NeRF can be seen as parameterizing the scene motion under the Eulerian view, i.e., focusing on specific locations in space through which the fluid flows as time passes. However, it is intractable to extract the motion of constituting objects or parts using the Eulerian view representation. In this work, we introduce the dual Lagrangian view and enforce representations under the Eulerian/Lagrangian views to be cycle-consistent. Under the Lagrangian view, we parameterize the scene motion by tracking the trajectory of particles on objects. The Lagrangian view makes it convenient to discover parts by factorizing the scene motion as a composition of part-level rigid motions. Experimentally, our method can achieve fast and high-quality dynamic scene reconstruction from even a single moving camera, and the induced part-based representation allows direct applications of part tracking, animation, 3D scene editing, etc.

****
Dehao Yuan,Furong Huang,Cornelia Fermuller,Yiannis Aloimonos
Decodable and Sample Invariant Continuous Object Encoder
We propose Hyper-Dimensional Function Encoding (HDFE). Given samples of a continuous object (e.g. a function), HDFE produces an explicit vector representation of the given object, invariant to the sample distribution and density. Sample distribution and density invariance enables HDFE to consistently encode continuous objects regardless of their sampling, and therefore allows neural networks to receive continuous objects as inputs for machine learning tasks, such as classification and regression. Besides, HDFE does not require any training and is proved to map the object into an organized embedding space, which facilitates the training of the downstream tasks. In addition, the encoding is decodable, which enables neural networks to regress continuous objects
by regressing their encodings. Therefore, HDFE serves as an interface for processing continuous objects.

We apply HDFE to function-to-function mapping, where vanilla HDFE achieves competitive performance with the state-of-the-art algorithm. We apply HDFE to point cloud surface normal estimation, where a simple replacement from PointNet to HDFE leads to 12\% and 15\% error reductions in two benchmarks.
In addition, by integrating HDFE into the PointNet-based SOTA network, we improve the SOTA baseline by 2.5\% and 1.7\% on the same benchmarks.
****
Renbo Tu,Colin White,Jean Kossaifi,Boris Bonev,Gennady Pekhimenko,Kamyar Azizzadenesheli,Anima Anandkumar
Guaranteed Approximation Bounds for Mixed-Precision Neural Operators
Neural operators, such as Fourier Neural Operators (FNO), form a principled approach for learning solution operators for partial differential equations (PDE) and other mappings between function spaces. However, many real-world problems require high-resolution training data, and the training time and limited GPU memory pose big barriers. One solution is to train neural operators in mixed precision to reduce the memory requirement and increase training speed. However, existing mixed-precision training techniques are designed for standard neural networks, and we find that their direct application to FNO leads to numerical overflow and poor memory efficiency. Further, at first glance, it may appear that mixed precision in FNO will lead to drastic accuracy degradation since reducing the precision of the Fourier transform yields poor results in classical numerical solvers. We show that this is not the case; in fact, we prove that reducing the precision in FNO still guarantees a good approximation bound, when done in a targeted manner. Specifically, we build on the intuition that neural operator learning inherently induces an approximation error, arising from discretizing the infinite-dimensional ground-truth input function, implying that training in full precision is not needed. We formalize this intuition by rigorously characterizing the approximation and precision errors of FNO and bounding these errors for general input functions. We prove that the precision error is asymptotically comparable to the approximation error. Based on this, we design a simple method to optimize the memory-intensive half-precision tensor contractions by greedily finding the optimal contraction order. Through extensive experiments on different state-of-the-art neural operators, datasets, and GPUs, we demonstrate that our approach reduces GPU memory usage by up to 50% and improves throughput by 58% with little or no reduction in accuracy.
****
Zhanke Zhou,Yongqi Zhang,Jiangchao Yao,quanming yao,Bo Han
Less is More: One-shot Subgraph Reasoning on Large-scale Knowledge Graphs
To deduce new facts on a knowledge graph (KG), a link predictor learns from the graph structure and collects local evidence to find the answer to a given query. However, existing methods suffer from a severe scalability problem due to the utilization of the whole KG for prediction, which hinders their promise on large scale KGs and cannot be directly addressed by vanilla sampling methods. In this work, we propose the one-shot-subgraph link prediction to achieve efficient and

adaptive prediction. The design principle is that, instead of directly acting on the whole KG, the prediction procedure is decoupled into two steps, i.e., (i) extracting only one subgraph according to the query and (ii) predicting on this single, query dependent subgraph. We reveal that the non-parametric and computation-efficient heuristics Personalized PageRank (PPR) can effectively identify the potential answers and supporting evidence. With efficient subgraph-based prediction, we further introduce the automated searching of the optimal configurations in both data and model spaces. Empirically, we achieve promoted efficiency and leading performances on five large-scale benchmarks. The code is publicly available at: https://github.com/tmlr-group/one-shot-subgraph.
****
Anna Bair,Hongxu Yin,Maying Shen,Pavlo Molchanov,Jose M. Alvarez
Adaptive Sharpness-Aware Pruning for Robust Sparse Networks
Robustness and compactness are two essential attributes of deep learning models that are deployed in the real world.
The goals of robustness and compactness may seem to be at odds, since robustness requires generalization across domains, while the process of compression exploits specificity in one domain.
We introduce \textit{Adaptive Sharpness-Aware Pruning (AdaSAP)}, which unifies these goals through the lens of network sharpness.
The AdaSAP method produces sparse networks that are robust to input variations which are \textit{unseen at training time}.
We achieve this by strategically incorporating weight perturbations in order to optimize the loss landscape. This allows the model to be both primed for pruning and regularized for improved robustness.
AdaSAP improves the robust accuracy of pruned models on image classification by up to +6\% on ImageNet C and +4\% on ImageNet V2, and on object detection by +4\% on a corrupted Pascal VOC dataset, over a wide range of compression ratios, pruning criteria, and network architectures, outperforming recent pruning art by large margins.
****
Sarthak Yadav,Sergios Theodoridis,Lars Kai Hansen,Zheng-Hua Tan
Masked Autoencoders with Multi-Window Local-Global Attention Are Better Audio Learners
In this work, we propose a Multi-Window Masked Autoencoder (MW-MAE) fitted with a novel Multi-Window Multi-Head Attention (MW-MHA) module that facilitates the modelling of local-global interactions in every decoder transformer block through attention heads of several distinct local and global windows. Empirical results on ten downstream audio tasks show that MW-MAEs consistently outperform standard MAEs in overall performance and learn better general-purpose audio representations, along with demonstrating considerably better scaling characteristics. Investigating attention distances and entropies reveals that MW-MAE encoders learn heads with broader local and global attention. Analyzing attention head feature representations through Projection Weighted Canonical Correlation Analysis (PWCCA) shows that attention heads with the same window sizes across the decoder layers of the MW-MAE learn correlated feature representations which enables each block to independently capture local and global information, leading to a decoupled decoder feature hierarchy.
****
Haoze Wu,Clark Barrett,Nina Narodytska
Lemur: Integrating Large Language Models in Automated Program Verification
The demonstrated code-understanding capability of LLMs raises the question of whether they can be used for automated program verification, a task that demands high-level abstract reasoning about program properties that is challenging for verification tools. We propose a general methodology to combine the power of LLMs and automated reasoners for automated program verification. We formally describe this methodology as a set of derivation rules and prove its soundness. We instantiate the calculus as a sound automated verification procedure, which led to practical improvements on a set of synthetic and competition benchmarks.
****

Yuxue Yang,Lue Fan,Zhaoxiang Zhang
MixSup: Mixed-grained Supervision for Label-efficient LiDAR-based 3D Object Detection
Label-efficient LiDAR-based 3D object detection is currently dominated by weakly/semi-supervised methods. Instead of exclusively following one of them, we propose MixSup, a more practical paradigm simultaneously utilizing massive cheap coarse labels and a limited number of accurate labels for Mixed-grained Supervision. We start by observing that point clouds are usually textureless, making it hard to learn semantics. However, point clouds are geometrically rich and scale-invariant to the distances from sensors, making it relatively easy to learn the geometry of objects, such as poses and shapes. Thus, MixSup leverages massive coarse cluster-level labels to learn semantics and a few expensive box-level labels to learn accurate poses and shapes. We redesign the label assignment in mainstream detectors, which allows them seamlessly integrated into MixSup, enabling practicality and universality. We validate its effectiveness in nuScenes, Waymo Open Dataset, and KITTI, employing various detectors. MixSup achieves up to 97.31% of fully supervised performance, using cheap cluster annotations and only 10% box annotations. Furthermore, we propose PointSAM based on the Segment Anything Model for automated coarse labeling, further reducing the annotation burden. The code is available at https://github.com/BraveGroup/PointSAM-for-MixSup.
****
Tim Dettmers,Ruslan A. Svirschevski,Vage Egiazarian,Denis Kuznedelev,Elias Frantar,Saleh Ashkboos,Alexander Borzunov,Torsten Hoefler,Dan Alistarh
SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression
Recent advances in large language model (LLM) pretraining have led to high-quality LLMs with impressive abilities. By compressing such LLMs via quantization to 3-4 bits per parameter, they can fit into memory-limited devices such as laptops and mobile phones, enabling personalized use. Quantizing models to 3-4 bits per parameter can lead to moderate to high accuracy losses, especially for smaller models (1-10B parameters), which are suitable for edge deployment. To address this accuracy issue, we introduce the Sparse-Quantized Representation (SpQR), a new compressed format and quantization technique that enables for the first time \emph{near-lossless} compression of LLMs across model scales while reaching similar compression levels to previous methods. SpQR works by identifying and isolating \emph{outlier weights}, which cause particularly large quantization errors, and storing them in higher precision while compressing all other weights to 3-4 bits, and achieves relative accuracy losses of less than 1% in perplexity for highly-accurate LLaMA and Falcon LLMs. This makes it possible to run a 33B parameter LLM on a single 24 GB consumer GPU without performance degradation at 15\% speedup, thus making powerful LLMs available to consumers without any downsides. SpQR comes with efficient algorithms for both encoding weights into its format, as well as decoding them efficiently at runtime. Specifically, we provide an efficient GPU inference algorithm for SpQR, which yields faster inference than 16-bit baselines at similar accuracy while enabling memory compression gains of more than 4x.
****
Mingjie Sun,Zhuang Liu,Anna Bair,J Zico Kolter
A Simple and Effective Pruning Approach for Large Language Models
As their size increases, Large Languages Models (LLMs) are natural candidates for network pruning methods: approaches that drop a subset of network weights while striving to preserve performance. Existing methods, however, require either retraining, which is rarely affordable for billion-scale LLMs, or solving a weight reconstruction problem reliant on second-order information, which may also be computationally expensive. In this paper, we introduce a novel, straightforward yet effective pruning method, termed Wanda (Pruning by Weights and activations), designed to induce sparsity in pretrained LLMs. Motivated by the recent observation of emergent large magnitude features in LLMs, our approach prunes weights with the smallest magnitudes multiplied by the corresponding input activations, on a per-output basis. Notably, Wanda requires no retraining or weight update, and the pruned LLM can be used as is. We conduct a thorough evaluation of our metho

d Wanda on LLaMA and LLaMA-2 across various language benchmarks. Wanda significantly outperforms the established baseline of magnitude pruning and performs competitively against recent method involving intensive weight update.
****
Druv Pai,Sam Buchanan,Ziyang Wu,Yaodong Yu,Yi Ma
Masked Completion via Structured Diffusion with White-Box Transformers
Modern learning frameworks often train deep neural networks with massive amounts of unlabeled data to learn representations by solving simple pretext tasks, then use the representations as foundations for downstream tasks. These networks are empirically designed; as such, they are usually not interpretable, their representations are not structured, and their designs are potentially redundant. White-box deep networks, in which each layer explicitly identifies and transforms structures in the data, present a promising alternative. However, existing white-box architectures have only been shown to work at scale in supervised settings with labeled data, such as classification. In this work, we provide the first instantiation of the white-box design paradigm that can be applied to large-scale unsupervised representation learning. We do this by exploiting a fundamental connection between diffusion, compression, and (masked) completion, deriving a deep transformer-like masked autoencoder architecture, called CRATE-MAE, in which the role of each layer is mathematically fully interpretable: they transform the data distribution to and from a structured representation. Extensive empirical evaluations confirm our analytical insights. CRATE-MAE demonstrates highly promising performance on large-scale imagery datasets while using only ~30% of the parameters compared to the standard masked autoencoder with the same model configuration. The representations learned by CRATE-MAE have explicit structure and also contain semantic meaning.
****
Libin Zhu,Chaoyue Liu,Adityanarayanan Radhakrishnan,Mikhail Belkin
Quadratic models for understanding catapult dynamics of neural networks
While neural networks can be approximated by linear models as their width increases, certain properties of wide neural networks cannot be captured by linear models. In this work we show that recently proposed Neural Quadratic Models can exhibit the "catapult phase" Lewkowycz et al. (2020) that arises when training such models with large learning rates. We then empirically show that the behaviour of quadratic models parallels that of neural networks in generalization, especially in the catapult phase regime. Our analysis further demonstrates that quadratic models are an effective tool for analysis of neural networks.
****
Marc Rußwurm,Konstantin Klemmer,Esther Rolf,Robin Zbinden,Devis Tuia
Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks
Learning representations of geographical space is vital for any machine learning model that integrates geolocated data, spanning application domains such as remote sensing, ecology, or epidemiology. Recent work embeds coordinates using sine and cosine projections based on Double Fourier Sphere (DFS) features. These embeddings assume a rectangular data domain even on global data, which can lead to artifacts, especially at the poles. At the same time, little attention has been paid to the exact design of the neural network architectures with which these functional embeddings are combined. This work proposes a novel location encoder for globally distributed geographic data that combines spherical harmonic basis functions, natively defined on spherical surfaces, with sinusoidal representation networks (SirenNets) that can be interpreted as learned Double Fourier Sphere embedding. We systematically evaluate positional embeddings and neural network architectures across various benchmarks and synthetic evaluation datasets. In contrast to previous approaches that require the combination of both positional encoding and neural networks to learn meaningful representations, we show that both spherical harmonics and sinusoidal representation networks are competitive on their own but set state-of-the-art performances across tasks when combined. The model code and experiments are available at https://github.com/marccoru/locationencoder.

****
Lunjun Zhang,Yuwen Xiong,Ze Yang,Sergio Casas,Rui Hu,Raquel Urtasun
Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion
Learning world models can teach an agent how the world works in an unsupervised manner. Even though it can be viewed as a special case of sequence modeling, progress for scaling world models on robotic applications such as autonomous driving has been somewhat less rapid than scaling language models with Generative Pre-trained Transformers (GPT). We identify two reasons as major bottlenecks: dealing with complex and unstructured observation space, and having a scalable generative model. Consequently, we propose a novel world modeling approach that first tokenizes sensor observations with VQVAE, then predicts the future via discrete diffusion. To efficiently decode and denoise tokens in parallel, we recast Masked Generative Image Transformer into the discrete diffusion framework with a few simple changes, resulting in notable improvement. When applied to learning world models on point cloud observations, our model reduces prior SOTA Chamfer distance by more than 65% for 1s prediction, and more than 50% for 3s prediction, across NuScenes, KITTI Odometry, and Argoverse2 datasets. Our results demonstrate that discrete diffusion on tokenized agent experience can unlock the power of GPT-like unsupervised learning for robotic agents.
****
Bo Zhou,Ruiwei Jiang,Siqian Shen
Learning to Solve Bilevel Programs with Binary Tender
Bilevel programs (BPs) find a wide range of applications in fields such as energy, transportation, and machine learning. As compared to BPs with continuous (linear/convex) optimization problems in both levels, the BPs with discrete decision variables have received much less attention, largely due to the ensuing computational intractability and the incapability of gradient-based algorithms for handling discrete optimization formulations. In this paper, we develop deep learning techniques to address this challenge. Specifically, we consider a BP with binary tender, wherein the upper and lower levels are linked via binary variables. We train a neural network to approximate the optimal value of the lower-level problem, as a function of the binary tender. Then, we obtain a single-level reformulation of the BP through a mixed-integer representation of the value function. Furthermore, we conduct a comparative analysis between two types of neural networks: general neural networks and the novel input supermodular neural networks, studying their representational capacities. To solve high-dimensional BPs, we introduce an enhanced sampling method to generate higher-quality samples and implement an iterative process to refine solutions. We demonstrate the performance of these approaches through extensive numerical experiments, whose lower-level problems are linear and mixed-integer programs, respectively.
****
Jong-Hoon Ahn,Akshay Vashist
A Linear Algebraic Framework for Counterfactual Generation
Estimating individual treatment effects in clinical data is essential for understanding how different patients uniquely respond to treatments and identifying the most effective interventions for specific patient subgroups, thereby enhancing the precision and personalization of healthcare. However, counterfactual data are not accessible, and the true calculation of causal effects cannot be performed at the individual level. This paper proposes a linear algebraic framework to generate counterfactual longitudinal data that exactly matches pre-treatment factual data. Because causation travels forward in time, not in reverse, counterfactual predictability is further strengthened by blocking causal effects from flowing back to the past, thus limiting counterfactual dependence on the future. Using simulated LDL cholesterol datasets, we show that our method significantly outperforms the most cited methods of counterfactual generation. We also provide a formula that can estimate the time-varying variance of individual treatment effects, interpreted as a confidence level in the generated counterfactuals compared to true values.
****
Ce Ju,Reinmar J Kobler,Liyao Tang,Cuntai Guan,Motoaki Kawanabe

Deep Geodesic Canonical Correlation Analysis for Covariance-Based Neuroimaging Data
In human neuroimaging, multi-modal imaging techniques are frequently combined to enhance our comprehension of whole-brain dynamics and improve diagnosis in clinical practice. Modalities like electroencephalography and functional magnetic resonance imaging provide distinct views to the brain dynamics due to diametral spatiotemporal sensitivities and underlying neurophysiological coupling mechanisms. These distinct views pose a considerable challenge to learning a shared representation space, especially when dealing with covariance-based data characterized by their geometric structure. To capitalize on the geometric structure, we introduce a measure called geodesic correlation which expands traditional correlation consistency to covariance-based data on the symmetric positive definite (SPD) manifold. This measure is derived from classical canonical correlation analysis and serves to evaluate the consistency of latent representations obtained from paired views. For multi-view, self-supervised learning where one or both latent views are SPD we propose an innovative geometric deep learning framework termed DeepGeoCCA. Its primary objective is to enhance the geodesic correlation of unlabeled, paired data, thereby generating novel representations while retaining the geometric structures. In simulations and experiments with multi-view and multi-modal human neuroimaging data, we find that DeepGeoCCA learns latent representations with high geodesic correlation for unseen data while retaining relevant information for downstream tasks.
****
Hongyi Wang,Felipe Maia Polo,Yuekai Sun,Souvik Kundu,Eric Xing,Mikhail Yurochkin
Fusing Models with Complementary Expertise
Training AI models that generalize across tasks and domains has long been among the open problems driving AI research. The emergence of Foundation Models made it easier to obtain expert models for a given task, but the heterogeneity of data that may be encountered at test time often means that any single expert is insufficient. We consider the Fusion of Experts (FoE) problem of fusing outputs of expert models with complementary knowledge of the data distribution and formulate it as an instance of supervised learning. Our method is applicable to both discriminative and generative tasks and leads to significant performance improvements in image and text classification, text summarization, multiple-choice QA, and automatic evaluation of generated text. We also extend our method to the "frugal" setting where it is desired to reduce the number of expert model evaluations at test time. Our implementation is publicly available at https://github.com/hwang595/FoE-ICLR2024.
****
Nima Shoghi,Adeesh Kolluru,John R. Kitchin,Zachary Ward Ulissi,C. Lawrence Zitnick,Brandon M Wood
From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction
Foundation models have been transformational in machine learning fields such as natural language processing and computer vision. Similar success in atomic property prediction has been limited due to the challenges of training effective models across multiple chemical domains. To address this, we introduce Joint Multi-domain Pre-training (JMP), a supervised pre-training strategy that simultaneously trains on multiple datasets from different chemical domains, treating each dataset as a unique pre-training task within a multi-task framework. Our combined training dataset consists of $\sim$120M systems from OC20, OC22, ANI-1x, and Transition-1x. We evaluate performance and generalization by fine-tuning over a diverse set of downstream tasks and datasets including: QM9, rMD17, MatBench, QMOF, SPICE, and MD22. JMP demonstrates an average improvement of 59% over training from scratch and matches or sets state-of-the-art on 34 out of 40 tasks. Our work highlights the potential of pre-training strategies that utilize diverse data to advance property prediction across chemical domains, especially for low-data tasks.
****
Runzhe Wu,Wen Sun

Making RL with Preference-based Feedback Efficient via Randomization
Reinforcement Learning algorithms that learn from human feedback (RLHF) need to
be efficient in terms of *statistical complexity, computational complexity, and
query complexity*. In this work, we consider the RLHF setting where the feedback
is given in the format of preferences over pairs of trajectories. In the linear
MDP model, using randomization in algorithm design, we present an algorithm tha
t is sample efficient (i.e., has near-optimal worst-case regret bounds) and has
polynomial running time (i.e., computational complexity is polynomial with respe
ct to relevant parameters). Our algorithm further minimizes the query complexity
through a novel randomized active learning procedure. In particular, our algori
thm demonstrates a near-optimal tradeoff between the regret bound and the query
complexity. To extend the results to more general nonlinear function approximati
on, we design a model-based randomized algorithm inspired by the idea of Thompso
n sampling. Our algorithm minimizes Bayesian regret bound and query complexity,
again achieving a near-optimal tradeoff between these two quantities. Computatio
n-wise, similar to the prior Thompson sampling algorithms under the regular RL s
etting, the main computation primitives of our algorithm are Bayesian supervised
learning oracles which have been heavily investigated on the empirical side whe
n applying Thompson sampling algorithms to RL benchmark problems.
****
Ido Amos,Jonathan Berant,Ankit Gupta
Never Train from Scratch: Fair Comparison of Long-Sequence Models Requires Data-
Driven Priors
Modeling long-range dependencies across sequences is a longstanding goal in mach
ine learning and has led to architectures, such as state space models, that dram
atically outperform Transformers on long sequences. However, these impressive em
pirical gains have been by and large demonstrated on benchmarks (e.g. Long Range
Arena), where models are randomly initialized and trained to predict a target l
abel from an input sequence. In this work, we show that random initialization le
ads to gross overestimation of the differences between architectures and that pr
etraining with standard denoising objectives, *using only the downstream task da
ta*, leads to dramatic gains across multiple architectures and to very small gap
s between Transformers and state space models (SSMs). In stark contrast to prior
works, we find vanilla Transformers to match the performance of S4 on Long Rang
e Arena when properly pretrained, and we improve the best reported results of SS
Ms on the PathX-256 task by 20 absolute points. Subsequently, we analyze the uti
lity of previously-proposed structured parameterizations for SSMs and show they
become mostly redundant in the presence of data-driven initialization obtained t
hrough pretraining. Our work shows that, when evaluating different architectures
on supervised tasks, incorporation of data-driven priors via pretraining is ess
ential for reliable performance estimation, and can be done efficiently.
****
Aditya Bhatt,Daniel Palenicek,Boris Belousov,Max Argus,Artemij Amiranashvili,Tho
mas Brox,Jan Peters
CrossQ: Batch Normalization in Deep Reinforcement Learning for Greater Sample Ef
ficiency and Simplicity
Sample efficiency is a crucial problem in deep reinforcement learning. Recent al
gorithms, such as REDQ and DroQ, found a way to improve the sample efficiency by
increasing the update-to-data (UTD) ratio to 20 gradient update steps on the cr
itic per environment sample.
However, this comes at the expense of a greatly increased computational cost. To
reduce this computational burden, we introduce CrossQ:
A lightweight algorithm for continuous control tasks that makes careful use of B
atch Normalization and removes target networks to surpass the current state-of-t
he-art in sample efficiency while maintaining a low UTD ratio of 1. Notably, Cro
ssQ does not rely on advanced bias-reduction schemes used in current methods. Cr
ossQ's contributions are threefold: (1) it matches or surpasses current state-of
-the-art methods in terms of sample efficiency, (2) it substantially reduces the
computational cost compared to REDQ and DroQ, (3) it is easy to implement, requ
iring just a few lines of code on top of SAC.

****
Zecheng Wang,Che Wang,Zixuan Dong,Keith W. Ross
Pre-training with Synthetic Data Helps Offline Reinforcement Learning
Recently, it has been shown that for offline deep reinforcement learning (DRL),
pre-training Decision Transformer with a large language corpus can improve downs
tream performance (Reid et al., 2022). A natural question to ask is whether this
 performance gain can only be achieved with language pre-training, or can be ach
ieved with simpler pre-training schemes which do not involve language. In this p
aper, we first show that language is not essential for improved performance, and
 indeed pre-training with synthetic IID data for a small number of updates can m
atch the performance gains from pre-training with a large language corpus; moreo
ver, pre-training with data generated by a one-step Markov chain can further imp
rove the performance. Inspired by these experimental results, we then consider p
re-training Conservative Q-Learning (CQL), a popular offline DRL algorithm, whic
h is Q-learning-based and typically employs a Multi-Layer Perceptron (MLP) backb
one. Surprisingly, pre-training with simple synthetic data for a small number of
 updates can also improve CQL, providing consistent performance improvement on D
4RL Gym locomotion datasets. The results of this paper not only illustrate the i
mportance of pre-training for offline DRL but also show that the pre-training da
ta can be synthetic and generated with remarkably simple mechanisms.
****
Longtao Zheng,Rundong Wang,Xinrun Wang,Bo An
Synapse: Trajectory-as-Exemplar Prompting with Memory for Computer Control
Building agents with large language models (LLMs) for computer control is a burg
eoning research area, where the agent receives computer states and performs acti
ons to complete complex tasks. Previous computer agents have demonstrated the be
nefits of in-context learning (ICL); however, their performance is hindered by s
everal issues. First, the limited context length of LLMs and complex computer st
ates restrict the number of exemplars, as a single webpage can consume the entir
e context. Second, the exemplars in current methods, such as high-level plans an
d multi-choice questions, cannot represent complete trajectories, leading to sub
optimal performance in long-horizon tasks. Third, existing computer agents rely
on task-specific exemplars and overlook the similarity among tasks, resulting in
 poor generalization to novel tasks. To address these challenges, we introduce S
ynapse, a computer agent featuring three key components: i) state abstraction, w
hich filters out task-irrelevant information from raw states, allowing more exem
plars within the limited context, ii) trajectory-as-exemplar prompting, which pr
ompts the LLM with complete trajectories of the abstracted states and actions to
 improve multi-step decision-making, and iii) exemplar memory, which stores the
embeddings of exemplars and retrieves them via similarity search for generalizat
ion to novel tasks. We evaluate Synapse on MiniWoB++, a standard task suite, and
 Mind2Web, a real-world website benchmark. In MiniWoB++, Synapse achieves a 99.2
% average success rate (a 10% relative improvement) across 64 tasks using demons
trations from only 48 tasks. Notably, Synapse is the first ICL method to solve t
he book-flight task in MiniWoB++. Synapse also exhibits a 56% relative improveme
nt in average step success rate over the previous state-of-the-art prompting sch
eme in Mind2Web.
****
Xinyan Chen,Yang Li,Runzhong Wang,Junchi Yan
MixSATGEN: Learning Graph Mixing for SAT Instance Generation
The Boolean satisfiability problem (SAT) stands as a canonical NP-complete task.
 In particular, the scarcity of real-world SAT instances and their usefulness fo
r tuning SAT solvers underscore the necessity for effective and efficient ways o
f hard instance generation, whereas existing methods either struggle to maintain
 plausible hardness or suffer from limited applicability. Different from the typ
ical construction-based methods, this paper introduces an adaptive and efficient
 graph interpolation approach that in place modifies the raw structure of graph-
represented SAT instance by replacing it with a counterpart from another instanc
e. Specifically, it involves a two-stage matching and mixing pipeline. The match
ing aims to find a correspondence map of literal nodes from two instance graphs

via learned features from a matching network; while the mixing stage involves it eratively exchanging clause pairs with the highest correspondence scores until a specified replacement ratio is achieved. We further show that under our matchin g-mixing framework, moderate randomness can avoid hardness degradation of instan ces by introducing Gumbel noise. Experimental results show the superiority of ou r method with both resemblance in structure and hardness, and general applicabil ity.
****
Shengcao Cao,Jiuxiang Gu,Jason Kuen,Hao Tan,Ruiyi Zhang,Handong Zhao,Ani Nenkova ,Liangyan Gui,Tong Sun,Yu-Xiong Wang
SOHES: Self-supervised Open-world Hierarchical Entity Segmentation
Open-world entity segmentation, as an emerging computer vision task, aims at seg menting entities in images without being restricted by pre-defined classes, offe ring impressive generalization capabilities on unseen images and concepts. Despi te its promise, existing entity segmentation methods like Segment Anything Model (SAM) rely heavily on costly expert annotators. This work presents Self-supervi sed Open-world Hierarchical Entity Segmentation (SOHES), a novel approach that e liminates the need for human annotations. SOHES operates in three phases: self-e xploration, self-instruction, and self-correction. Given a pre-trained self-supe rvised representation, we produce abundant high-quality pseudo-labels through vi sual feature clustering. Then, we train a segmentation model on the pseudo-label s, and rectify the noises in pseudo-labels via a teacher-student mutual-learning procedure. Beyond segmenting entities, SOHES also captures their constituent pa rts, providing a hierarchical understanding of visual entities. Using raw images as the sole training data, our method achieves unprecedented performance in sel f-supervised open-world segmentation, marking a significant milestone towards hi gh-quality open-world entity segmentation in the absence of human-annotated mask s. Project page: https://SOHES.github.io.
****
Robert Kirk,Ishita Mediratta,Christoforos Nalmpantis,Jelena Luketina,Eric Hambro ,Edward Grefenstette,Roberta Raileanu
Understanding the Effects of RLHF on LLM Generalisation and Diversity
Large language models (LLMs) fine-tuned with reinforcement learning from human f eedback (RLHF) have been used in some of the most widely deployed AI models to d ate, such as OpenAI's ChatGPT or Anthropic's Claude. While there has been signif icant work developing these methods, our understanding of the benefits and downs ides of each stage in RLHF is still limited. To fill this gap, we present an ext ensive analysis of how each stage of the process (i.e. supervised fine-tuning (S FT), reward modelling, and RLHF) affects two key properties: out-of-distribution (OOD) generalisation and output diversity. OOD generalisation is crucial given the wide range of real-world scenarios in which these models are being used, whi le output diversity refers to the model's ability to generate varied outputs and is important for a variety of use cases. We perform our analysis across two bas e models on both summarisation and instruction following tasks, the latter being highly relevant for current LLM use cases. We find that RLHF generalises better than SFT to new inputs, particularly as the distribution shift between train an d test becomes larger. However, RLHF significantly reduces output diversity comp ared to SFT across a variety of measures, implying a tradeoff in current LLM fin e-tuning methods between generalisation and diversity. Our results provide guida nce on which fine-tuning method should be used depending on the application, and show that more research is needed to improve the tradeoff between generalisatio n and diversity.
****
Jung-Chun Liu,Chi-Hsien Chang,Shao-Hua Sun,Tian-Li Yu
Integrating Planning and Deep Reinforcement Learning via Automatic Induction of Task Substructures
Despite the recent advancement in deep reinforcement learning (DRL), it still st ruggles at learning sparse-reward goal-directed tasks. On the other hand, classi cal planning approaches excel at addressing tasks with hierarchical structures b y employing symbolic knowledge for high-level planning. Yet, most classical plan

ning methods rely on assumptions about pre-defined subtasks, making them inapplicable in domains without domain knowledge or models. To bridge the best of both worlds, we propose a framework that integrates DRL with classical planning by automatically inducing task structures and substructures from a few demonstrations. Specifically, we use symbolic regression for substructure induction by adopting genetic programming where the program model reflects prior domain knowledge of effect rules. We compare our proposed framework to DRL algorithms, imitation learning methods, and an exploration approach
in various domains. The experimental results show that our framework outperforms the baselines in terms of sample efficiency and task performance. Moreover, our framework achieves strong generalization performance by inducing the new rules and composing the task structures. Ablation studies justify the design of the induction module and the proposed genetic programming procedure.
****
Jun Chen,Haishan Ye,Mengmeng Wang,Tianxin Huang,Guang Dai,Ivor Tsang,Yong Liu
Decentralized Riemannian Conjugate Gradient Method on the Stiefel Manifold
The conjugate gradient method is a crucial first-order optimization method that generally converges faster than the steepest descent method, and its computational cost is much lower than that of second-order methods. However, while various types of conjugate gradient methods have been studied in Euclidean spaces and on Riemannian manifolds, there is little study for those in distributed scenarios. This paper proposes a decentralized Riemannian conjugate gradient descent (DRCGD) method that aims at minimizing a global function over the Stiefel manifold. The optimization problem is distributed among a network of agents, where each agent is associated with a local function, and the communication between agents occurs over an undirected connected graph. Since the Stiefel manifold is a non-convex set, a global function is represented as a finite sum of possibly non-convex (but smooth) local functions. The proposed method is free from expensive Riemannian geometric operations such as retractions, exponential maps, and vector transports, thereby reducing the computational complexity required by each agent. To the best of our knowledge, DRCGD is the first decentralized Riemannian conjugate gradient algorithm to achieve global convergence over the Stiefel manifold.
****
Ali Shahin Shamsabadi,Gefei Tan,Tudor Ioan Cebere,Aurélien Bellet,Hamed Haddadi,Nicolas Papernot,Xiao Wang,Adrian Weller
Confidential-DPproof: Confidential Proof of Differentially Private Training
Post hoc privacy auditing techniques can be used to test the privacy guarantees of a model, but come with several limitations: (i) they can only establish lower bounds on the privacy loss, (ii) the intermediate model updates and some data must be shared with the auditor to get a better approximation of the privacy loss, and (iii) the auditor typically faces a steep computational cost to run a large number of attacks. In this paper, we propose to proactively generate a cryptographic certificate of privacy during training to forego such auditing limitations. We introduce Confidential-DPproof , a framework for Confidential Proof of Differentially Private Training, which enhances training with a certificate of the (ε,δ)-DP guarantee achieved. To obtain this certificate without revealing information about the training data or model, we design a customized zero-knowledge proof protocol tailored to the requirements introduced by differentially private training, including random noise addition and privacy amplification by subsampling. In experiments on CIFAR-10, Confidential-DPproof trains a model achieving state-of-the-art 91% test accuracy with a certified privacy guarantee of $(\varepsilon=0.55,\delta=10^{-5})$-DP in approximately 100 hours.
****
Francisco Vargas,Shreyas Padhy,Denis Blessing,Nikolas Nüsken
Transport meets Variational Inference: Controlled Monte Carlo Diffusions
Connecting optimal transport and variational inference, we present a principled and systematic framework for sampling and generative modelling centred around divergences on path space. Our work culminates in the development of the Controlled Monte Carlo Diffusion sampler (CMCD) for Bayesian computation, a score-based annealing technique that crucially adapts both forward and backward dynamics in a

diffusion model. On the way, we clarify the relationship between the EM-algorithm and iterative proportional fitting (IPF) for Schroedinger bridges, deriving as well a regularised objective that bypasses the iterative bottleneck of standard IPF-updates. Finally, we show that CMCD has a strong foundation in the Jarzinsky and Crooks identities from statistical physics, and that it convincingly outperforms competing approaches across a wide array of experiments.
****

Guangchi Fang,Qingyong Hu,Longguang Wang,Yulan Guo
ACRF: Compressing Explicit Neural Radiance Fields via Attribute Compression
In this work, we study the problem of explicit NeRF compression. Through analyzing recent explicit NeRF models, we reformulate the task of explicit NeRF compression as 3D data compression. We further introduce our NeRF compression framework, Attributed Compression of Radiance Field (ACRF), which focuses on the compression of the explicit neural 3D representation. The neural 3D structure is pruned and converted to points with features, which are further encoded using importance-guided feature encoding. Furthermore, we employ an importance-prioritized entropy model to estimate the probability distribution of transform coefficients, which are then entropy coded with an arithmetic coder using the predicted distribution. Within this framework, we present two models, ACRF and ACRF-F, to strike a balance between compression performance and encoding time budget. Our experiments, which include both synthetic and real-world datasets such as Synthetic-NeRF and Tanks&Temples, demonstrate the superior performance of our proposed algorithm.
****

Yunzhen Feng,Shanmukha Ramakrishna Vedantam,Julia Kempe
Embarrassingly Simple Dataset Distillation
Dataset distillation extracts a small set of synthetic training samples from a large dataset with the goal of achieving competitive performance on test data when trained on this sample. In this work, we tackle dataset distillation at its core by treating it directly as a bilevel optimization problem. Re-examining the foundational back-propagation through time method, we study the pronounced variance in the gradients, computational burden, and long-term dependencies. We introduce an improved method: Random Truncated Backpropagation Through Time (RaT-BPTT) to address them. RaT-BPTT incorporates a truncation coupled with a random window, effectively stabilizing the gradients and speeding up the optimization while covering long dependencies. This allows us to establish new state-of-the-art for a variety of standard dataset benchmarks. A deeper dive into the nature of distilled data unveils pronounced intercorrelation. In particular, subsets of distilled datasets tend to exhibit much worse performance than directly distilled smaller datasets of the same size. Leveraging RaT-BPTT, we devise a boosting mechanism that generates distilled datasets that contain subsets with near optimal performance across different data budgets.
****

Yuhang Zang,Hanlin Goh,Joshua M. Susskind,Chen Huang
Overcoming the Pitfalls of Vision-Language Model Finetuning for OOD Generalization
Existing vision-language models exhibit strong generalization on a variety of visual domains and tasks. However, such models mainly perform zero-shot recognition in a closed-set manner, and thus struggle to handle open-domain visual concepts by design. There are recent finetuning methods, such as prompt learning, that not only study the discrimination between in-distribution (ID) and out-of-distribution (OOD) samples, but also show some improvements in both ID and OOD accuracies. In this paper, we first demonstrate that vision-language models, after long enough finetuning but without proper regularization, tend to overfit the known classes in the given dataset, with degraded performance on unknown classes. Then we propose a novel approach OGEN to address this pitfall, with the main focus on improving the OOD GENeralization of finetuned models. Specifically, a class-conditional feature generator is introduced to synthesize OOD features using just the class name of any unknown class. Such synthesized features will provide useful knowledge about unknowns and help regularize the decision boundary between ID

and OOD data when optimized jointly. Equally important is our adaptive self-distillation mechanism to regularize our feature generation model during joint optimization, i.e., adaptively transferring knowledge between model states to further prevent overfitting. Experiments validate that our method yields convincing gains in OOD generalization performance in different settings.
****

Siqiao Xue,Xiaoming Shi,Zhixuan Chu,Yan Wang,Hongyan Hao,Fan Zhou,Caigao JIANG,Chen Pan,James Y. Zhang,Qingsong Wen,JUN ZHOU,Hongyuan Mei

EasyTPP: Towards Open Benchmarking Temporal Point Processes

Continuous-time event sequences play a vital role in real-world domains such as healthcare, finance, online shopping, social networks, and so on. To model such data, temporal point processes (TPPs) have emerged as the most natural and competitive models, making a significant impact in both academic and application communities. Despite the emergence of many powerful models in recent years, there hasn't been a central benchmark for these models and future research endeavors. This lack of standardization impedes researchers and practitioners from comparing methods and reproducing results, potentially slowing down progress in this field.

In this paper, we present EasyTPP, the first central repository of research assets (e.g., data, models, evaluation programs, documentations) in the area of event sequence modeling. Our EasyTPP makes several unique contributions to this area: a unified interface of using existing datasets and adding new datasets; a wide range of evaluation programs that are easy to use and extend as well as facilitate reproducible research; implementations of popular neural TPPs, together with a rich library of modules by composing which one could quickly build complex models. We will actively maintain this benchmark and welcome contributions from other researchers and practitioners.

Our benchmark will help promote reproducible research in this field, thus accelerating research progress as well as making more significant real-world impacts.

The code and data are available at \url{https://github.com/ant-research/EasyTemporalPointProcess}.
****

Yilong Xu,Yang Liu,Hao Sun

Reinforcement Symbolic Regression Machine

In nature, the behavior of many complex systems can be described by parsimonious math equations. Symbolic Regression (SR) is defined as the task of automatically distilling equations from limited data. Keen efforts have been placed on tackling this issue and demonstrated success in SR. However, there still exist bottlenecks that current methods struggle to break, when the expressions we need to explore tend toward infinity and especially when the underlying math formula is intricate. To this end, we propose a novel Reinforcement Symbolic Regression Machine (RSRM) that masters the capability of uncovering complex math equations from only scarce data. The RSRM model is composed of three key modules: (1) a Monte Carlo tree search (MCTS) agent, designed for exploration, that explores optimal math expression trees consisting of pre-defined math operators and variables, (2) a Double Q-learning block, designed for exploitation, that helps reduce the feasible search space of MCTS via properly understanding the distribution of reward, and (3) a modulated sub-tree discovery block that heuristically learns and defines new math operators to improve representation ability of math expression trees. Binding of these modules yields the SOTA performance of RSRM in SR as demonstrated by multiple benchmark datasets. The RSRM shows clear superiority over several representative baseline models.
****

James Chapman,Lennie Wells,Ana Lawry Aguila

Unconstrained Stochastic CCA: Unifying Multiview and Self-Supervised Learning

The Canonical Correlation Analysis (CCA) family of methods is foundational in multiview learning.

Regularised linear CCA methods can be seen to generalise Partial Least Squares (PLS) and be unified with a Generalized Eigenvalue Problem (GEP) framework.

However, classical algorithms for these linear methods are computationally infea

sible for large-scale data.
Extensions to Deep CCA show great promise, but current training procedures are s
low and complicated.
First we propose a novel unconstrained objective that characterizes the top subs
pace of GEPs.
Our core contribution is a family of fast algorithms for stochastic PLS, stochas
tic CCA, and Deep CCA, simply obtained by applying stochastic gradient descent (
SGD) to the corresponding CCA objectives.
Our algorithms show far faster convergence and recover higher correlations than
the previous state-of-the-art on all standard CCA and Deep CCA benchmarks.
These improvements allow us to perform a first-of-its-kind PLS analysis of an ex
tremely large biomedical dataset from the UK Biobank, with over 33,000 individua
ls and 500,000 features.
Finally, we apply our algorithms to match the performance of `CCA-family' Self-S
upervised Learning (SSL) methods on CIFAR-10 and CIFAR-100 with minimal hyper-pa
rameter tuning, and also present theory to clarify the links between these metho
ds and classical CCA, laying the groundwork for future insights.
****
Junyan Li,Delin Chen,Yining Hong,Zhenfang Chen,Peihao Chen,Yikang Shen,Chuang Ga
n
Compositional VLM: Composing Visual Entities and Relationships in Large Language
 Models Via Communicative Decoding
A remarkable ability of human beings resides in compositional reasoning, i.e., t
he capacity to make "infinite use of finite means". However, current large visio
n-language foundation models (VLMs) fall short of such compositional abilities
due to their ``bag-of-words" behaviors and inability to construct words that cor
rectly represent visual entities and the relations among the entities. To this e
nd, we propose Compositional VLM, which can guide the LLM to explicitly compose
visual entities and relationships among the text and dynamically communicate wit
h the vision encoder and detection network to achieve vision-language communicat
ive decoding. Specifically, we first devise a set of novel communication tokens
for the LLM, for dynamic communication between the visual detection system and t
he language system. A communication token is generated by the LLM following a vi
sual entity or a relation, to inform the detection network to propose regions th
at are relevant to the sentence generated so far. The proposed regions-of-intere
sts (ROIs) are then fed back into the LLM for better language generation conting
ent on the relevant regions. The LLM is thus able to compose the visual entities
 and relationships through the communication tokens. The vision-to-language and
language-to-vision communication are iteratively performed until the entire sent
ence is generated. Our framework seamlessly bridges the gap between visual perce
ption and LLMs and outperforms previous VLMs by a large margin on compositional
reasoning benchmarks (e.g., ~20% in HICO-DET mAP, ~14% in Cola top-1 accuracy, a
nd ~3% on ARO top-1 accuracy). We also achieve state-of-the-art performances on
traditional vision-language tasks such as referring expression comprehension and
 visual question answering.
****
Samuel Pegg,Kai Li,Xiaolin Hu
RTFS-Net: Recurrent Time-Frequency Modelling for Efficient Audio-Visual Speech S
eparation
Audio-visual speech separation methods aim to integrate different modalities to
generate high-quality separated speech, thereby enhancing the performance of dow
nstream tasks such as speech recognition. Most existing state-of-the-art (SOTA)
models operate in the time domain. However, their overly simplistic approach to
modeling acoustic features often necessitates larger and more computationally in
tensive models in order to achieve SOTA performance. In this paper, we present a
 novel time-frequency domain audio-visual speech separation method: Recurrent Ti
me-Frequency Separation Network (RTFS-Net), which applies its algorithms on the
complex time-frequency bins yielded by the Short-Time Fourier Transform. We mode
l and capture the time and frequency dimensions of the audio independently using
 a multi-layered RNN along each dimension. Furthermore, we introduce a unique at

tention-based fusion technique for the efficient integration of audio and visual information, and a new mask separation approach that takes advantage of the intrinsic spectral nature of the acoustic features for a clearer separation. RTFS-Net outperforms the prior SOTA method in both inference speed and separation quality while reducing the number of parameters by 90% and MACs by 83%. This is the first time-frequency domain audio-visual speech separation method to outperform all contemporary time-domain counterparts.
****
Shuyang Yu,Junyuan Hong,Haobo Zhang,Haotao Wang,Zhangyang Wang,Jiayu Zhou
Safe and Robust Watermark Injection with a Single OoD Image
Training a high-performance deep neural network requires large amounts of data and computational resources.
Protecting the intellectual property (IP) and commercial ownership of a deep model is challenging yet increasingly crucial.
A major stream of watermarking strategies implants verifiable backdoor triggers by poisoning training samples, but these are often unrealistic due to data privacy and safety concerns and are vulnerable to minor model changes such as fine-tuning.
To overcome these challenges, we propose a safe and robust backdoor-based watermark injection technique that leverages the diverse knowledge from a single out-of-distribution (OoD) image, which serves as a secret key for IP verification.
The independence of training data makes it agnostic to third-party promises of IP security.
We induce robustness via random perturbation of model parameters during watermark injection to defend against common watermark removal attacks, including fine-tuning, pruning, and model extraction.
Our experimental results demonstrate that the proposed watermarking approach is not only time- and sample-efficient without training data, but also robust against the watermark removal attacks above.
****
Quoc Viet Vo,Ehsan Abbasnejad,Damith Ranasinghe
BRUSLEATTACK: A QUERY-EFFICIENT SCORE- BASED BLACK-BOX SPARSE ADVERSARIAL ATTACK
We study the unique, less-well understood problem of generating sparse adversarial samples simply by observing the score-based replies to model queries. Sparse attacks aim to discover a minimum number—the l_0 bounded—perturbations to model inputs to craft adversarial examples and misguide model decisions. But, in contrast to query-based dense attack counterparts against black-box models, constructing sparse adversarial perturbations, even when models serve confidence score information to queries in a score-based setting, is non-trivial. Because, such an attack leads to: i) an NP-hard problem; and ii) a non-differentiable search space. We develop the BRUSLEATTACK—a new, faster (more query-efficient) algorithm formulation for the problem. We conduct extensive attack evaluations including an attack demonstration against a Machine Learning as a Service (MLaaS) offering exemplified by __Google Cloud Vision__ and robustness testing of adversarial training regimes and a recent defense against black-box attacks. The proposed attack scales to achieve state-of-the-art attack success rates and query efficiency on standard computer vision tasks such as ImageNet across different model architectures. Our artifacts and DIY attack samples are available on GitHub. Importantly, our work facilitates faster evaluation of model vulnerabilities and raises our vigilance on the safety, security and reliability of deployed systems.
****
Zihan Wang,Arthur Jacot
Implicit bias of SGD in L_2-regularized linear DNNs: One-way jumps from high to low rank
The L_{2}-regularized loss of Deep Linear Networks (DLNs) with
more than one hidden layers has multiple local minima, corresponding
to matrices with different ranks. In tasks such as matrix completion,
the goal is to converge to the local minimum with the smallest rank
that still fits the training data. While rank-underestimating minima
can be avoided since they do not fit the data, GD might get

stuck at rank-overestimating minima. We show that with SGD, there is always a pr obability to jump
from a higher rank minimum to a lower rank one, but the probability
of jumping back is zero. More precisely, we define a sequence of sets
$B_{1}\subset B_{2}\subset\cdots\subset B_{R}$ so that B_{r}
contains all minima of rank r or less (and not more) that are absorbing
for small enough ridge parameters λ and learning rates η:
SGD has prob. 0 of leaving B_{r}, and from any starting point there
is a non-zero prob. for SGD to go in B_{r}.
****
Minyoung Park,Mirae Do,Yeon Jae Shin,Jaeseok Yoo,Jongkwang Hong,Joongrock Kim,Ch ul Lee
H2O-SDF: Two-phase Learning for 3D Indoor Reconstruction using Object Surface Fi elds
Advanced techniques using Neural Radiance Fields (NeRF), Signed Distance Fields (SDF), and Occupancy Fields have recently emerged as solutions for 3D indoor sce ne reconstruction. We introduce a novel two-phase learning approach, H2O-SDF, t hat discriminates between object and non-object regions within indoor environmen ts. This method achieves a nuanced balance, carefully preserving the geometric i ntegrity of room layouts while also capturing intricate surface details of speci fic objects. A cornerstone of our two-phase learning framework is the introducti on of the Object Surface Field (OSF), a novel concept designed to mitigate the p ersistent vanishing gradient problem that has previously hindered the capture of high-frequency details in other methods. Our proposed approach is validated thr ough several experiments that include ablation studies.
****
Hyosoon Jang,Minsu Kim,Sungsoo Ahn
Learning Energy Decompositions for Partial Inference in GFlowNets
This paper studies generative flow networks (GFlowNets) to sample objects from t he Boltzmann energy distribution via a sequence of actions. In particular, we fo cus on improving GFlowNet with partial inference: training flow functions with t he evaluation of the intermediate states or transitions. To this end, the recent ly developed forward-looking GFlowNet reparameterizes the flow functions based o n evaluating the energy of intermediate states. However, such an evaluation of i ntermediate energies may (i) be too expensive or impossible to evaluate and (ii) even provide misleading training signals under large energy fluctuations along the sequence of actions. To resolve this issue, we propose learning energy decom positions for GFlowNets (LED-GFN). Our main idea is to (i) decompose the energy of an object into learnable potential functions defined on state transitions and (ii) reparameterize the flow functions using the potential functions. In partic ular, to produce informative local credits, we propose to regularize the potenti al to change smoothly over the sequence of actions. It is also noteworthy that t raining GFlowNet with our learned potential can preserve the optimal policy. We empirically verify the superiority of LED-GFN in five problems including the gen eration of unstructured and maximum independent sets, molecular graphs, and RNA sequences.
****
Nicklas Hansen,Hao Su,Xiaolong Wang
TD-MPC2: Scalable, Robust World Models for Continuous Control
TD-MPC is a model-based reinforcement learning (RL) algorithm that performs loca l trajectory optimization in the latent space of a learned implicit (decoder-fre e) world model. In this work, we present TD-MPC2: a series of improvements upon the TD-MPC algorithm. We demonstrate that TD-MPC2 improves significantly over ba selines across 104 online RL tasks spanning 4 diverse task domains, achieving co nsistently strong results with a single set of hyperparameters. We further show that agent capabilities increase with model and data size, and successfully trai n a single 317M parameter agent to perform 80 tasks across multiple task domains , embodiments, and action spaces. We conclude with an account of lessons, opport unities, and risks associated with large TD-MPC2 agents.

Explore videos, models, data, code, and more at https://tdmpc2.com
****
Alaa Saade,Steven Kapturowski,Daniele Calandriello,Charles Blundell,Pablo Sprechmann,Leopoldo Sarra,Oliver Groth,Michal Valko,Bilal Piot
Unlocking the Power of Representations in Long-term Novelty-based Exploration
We introduce Robust Exploration via Clustering-based Online Density Estimation (RECODE), a non-parametric method for novelty-based exploration that estimates visitation counts for clusters of states based on their similarity in a chosen embedding space. By adapting classical clustering to the nonstationary setting of Deep RL, RECODE can efficiently track state visitation counts over thousands of episodes. We further propose a novel generalization of the inverse dynamics loss, which leverages masked transformer architectures for multi-step prediction; which in conjunction with \DETOCS achieves a new state-of-the-art in a suite of challenging 3D-exploration tasks in DM-Hard-8. RECODE also sets new state-of-the-art in hard exploration Atari games, and is the first agent to reach the end screen in "Pitfall!"
****
Mahdi Kallel,Debabrota Basu,Riad Akrour,Carlo D'Eramo
Augmented Bayesian Policy Search
Deterministic policies are often preferred over stochastic ones when implemented on physical systems. They can prevent erratic and harmful behaviors while being easier to implement and interpret. However, in practice, exploration is largely performed by stochastic policies.
First-order Bayesian Optimization (BO) methods offer a principled way of performing exploration using deterministic policies. This is done through a learned probabilistic model of the objective function and its gradient. Nonetheless, such approaches treat policy search as a black-box problem, and thus, neglect the reinforcement learning nature of the problem. In this work, we leverage the performance difference lemma to introduce a novel mean function for the probabilistic model. This results in augmenting BO methods with the action-value function. Hence, we call our method Augmented Bayesian Search (ABS).
Interestingly, this new mean function enhances the posterior gradient with the deterministic policy gradient, effectively bridging the gap between BO and policy gradient methods. The resulting algorithm combines the convenience of the direct policy search with the scalability of reinforcement learning.
We validate ABS on high-dimensional locomotion problems and demonstrate competitive performance compared to existing direct policy search schemes.
****
Edward J Hu,Moksh Jain,Eric Elmoznino,Younesse Kaddar,Guillaume Lajoie,Yoshua Bengio,Nikolay Malkin
Amortizing intractable inference in large language models
Autoregressive large language models (LLMs) compress knowledge from their training data through next-token conditional distributions. This limits tractable querying of this knowledge to start-to-end autoregressive sampling. However, many tasks of interest---including sequence continuation, infilling, and other forms of constrained generation---involve sampling from intractable posterior distributions. We address this limitation by using amortized Bayesian inference to sample from these intractable posteriors. Such amortization is algorithmically achieved by fine-tuning LLMs via diversity-seeking reinforcement learning algorithms: generative flow networks (GFlowNets). We empirically demonstrate that this distribution-matching paradigm of LLM fine-tuning can serve as an effective alternative to maximum-likelihood training and reward-maximizing policy optimization. As an important application, we interpret chain-of-thought reasoning as a latent variable modeling problem and demonstrate that our approach enables data-efficient adaptation of LLMs to tasks that require multi-step rationalization and tool use.
****
Max Ku,Tianle Li,Kai Zhang,Yujie Lu,Xingyu Fu,Wenwen Zhuang,Wenhu Chen
ImagenHub: Standardizing the evaluation of conditional image generation models
Recently, a myriad of conditional image generation and editing models have been developed to serve different downstream tasks, including text-to-image generatio

n, text-guided image editing, subject-driven image generation, control-guided image generation, etc. However, we observe huge inconsistencies in experimental conditions: datasets, inference, and evaluation metrics -- render fair comparisons difficult.
This paper proposes ImagenHub, which is a one-stop library to standardize the inference and evaluation of all the conditional image generation models. Firstly, we define seven prominent tasks and curate high-quality evaluation datasets for them. Secondly, we built a unified inference pipeline to ensure fair comparison. Thirdly, we design two human evaluation scores, i.e. Semantic Consistency and Perceptual Quality, along with comprehensive guidelines to evaluate generated images. We train expert raters to evaluate the model outputs based on the proposed metrics. Our human evaluation achieves a high inter-worker agreement of Krippendorff’s alpha on 76\% models with a value higher than 0.4. We comprehensively evaluated a total of around 30 models and observed three key takeaways: (1) the existing models’ performance is generally unsatisfying except for Text-guided Image Generation and Subject-driven Image Generation, with 74\% models achieving an overall score lower than 0.5. (2) we examined the claims from published papers and found 83\% of them hold with a few exceptions. (3) None of the existing automatic metrics has a Spearman's correlation higher than 0.2 except subject-driven image generation. Moving forward, we will continue our efforts to evaluate newly published models and update our leaderboard to keep track of the progress in conditional image generation.
****
Florian Frantzen,Michael T Schaub
Learning From Simplicial Data Based on Random Walks and 1D Convolutions
Triggered by limitations of graph-based deep learning methods in terms of computational expressivity and model flexibility, recent years have seen a surge of interest in computational models that operate on higher-order topological domains such as hypergraphs and simplicial complexes. While the increased expressivity of these models can indeed lead to a better classification performance and a more faithful representation of the underlying system, the computational cost of these higher-order models can increase dramatically. To this end, we here explore a simplicial complex neural network learning architecture based on random walks and fast 1D convolutions (SCRaWl), in which we can adjust the increase in computational cost by varying the length and number of random walks considered while accounting for higher-order relationships. Importantly, due to the random walk-based design, the expressivity of the proposed architecture is provably incomparable to that of existing message-passing simplicial neural networks. We empirically evaluate SCRaWl on real-world datasets and show that it outperforms other simplicial neural networks.
****
Zhengyi Luo,Jinkun Cao,Josh Merel,Alexander Winkler,Jing Huang,Kris M. Kitani,Weipeng Xu
Universal Humanoid Motion Representations for Physics-Based Control
We present a universal motion representation that encompasses a comprehensive range of motor skills for physics-based humanoid control. Due to the high-dimensionality of humanoid control as well as the inherent difficulties in reinforcement learning, prior methods have focused on learning skill embeddings for a narrow range of movement styles (e.g. locomotion, game characters) from specialized motion datasets. This limited scope hampers its applicability in complex tasks. Our work closes this gap, significantly increasing the coverage of motion representation space. To achieve this, we first learn a motion imitator that can imitate all of human motion from a large, unstructured motion dataset. We then create our motion representation by distilling skills directly from the imitator. This is achieved using an encoder-decoder structure with a variational information bottleneck. Additionally, we jointly learn a prior conditioned on proprioception (humanoid's own pose and velocities) to improve model expressiveness and sampling efficiency for downstream tasks. Sampling from the prior, we can generate long, stable, and diverse human motions. Using this latent space for hierarchical RL, we show that our policies solve tasks using natural and realistic human behavior.

We demonstrate the effectiveness of our motion representation by solving generative tasks and motion tracking using VR controllers.
****
Licheng Wen,Daocheng Fu,Xin Li,Xinyu Cai,Tao MA,Pinlong Cai,Min Dou,Botian Shi,Liang He,Yu Qiao
DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models
Recent advancements in autonomous driving have relied on data-driven approaches, which are widely adopted but face challenges including dataset bias, overfitting, and uninterpretability.
Drawing inspiration from the knowledge-driven nature of human driving, we explore the question of how to instill similar capabilities into autonomous driving systems and summarize a paradigm that integrates an interactive environment, a driver agent, as well as a memory component to address this question.
Leveraging large language models (LLMs) with emergent abilities, we propose the DiLu framework, which combines a Reasoning and a Reflection module to enable the system to perform decision-making based on common-sense knowledge and evolve continuously.
Extensive experiments prove DiLu's capability to accumulate experience and demonstrate a significant advantage in generalization ability over reinforcement learning-based methods.
Moreover, DiLu is able to directly acquire experiences from real-world datasets which highlights its potential to be deployed on practical autonomous driving systems.
To the best of our knowledge, we are the first to leverage knowledge-driven capability in decision-making for autonomous vehicles. Through the proposed DiLu framework, LLM is strengthened to apply knowledge and to reason causally in the autonomous driving domain.

Project page: https://pjlab-adg.github.io/DiLu/
****
Zhuqing Liu,Xin Zhang,Jia Liu,Zhengyuan Zhu,Songtao Lu
PILOT: An $\mathcal{O}(1/K)$-Convergent Approach for Policy Evaluation with Nonlinear Function Approximation
Learning an accurate value function for a given policy is a critical step in solving reinforcement learning (RL) problems. So far, however, the convergence speed and sample complexity performances of most existing policy evaluation algorithms remain unsatisfactory, particularly with non-linear function approximation. This challenge motivates us to develop a new path-integrated primal-dual stochastic gradient (PILOT) method, that is able to achieve a fast convergence speed for RL policy evaluation with nonlinear function approximation. To further alleviate the periodic full gradient evaluation requirement, we further propose an enhanced method with an adaptive-batch adjustment called PILOT$^+$. The main advantages of our methods include: i) PILOT allows the use of {\em{constant}} step sizes and achieves the $\mathcal{O}(1/K)$ convergence rate to first-order stationary points of non-convex policy evaluation problems; ii) PILOT is a generic {\em{single}}-timescale algorithm that is also applicable for solving a large class of non-convex strongly-concave minimax optimization problems; iii) By adaptively adjusting the batch size via historical stochastic gradient information, PILOT$^+$ is more sample-efficient empirically without loss of theoretical convergence rate. Our extensive numerical experiments verify our theoretical findings and showcase the high efficiency of the proposed PILOT and PILOT$^+$ algorithms compared with the state-of-the-art methods.
****
Soham Gadgil,Ian Connick Covert,Su-In Lee
Estimating Conditional Mutual Information for Dynamic Feature Selection
Dynamic feature selection, where we sequentially query features to make accurate predictions with a minimal budget, is a promising paradigm to reduce feature acquisition costs and provide transparency into the prediction process. The problem is challenging, however, as it requires both making predictions with arbitrary

feature sets and learning a policy to identify the most valuable selections. He
re, we take an information-theoretic perspective and prioritize features based o
n their mutual information with the response variable. The main challenge is imp
lementing this policy, and we design a new approach that estimates the mutual in
formation in a discriminative rather than a generative fashion. Building on our
learning approach, we introduce several further improvements: allowing variable
feature budgets across samples, enabling non-uniform costs between features, inc
orporating prior information, and exploring modern architectures to handle parti
al input information. We find that our method provides consistent gains over rec
ent state-of-the-art methods across a variety of datasets.
****
Sunghyeon Woo,Sunwoo Lee,Dongsuk Jeon
ALAM: Averaged Low-Precision Activation for Memory-Efficient Training of Transfo
rmer Models
One of the key challenges in deep neural network training is the substantial amo
unt of GPU memory required to store activations obtained in the forward pass. Va
rious Activation-Compressed Training (ACT) schemes have been proposed to mitigat
e this issue; however, it is challenging to adopt those approaches in recent tra
nsformer-based large language models (LLMs), which experience significant perfor
mance drops when the activations are deeply compressed during training. In this
paper, we introduce ALAM, a novel ACT framework that utilizes average quantizati
on and a lightweight sensitivity calculation scheme, enabling large memory savin
g in LLMs while maintaining training performance. We first demonstrate that comp
ressing activations into their group average values minimizes the gradient varia
nce. Employing this property, we propose Average Quantization which provides hig
h-quality deeply compressed activations with an effective precision of less than
 1 bit and improved flexibility of precision allocation. In addition, we present
 a cost-effective yet accurate sensitivity calculation algorithm that solely rel
ies on the L2 norm of parameter gradients, substantially reducing memory overhea
d due to sensitivity calculation. In experiments, the ALAM framework significant
ly reduces activation memory without compromising accuracy, achieving up to a 10
$\times$ compression rate in LLMs.
****
Atsushi Nitanda,Kazusato Oko,Taiji Suzuki,Denny Wu
Anisotropy helps: improved statistical and computational complexity of the mean-
field Langevin dynamics under structured data
Recent works have shown that neural networks optimized by gradient-based methods
 can adapt to sparse or low-dimensional target functions through feature learnin
g; an often studied target is the sparse parity function defined on the unit hyp
ercube. However, such isotropic data setting does not capture the anisotropy and
 low intrinsic dimensionality exhibited in realistic datasets.
In this work, we address this shortcoming by studying how gradient-based feature
 learning interacts with structured (anisotropic) input data: we consider the sp
arse parity problem on high-dimensional orthotope where the feature coordinates
have varying magnitudes, and analyze the learning complexity of the mean-field L
angevin dynamics (MFLD), which describes the noisy gradient descent update on tw
o-layer neural network. We show that the statistical complexity (i.e. sample siz
e) and computational complexity (i.e. width of the neural network) of MFLD can b
oth be improved when prominent directions of the anisotropic input data aligns w
ith the support of the target function. Moreover, by employing an anisotropic we
ight decay regularization determined by the gradient covariance, the problem can
 be efficiently learned by a constant-width neural network.
****
Runtian Zhai,Rattana Pukdee,Roger Jin,Maria Florina Balcan,Pradeep Kumar Ravikum
ar
Spectrally Transformed Kernel Regression
Unlabeled data is a key component of modern machine learning. In general, the ro
le
of unlabeled data is to impose a form of smoothness, usually from the similarity
information encoded in a base kernel, such as the ■-neighbor kernel or the adjac

ency
matrix of a graph. This work revisits the classical idea of spectrally transform
ed
kernel regression (STKR), and provides a new class of general and scalable STKR
estimators able to leverage unlabeled data. Intuitively, via spectral transforma
tion,
STKR exploits the data distribution for which unlabeled data can provide additio
nal
information. First, we show that STKR is a principled and general approach,
by characterizing a universal type of “target smoothness”, and proving that any
sufficiently smooth function can be learned by STKR. Second, we provide scalable
STKR implementations for the inductive setting and a general transformation
function, while prior work is mostly limited to the transductive setting. Third,
we
derive statistical guarantees for two scenarios: STKR with a known polynomial
transformation, and STKR with kernel PCA when the transformation is unknown.
Overall, we believe that this work helps deepen our understanding of how to work
with unlabeled data, and its generality makes it easier to inspire new methods.
****
Jiaxu Zhang,Shaoli Huang,Zhigang Tu,Xin Chen,Xiaohang Zhan,Gang YU,Ying Shan
TapMo: Shape-aware Motion Generation of Skeleton-free Characters
Previous motion generation methods are limited to the pre-rigged 3D human model, hindering their applications in the animation of various non-rigged characters. In this work, we present TapMo, a Text-driven Animation PIpeline for synthesizing Motion in a broad spectrum of skeleton-free 3D characters. The pivotal innovation in TapMo is its use of shape deformation-aware features as a condition to guide the diffusion model, thereby enabling the generation of mesh-specific motions for various characters. Specifically, TapMo comprises two main components - Mesh Handle Predictor and Shape-aware Diffusion Module. Mesh Handle Predictor predicts the skinning weights and clusters mesh vertices into adaptive handles for deformation control, which eliminates the need for traditional skeletal rigging. Shape-aware Motion Diffusion synthesizes motion with mesh-specific adaptations. This module employs text-guided motions and mesh features extracted during the first stage, preserving the geometric integrity of the animations by accounting for the character's shape and deformation. Trained in a weakly-supervised manner, TapMo can accommodate a multitude of non-human meshes, both with and without associated text motions. We demonstrate the effectiveness and generalizability of TapMo through rigorous qualitative and quantitative experiments. Our results reveal that TapMo consistently outperforms existing auto-animation methods, delivering superior-quality animations for both seen or unseen heterogeneous 3D characters.
****
James B Simon,Dhruva Karkada,Nikhil Ghosh,Mikhail Belkin
More is Better: when Infinite Overparameterization is Optimal and Overfitting is Obligatory
In our era of enormous neural networks, empirical progress has been driven by the philosophy that *more is better.*
Recent deep learning practice has found repeatedly that larger model size, more data, and more computation (resulting in lower training loss) optimizing to near-interpolation improves performance. In this paper, we give theoretical backing to these empirical observations by showing that these three properties hold in random feature (RF) regression, a class of models equivalent to shallow networks with only the last layer trained.

Concretely, we first show that the test risk of RF regression decreases monotonically with both the number of features and samples, provided the ridge penalty is tuned optimally. In particular, this implies that infinite width RF architectures are preferable to those of any finite width. We then proceed to demonstrate that, for a large class of tasks characterized by powerlaw eigenstructure, training to near-zero training loss is *obligatory:* near-optimal performance can *on

ly* be achieved when the training error is much smaller than the test error. Gro
unding our theory in real-world data, we find empirically that standard computer
 vision tasks with convolutional neural kernels clearly fall into this class. Ta
ken together, our results tell a simple, testable story of the benefits of overp
arameterization and overfitting in random feature models.
****
Yili Wang,Kaixiong Zhou,Ninghao Liu,Ying Wang,Xin Wang
Efficient Sharpness-Aware Minimization for Molecular Graph Transformer Models
Sharpness-aware minimization (SAM) has received increasing attention in computer
 vision since it can effectively eliminate the sharp local minima from the train
ing trajectory and mitigate generalization degradation. However, SAM requires tw
o sequential gradient computations during the optimization of each step: one to
obtain the perturbation gradient and the other to obtain the updating gradient.
Compared with the base optimizer (e.g., Adam), SAM doubles the time overhead due
 to the additional perturbation gradient. By dissecting the theory of SAM and ob
serving the training gradient of the molecular graph transformer, we propose a n
ew algorithm named GraphSAM, which reduces the training cost of SAM and improve
s the generalization performance of graph transformer models.
There are two key factors that contribute to this result: (i) \textit{gradient a
pproximation}: we use the updating gradient of the previous step to approximate
the perturbation gradient at the intermediate steps smoothly (\textbf{increases
efficiency}); (ii) \textit{loss landscape approximation}: we theoretically prove
 that the loss landscape of GraphSAM is limited to a small range centered on the
 expected loss of SAM (\textbf{guarantees generalization performance}). The exte
nsive experiments on six datasets with different tasks demonstrate the superiori
ty of GraphSAM, especially in optimizing the model update process.
****
Chengming Hu,Haolun Wu,Xuan Li,Chen Ma,Xi Chen,Boyu Wang,Jun Yan,Xue Liu
Less or More From Teacher: Exploiting Trilateral Geometry For Knowledge Distilla
tion
Knowledge distillation aims to train a compact student network using soft superv
ision from a larger teacher network and hard supervision from ground truths. How
ever, determining an optimal knowledge fusion ratio that balances these supervis
ory signals remains challenging. Prior methods generally resort to a constant or
 heuristic-based fusion ratio, which often falls short of a proper balance. In t
his study, we introduce a novel adaptive method for learning a sample-wise knowl
edge fusion ratio, exploiting both the correctness of teacher and student, as we
ll as how well the student mimics the teacher on each sample. Our method natural
ly leads to the \textit{intra-sample} trilateral geometric relations among the s
tudent prediction ($\mathcal{S}$), teacher prediction ($\mathcal{T}$), and groun
d truth ($\mathcal{G}$). To counterbalance the impact of outliers, we further ex
tend to the \textit{inter-sample} relations, incorporating the teacher's global
average prediction ($\mathcal{\bar{T}})$ for samples within the same class. A s
imple neural network then learns the implicit mapping from the intra- and inter-
sample relations to an adaptive, sample-wise knowledge fusion ratio in a bilevel
-optimization manner. Our approach provides a simple, practical, and adaptable s
olution for knowledge distillation that can be employed across various architect
ures and model sizes. Extensive experiments demonstrate consistent improvements
over other loss re-weighting methods on image classification, attack detection,
and click-through rate prediction.
****
Xinyu Hu,Pengfei Tang,Simiao Zuo,Zihan Wang,Bowen Song,Qiang Lou,Jian Jiao,Denis
 X Charles
Evoke: Evoking Critical Thinking Abilities in LLMs via Reviewer-Author Prompt Ed
iting
Large language models (LLMs) have made impressive progress in natural language p
rocessing. These models rely on proper human instructions (or prompts) to genera
te suitable responses. However, the potential of LLMs are not fully harnessed by
 commonly-used prompting methods: many human-in-the-loop algorithms employ ad-ho
c procedures for prompt selection; while auto prompt generation approaches are e

ssentially searching all possible prompts randomly and inefficiently. We propose Evoke, an automatic prompt refinement framework. In Evoke, there are two instances of a same LLM: one as a reviewer (LLM-Reviewer), it scores the current prompt; the other as an author (LLM-Author), it edits the prompt by considering the edit history and the reviewer's feedback. Such an author-reviewer feedback loop ensures that the prompt is refined in each iteration. We further aggregate a data selection approach to Evoke, where only the hard samples are exposed to the LLM. The hard samples are more important because the LLM can develop deeper understanding of the tasks out of them, while the model may already know how to solve the easier cases. Experimental results show that Evoke significantly outperforms existing methods. For instance, in the challenging task of logical fallacy detection, Evoke scores above 80, while all other baseline methods struggle to reach 20.
****
Aleksa Sukovic,Goran Radanovic
Reward Design for Justifiable Sequential Decision-Making
Equipping agents with the capacity to justify made decisions using supporting evidence represents a cornerstone of accountable decision-making. Furthermore, ensuring that justifications are in line with human expectations and societal norms is vital, especially in high-stakes situations such as healthcare. In this work, we propose the use of a debate-based reward model for reinforcement learning agents, where the outcome of a zero-sum debate game quantifies the justifiability of a decision in a particular state. This reward model is then used to train a justifiable policy, whose decisions can be more easily corroborated with supporting evidence. In the debate game, two argumentative agents take turns providing supporting evidence for two competing decisions. Given the proposed evidence, a proxy of a human judge evaluates which decision is better justified. We demonstrate the potential of our approach in learning policies for prescribing and justifying treatment decisions of septic patients. We show that augmenting the reward with the feedback signal generated by the debate-based reward model yields policies highly favored by the judge when compared to the policy obtained solely from the environment rewards, while hardly sacrificing any performance. Moreover, in terms of the overall performance and justifiability of trained policies, the debate-based feedback is comparable to the feedback obtained from an ideal judge proxy that evaluates decisions using the full information encoded in the state. This suggests that the debate game outputs key information contained in states that is most relevant for evaluating decisions, which in turn substantiates the practicality of combining our approach with human-in-the-loop evaluations. Lastly, we showcase that agents trained via multi-agent debate learn to propose evidence that is resilient to refutations and closely aligns with human preferences.
****
Sam Bond-Taylor,Chris G. Willcocks
∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States
This paper introduces ∞-Diff, a generative diffusion model defined in an infinite-dimensional Hilbert space, which can model infinite resolution data. By training on randomly sampled subsets of coordinates and denoising content only at those locations, we learn a continuous function for arbitrary resolution sampling. Unlike prior neural field-based infinite-dimensional models, which use point-wise functions requiring latent compression, our method employs non-local integral operators to map between Hilbert spaces, allowing spatial context aggregation. This is achieved with an efficient multi-scale function-space architecture that operates directly on raw sparse coordinates, coupled with a mollified diffusion process that smooths out irregularities. Through experiments on high-resolution datasets, we found that even at an $8\times$ subsampling rate, our model retains high-quality diffusion. This leads to significant run-time and memory savings, delivers samples with lower FID scores, and scales beyond the training resolution while retaining detail.
****
Eslam Mohamed BAKR,Mohamed Ayman Mohamed,Mahmoud Ahmed,Habib Slim,Mohamed Elhoseiny

CoT3DRef: Chain-of-Thoughts Data-Efficient 3D Visual Grounding
3D visual grounding is the ability to localize objects in 3D scenes conditioned on
an input utterance. Most existing methods devote the referring head to localize the
referred object directly. However, this approach will fail in complex scenarios and
not illustrate how and why the network reaches the final decision. In this paper,
we address this question “Can we design an interpretable 3D visual grounding
framework that has the potential to mimic the human perception system?”. To this
end, we formulate the 3D visual grounding problem as a sequence-to-sequence
(Seq2Seq) task by first predicting a chain of anchors and then utilizing them to pre-
dict the final target. Following the chain of thoughts approach enables us to decom-
pose the referring task into interpretable intermediate steps, which in turn, boosts
the performance and makes our framework extremely data-efficient. Interpretabil-
ity not only improves the overall performance but also helps us identify failure
cases. Moreover, our proposed framework can be easily integrated into any existing
architecture. We validate our approach through comprehensive experiments on the
Nr3D and Sr3D benchmarks and show consistent performance gains compared to
existing methods without requiring any manually annotated data. Furthermore, our
proposed framework, dubbed CoT3DRef, is significantly data-efficient, whereas
when trained only on 10% of the data, we match the SOTA performance that trained
on the entire data. The code is available at https://cot3dref.github.io/.
****
Tennison Liu,Nicolás Astorga,Nabeel Seedat,Mihaela van der Schaar
Large Language Models to Enhance Bayesian Optimization
Bayesian optimization (BO) is a powerful approach for optimizing complex and expensive-to-evaluate black-box functions. Its importance is underscored in many applications, notably including hyperparameter tuning, but its efficacy depends on efficiently balancing exploration and exploitation. While there has been substantial progress in BO methods, striking this balance remains a delicate process. In this light, we present \texttt{LLAMBO}, a novel approach that integrates the capabilities of Large Language Models (LLM) within BO. At a high level, we frame the BO problem in natural language, enabling LLMs to iteratively \emph{propose} and \emph{evaluate} promising solutions conditioned on historical evaluations. More specifically, we explore how combining contextual understanding, few-shot learning proficiency, and domain knowledge of LLMs can improve model-based BO. Our findings illustrate that \texttt{LLAMBO} is effective at zero-shot warmstarting, and enhances surrogate modeling and candidate sampling, especially in the early stages of search when observations are sparse. Our approach is performed in context and does not require LLM finetuning. Additionally, it is modular by design, allowing individual components to be integrated into existing BO frameworks, or function cohesively as an end-to-end method. We empirically validate \texttt{LLAMBO}'s efficacy on the problem of hyperparameter tuning, highlighting strong empirical performance across a range of diverse benchmarks, proprietary, and synthetic tasks.
****
Nathan Godey,Éric Villemonte de la Clergerie,Benoît Sagot
Headless Language Models: Learning without Predicting with Contrastive Weight Tying
Self-supervised pre-training of language models usually consists in predicting probability distributions over extensive token vocabularies. In this study, we propose an innovative method that shifts away from probability prediction and instead focuses on reconstructing input embeddings in a contrastive fashion via Constrastive Weight Tying (CWT). We apply this approach to pretrain Headless Languag

e Models in both monolingual and multilingual contexts. Our method offers practi
cal advantages, substantially reducing training computational requirements by up
 to 20 times, while simultaneously enhancing downstream performance and data eff
iciency. We observe a significant +1.6 GLUE score increase and a notable +2.7 LA
MBADA accuracy improvement compared to classical LMs within similar compute budg
ets.
****
Robert Jenssen
MAP IT to Visualize Representations
MAP IT visualizes representations by taking a fundamentally different approach t
o dimensionality reduction. MAP IT aligns distributions over discrete marginal p
robabilities in the input space versus the target space, thus capturing informat
ion in local regions, as opposed to current methods which align based on individ
ual probabilities between pairs of data points (states) only. The MAP IT theory
reveals that alignment based on a projective divergence avoids normalization of
weights (to obtain true probabilities) entirely, and further reveals a dual view
point via continuous densities and kernel smoothing. MAP IT is shown to produce
visualizations which capture class structure better than the current state of th
e art while being inherently scalable.
****
Dinghuai Zhang,Ricky T. Q. Chen,Cheng-Hao Liu,Aaron Courville,Yoshua Bengio
Diffusion Generative Flow Samplers: Improving learning signals through partial t
rajectory optimization
We tackle the problem of sampling from intractable high-dimensional density func
tions, a fundamental task that often appears in machine learning and statistics.

We extend recent sampling-based approaches that leverage controlled stochastic p
rocesses to model approximate samples from these target densities.
The main drawback of these approaches is that the training objective requires fu
ll trajectories to compute, resulting in sluggish credit assignment issues due t
o use of entire trajectories and a learning signal present only at the terminal
time.
In this work, we present Diffusion Generative Flow Samplers (DGFS), a sampling-b
ased framework where the learning process can be tractably broken down into shor
t partial trajectory segments, via parameterizing an additional ``flow function'
'.
Our method takes inspiration from the theory developed for generative flow netwo
rks (GFlowNets), allowing us to make use of intermediate learning signals.
Through various challenging experiments, we demonstrate that DGFS achieves more
accurate estimates of the normalization constant than closely-related prior meth
ods.
****
Lirui Wang,Yiyang Ling,Zhecheng Yuan,Mohit Shridhar,Chen Bao,Yuzhe Qin,Bailin Wa
ng,Huazhe Xu,Xiaolong Wang
GenSim: Generating Robotic Simulation Tasks via Large Language Models
Collecting large amounts of real-world interaction data to train general robotic
 policies is often prohibitively expensive, thus motivating the use of simulatio
n data. However, existing methods for data generation have generally focused on
scene-level diversity (e.g., object instances and poses) rather than task-level
diversity, due to the human effort required to come up with and verify novel tas
ks. This has made it challenging for policies trained on simulation data to demo
nstrate significant task-level generalization. In this paper, we propose to auto
matically generate rich simulation environments and expert demonstrations by exp
loiting a large language models' (LLM) grounding and coding ability. Our approac
h, dubbed GenSim, has two modes: goal-directed generation, wherein a target task
 is given to the LLM and the LLM proposes a task curriculum to solve the target
task, and exploratory generation, wherein the LLM bootstraps from previous task
s and iteratively proposes novel tasks that would be helpful in solving more com
plex tasks. We use GPT4 to expand the existing benchmark by ten times to over 10
0 tasks, on which we conduct supervised finetuning and evaluate several LLMs inc

luding finetuned GPTs and Code Llama on code generation for robotic simulation tasks. Furthermore, we observe that LLMs-generated simulation programs can enhance task-level generalization significantly when used for multitask policy training. We further find that with minimal sim-to-real adaptation, the multitask policies pretrained on GPT4-generated simulation tasks exhibit stronger transfer to unseen long-horizon tasks in the real world and outperform baselines by 25%. See our project website (https://gen-sim.github.io) and demo (https://huggingface.co/spaces/Gen-Sim/Gen-Sim) for visualizations and open-source models and datasets.
****
Ziqi Gao,Xiangguo Sun,Zijing Liu,Yu Li,Hong Cheng,Jia Li
Protein Multimer Structure Prediction via Prompt Learning
Understanding the 3D structures of protein multimers is crucial, as they play a vital role in regulating various cellular processes. It has been empirically confirmed that the multimer structure prediction (MSP) can be well handled in a step-wise assembly fashion using provided dimer structures and predicted protein-protein interactions (PPIs). However, due to the biological gap in the formation of dimers and larger multimers, directly applying PPI prediction techniques can often cause a poor generalization to the MSP task. To address this challenge, we aim to extend the PPI knowledge to multimers of different scales (i.e., chain numbers). Specifically, we propose PromptMSP, a pre-training and Prompt tuning framework for Multimer Structure Prediction. First, we tailor the source and target tasks for effective PPI knowledge learning and efficient inference, respectively. We design PPI-inspired prompt learning to narrow the gaps of two task formats and generalize the PPI knowledge to multimers of different scales. We provide a meta-learning strategy to learn a reliable initialization of the prompt model, enabling our prompting framework to effectively adapt to limited data for large-scale multimers. Empirically, we achieve both significant accuracy (RMSD and TM-Score) and efficiency improvements compared to advanced MSP models.
****
Dingling Yao,Danru Xu,Sebastien Lachapelle,Sara Magliacane,Perouz Taslakian,Georg Martius,Julius von Kügelgen,Francesco Locatello
Multi-View Causal Representation Learning with Partial Observability
We present a unified framework for studying the identifiability of representations learned from simultaneously observed views, such as different data modalities. We allow a partially observed setting in which each view constitutes a nonlinear mixture of a subset of underlying latent variables, which can be causally related.
We prove that the information shared across all subsets of any number of views can be learned up to a smooth bijection using contrastive learning and a single encoder per view.
We also provide graphical criteria indicating which latent variables can be identified through a simple set of rules, which we refer to as identifiability algebra. Our general framework and theoretical results unify and extend several previous work on multi-view nonlinear ICA, disentanglement, and causal representation learning. We experimentally validate our claims on numerical, image, and multi-modal data sets. Further, we demonstrate that the performance of prior methods is recovered in different special cases of our setup.
Overall, we find that access to multiple partial views offers unique opportunities for identifiable representation learning, enabling the discovery of latent structures from purely observational data.
****
Daouda Sow,Sen Lin,Zhangyang Wang,Yingbin Liang
Doubly Robust Instance-Reweighted Adversarial Training
Assigning importance weights to adversarial data has achieved great success in training adversarially robust networks under limited model capacity. However, existing instance-reweighted adversarial training (AT) methods heavily depend on heuristics and/or geometric interpretations to determine those importance weights, making these algorithms lack rigorous theoretical justification/guarantee. Moreover, recent research has shown that adversarial training suffers from a severe non-uniform robust performance across the training distribution, e.g., data poin

ts belonging to some classes can be much more vulnerable to adversarial attacks than others. To address both issues, in this paper, we propose a novel doubly-ro bust instance reweighted AT framework, which allows to obtain the importance wei ghts via exploring distributionally robust optimization (DRO) techniques, and at the same time boosts the robustness on the most vulnerable examples. In particu lar, our importance weights are obtained by optimizing the KL-divergence regular ized loss function, which allows us to devise new algorithms with a theoretical convergence guarantee.
Experiments on standard classification datasets demonstrate that our proposed ap proach outperforms related state-of-the-art baseline methods in terms of average robust performance, and at the same time improves the robustness against attack s on the weakest data points. Codes can be found in the Supplement.
****
Chong Mou,Xintao Wang,Jiechong Song,Ying Shan,Jian Zhang
DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models
Despite the ability of text-to-image (T2I) diffusion models to generate high-qua lity images, transferring this ability to accurate image editing remains a chall enge. In this paper, we propose a novel image editing method, DragonDiffusion, e nabling Drag-style manipulation on Diffusion models. Specifically, we treat imag e editing as the change of feature correspondence in a pre-trained diffusion mod el. By leveraging feature correspondence, we develop energy functions that align with the editing target, transforming image editing operations into gradient gu idance. Based on this guidance approach, we also construct multi-scale guidance that considers both semantic and geometric alignment. Furthermore, we incorporat e a visual cross-attention strategy based on a memory bank design to ensure cons istency between the edited result and original image. Benefiting from these effi cient designs, all content editing and consistency operations come from the feat ure correspondence without extra model fine-tuning. Extensive experiments demons trate that our method has promising performance on various image editing tasks, including within a single image (e.g., object moving, resizing, and content drag ging) or across images (e.g., appearance replacing and object pasting). Code is available at https://github.com/MC-E/DragonDiffusion.
****
Yaning Jia,Chunhui Zhang,Soroush Vosoughi
Aligning Relational Learning with Lipschitz Fairness
Relational learning has gained significant attention, led by the expressiveness of Graph Neural Networks (GNNs) on graph data. While the inherent biases in comm on graph data are involved in GNN training, it poses a serious challenge to cons training the GNN output perturbations induced by input biases, thereby safeguard ing fairness during training. The Lipschitz bound, a technique from robust stati stics, can limit the maximum changes in the output concerning the input, taking into account associated irrelevant biased factors. It is an efficient and provab le method to examine the output stability of machine learning models without inc urring additional computational costs. Recently, its use in controlling the stab ility of Euclidean neural networks, the calculation of the precise Lipschitz bou nd remains elusive for non-Euclidean neural networks like GNNs, especially withi n fairness contexts. However, no existing research has investigated Lipschitz bo unds to shed light on stabilizing the GNN outputs, especially when working on gr aph data with implicit biases. To narrow this gap, we begin with the general GNN s operating on relational data, and formulate a Lipschitz bound to limit the cha nges in the output regarding biases associated with the input. Additionally, we theoretically analyze how the Lipschitz bound of a GNN model could constrain the output perturbations induced by biases learned from data for fairness training. We experimentally validate the Lipschitz bound's effectiveness in limiting bias es of the model output. Finally, from a training dynamics perspective, we demons trate why the theoretical Lipschitz bound can effectively guide the GNN training to better trade-off between accuracy and fairness.
****
Lu Chen,Siyu Lou,Benhao Huang,Quanshi Zhang
Defining and extracting generalizable interaction primitives from DNNs

Faithfully summarizing the knowledge encoded by a deep neural network (DNN) into
 a few symbolic primitive patterns without losing much information represents a
core challenge in explainable AI. To this end, Ren et al. (2024) have derived a
series of theorems to prove that the inference score of a DNN can be explained a
s a small set of interactions between input variables. However, the lack of gene
ralization power makes it still hard to consider such interactions as faithful p
rimitive patterns encoded by the DNN. Therefore, given different DNNs trained fo
r the same task, we develop a new method to extract interactions that are shared
 by these DNNs. Experiments show that the extracted interactions can better refl
ect common knowledge shared by different DNNs.
****
Xiaodan Chen,Xiucheng Li,Bo Liu,Zhijun Li
Biased Temporal Convolution Graph Network for Time Series Forecasting with Missi
ng Values
Multivariate time series forecasting plays an important role in various applicat
ions ranging from meteorology study, traffic management to economics planning. I
n the past decades, many efforts have been made toward accurate and reliable for
ecasting methods development under the assumption of intact input data. However,
 the time series data from real-world scenarios is often partially observed due
to device malfunction or costly data acquisition, which can seriously impede the
 performance of the existing approaches. A naive employment of imputation method
s unavoidably involves error accumulation and leads to suboptimal solutions. Mot
ivated by this, we propose a Biased Temporal Convolution Graph Network that join
tly captures the temporal dependencies and spatial structure. In particular, we
inject bias into the two carefully developed modules, the Multi-Scale Instance P
artialTCN and Biased GCN, to account for missing patterns. The experimental resu
lts show that our proposed model is able to achieve up to 9.93\% improvements
over the existing methods on five real-world benchmark datasets. Our code is ava
ilable at: https://github.com/chenxiaodanhit/BiTGraph.
****
Nils Lukas,Abdulrahman Diaa,Lucas Fenaux,Florian Kerschbaum
Leveraging Optimization for Adaptive Attacks on Image Watermarks
Untrustworthy users can misuse image generators to synthesize high-quality deepf
akes and engage in unethical activities. Watermarking deters misuse by marking g
enerated content with a hidden message, enabling its detection using a secret wa
termarking key. A core security property of watermarking is robustness, which st
ates that an attacker can only evade detection by substantially degrading image
quality. Assessing robustness requires designing an adaptive attack for the spec
ific watermarking algorithm. When evaluating watermarking algorithms and their (
adaptive) attacks, it is challenging to determine whether an adaptive attack is
optimal, i.e., the best possible attack. We solve this problem by defining an ob
jective function and then approach adaptive attacks as an optimization problem.
The core idea of our adaptive attacks is to replicate secret watermarking keys l
ocally by creating surrogate keys that are differentiable and can be used to opt
imize the attack's parameters. We demonstrate for Stable Diffusion models that s
uch an attacker can break all five surveyed watermarking methods at no visible d
egradation in image quality. Optimizing our attacks is efficient and requires le
ss than 1 GPU hour to reduce the detection accuracy to 6.3% or less. Our finding
s emphasize the need for more rigorous robustness testing against adaptive, lear
nable attackers.
****
Eric Qu,Yansen Wang,Xufang Luo,Wenqiang He,Kan Ren,Dongsheng Li
CNN Kernels Can Be the Best Shapelets
Shapelets and CNN are two typical approaches to model time series. Shapelets aim
 at finding a set of sub-sequences that extract feature-based interpretable shap
es, but may suffer from accuracy and efficiency issues. CNN performs well by enc
oding sequences with a series of hidden representations, but lacks interpretabil
ity. In this paper, we demonstrate that shapelets are essentially equivalent to
a specific type of CNN kernel with a squared norm and pooling. Based on this fin
ding, we propose ShapeConv, an interpretable CNN layer with its kernel serving a

s shapelets to conduct time-series modeling tasks in both supervised and unsuper vised settings. By incorporating shaping regularization, we enforce the similari ty for maximum interpretability. We also find human knowledge can be easily inje cted to ShapeConv by adjusting its initialization and model performance is boost ed with it. Experiments show that ShapeConv can achieve state-of-the-art perform ance on time-series benchmarks without sacrificing interpretability and controll ability.
****
Alexander Ashcroft,Ayan Das,Yulia Gryaditskaya,Zhiyu Qu,Yi-Zhe Song
Modelling complex vector drawings with stroke-clouds
Vector drawings are innately interactive as they preserve creational cues. Despi te
this desirable property they remain relatively under explored due to the difficu lties
in modeling complex vector drawings. This is in part due to the primarily _seque ntial and auto-regressive nature_ of existing approaches failing to scale beyond simple
drawings. In this paper, we define generative models over _highly complex_ vecto r
drawings by first representing them as "stroke-clouds" – _sets_ of arbitrary car dinality comprised of semantically meaningful strokes. The dimensionality of the strokes is a design choice that allows the model to adapt to a range of complexi ties.
We learn to encode these _set of strokes_ into compact latent codes by a probabi listic
reconstruction procedure backed by _De-Finetti's Theorem of Exchangability_. The parametric generative model is then defined over the latent vectors of the encod ed
stroke-clouds. The resulting "Latent stroke-cloud generator (LSG)" thus captures the distribution of complex vector drawings on an implicit _set space_. We demon strate the efficacy of our model on complex drawings (a newly created Anime line-art dataset) through a range
of generative tasks.
****
Xinhua Cheng,Tianyu Yang,Jianan Wang,Yu Li,Lei Zhang,Jian Zhang,Li Yuan
Progressive3D: Progressively Local Editing for Text-to-3D Content Creation with Complex Semantic Prompts
Recent text-to-3D generation methods achieve impressive 3D content creation capa city thanks to the advances in image diffusion models and optimizing strategies. However, current methods struggle to generate correct 3D content for a complex prompt in semantics, i.e., a prompt describing multiple interacted objects bindi ng with different attributes. In this work, we propose a general framework named Progressive3D, which decomposes the entire generation into a series of locally progressive editing steps to create precise 3D content for complex prompts, and we constrain the content change to only occur in regions determined by user-defi ned region prompts in each editing step. Furthermore, we propose an overlapped s emantic component suppression technique to encourage the optimization process to focus more on the semantic differences between prompts. Extensive experiments d emonstrate that the proposed Progressive3D framework generates precise 3D conten t for prompts with complex semantics through progressive editing steps and is ge neral for various text-to-3D methods driven by different 3D representations.
****
Haotian Xue,Chumeng Liang,Xiaoyu Wu,Yongxin Chen
Toward effective protection against diffusion-based mimicry through score distil lation
While generative diffusion models excel in producing high-quality images, they c an also be misused to mimic authorized images, posing a significant threat to AI systems. Efforts have been made to add calibrated perturbations to protect imag es from diffusion-based mimicry pipelines. However, most of the existing methods are too ineffective and even impractical to be used by individual users due to

their high computation and memory requirements. In this work, we present novel f
indings on attacking latent diffusion models (LDM) and propose new plug-and-play
 strategies for more effective protection. In particular, we explore the bottlen
eck in attacking an LDM, discovering that the encoder module rather than the den
oiser module is the vulnerable point. Based on this insight, we present our stra
tegy using Score Distillation Sampling (SDS) to double the speed of protection a
nd reduce memory occupation by half without compromising its strength. Additiona
lly, we provide a robust protection strategy by counterintuitively minimizing th
e semantic loss, which can assist in generating more natural perturbations. Fina
lly, we conduct extensive experiments to substantiate our findings and comprehen
sively evaluate our newly proposed strategies. We hope our insights and protecti
ve measures can contribute to better defense against malicious diffusion-based m
imicry, advancing the development of secure AI systems.
****
Alon Ziv,Itai Gat,Gael Le Lan,Tal Remez,Felix Kreuk,Jade Copet,Alexandre Défosse
z,Gabriel Synnaeve,Yossi Adi
Masked Audio Generation using a Single Non-Autoregressive Transformer
We introduce MAGNeT, a masked generative sequence modeling method that operates
directly over several streams of audio tokens. Unlike prior work, MAGNeT is comp
rised of a single-stage, non-autoregressive transformer. During training, we pre
dict spans of masked tokens obtained from a masking scheduler, while during infe
rence we gradually construct the output sequence using several decoding steps. T
o further enhance the quality of the generated audio, we introduce a novel resco
ring method in which, we leverage an external pre-trained model to rescore and r
ank predictions from MAGNeT, which will be then used for later decoding steps. L
astly, we explore a hybrid version of MAGNeT, in which we fuse between autoregre
ssive and non-autoregressive models to generate the first few seconds in an auto
regressive manner while the rest of the sequence is being decoded in parallel. W
e demonstrate the efficiency of MAGNeT for the task of text-to-music and text-to
-audio generation and conduct an extensive empirical evaluation, considering bot
h objective metrics and human studies. The proposed approach is comparable to th
e evaluated baselines, while being significantly faster (x7 faster than the au
toregressive baseline). Through ablation studies and analysis, we shed light on
the importance of each of the components comprising MAGNeT, together with pointi
ng to the trade-offs between autoregressive and non-autoregressive modeling, con
sidering latency, throughput, and generation quality. Samples are available on o
ur demo page https://pages.cs.huji.ac.il/adiyoss-lab/MAGNeT.
****
Mingzhen Huang,Shan Jia,Zhou Zhou,Yan Ju,Jialing Cai,Siwei Lyu
Exposing Text-Image Inconsistency Using Diffusion Models
In the battle against widespread online misinformation, a growing problem is tex
t-image inconsistency, where images are misleadingly paired with texts with diff
erent intent or meaning. Existing classification-based methods for text-image in
consistency can identify contextual inconsistencies but fail to provide explaina
ble justifications for their decisions that humans can understand. Although more
 nuanced, human evaluation is impractical at scale and susceptible to errors. To
 address these limitations, this study introduces D-TIIL (Diffusion-based Text-I
mage Inconsistency Localization), which employs text-to-image diffusion models t
o localize semantic inconsistencies in text and image pairs. These models, train
ed on large-scale datasets act as ``omniscient" agents that filter out irrelevan
t information and incorporate background knowledge to identify inconsistencies.
In addition, D-TIIL uses text embeddings and modified image regions to visualize
 these inconsistencies. To evaluate D-TIIL's efficacy, we introduce a new TIIL d
ataset containing 14K consistent and inconsistent text-image pairs. Unlike exist
ing datasets, TIIL enables assessment at the level of individual words and image
 regions and is carefully designed to represent various inconsistencies. D-TIIL
offers a scalable and evidence-based approach to identifying and localizing text
-image inconsistency, providing a robust framework for future research combating
 misinformation.
****

Hongtao Wu,Ya Jing,Chilam Cheang,Guangzeng Chen,Jiafeng Xu,Xinghang Li,Minghuan Liu,Hang Li,Tao Kong
Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation
Generative pre-trained models have demonstrated remarkable effectiveness in language and vision domains by learning useful representations. In this paper, we extend the scope of this effectiveness by showing that visual robot manipulation can significantly benefit from large-scale video generative pre-training. We introduce GR-1, a GPT-style model designed for multi-task language-conditioned visual robot manipulation. GR-1 takes as inputs a language instruction, a sequence of observation images, and a sequence of robot states. It predicts robot actions as well as future images in an end-to-end manner. Thanks to a flexible design, GR-1 can be seamlessly finetuned on robot data after pre-trained on a large-scale video dataset. We perform extensive experiments on the challenging CALVIN benchmark and a real robot. On CALVIN benchmark, our method outperforms state-of-the-art baseline methods and improves the success rate from 88.9% to 94.9%. In the setting of zero-shot unseen scene generalization, GR-1 improves the success rate from 53.3% to 85.4%. In real robot experiments, GR-1 also outperforms baseline methods and shows strong potentials in generalization to unseen scenes and objects. We provide inaugural evidence that a unified GPT-style transformer, augmented with large-scale video generative pre-training, exhibits remarkable generalization to multi-task visual robot manipulation. Project page: https://GR1-Manipulation.github.io
****
Hadar Sivan,Moshe Gabel,Assaf Schuster
FOSI: Hybrid First and Second Order Optimization
Popular machine learning approaches forgo second-order information due to the difficulty of computing curvature in high dimensions.
We present FOSI, a novel meta-algorithm that improves the performance of any base first-order optimizer by efficiently incorporating second-order information during the optimization process.
In each iteration, FOSI implicitly splits the function into two quadratic functions defined on orthogonal subspaces, then uses a second-order method to minimize the first, and the base optimizer to minimize the other.
We formally analyze FOSI's convergence and the conditions under which it improves a base optimizer.
Our empirical evaluation
demonstrates that FOSI improves the convergence rate and optimization time of first-order methods such as Heavy-Ball and Adam, and outperforms second-order methods (K-FAC and L-BFGS).
****
Yulu Gan,Sungwoo Park,Alexander Marcel Schubert,Anthony Philippakis,Ahmed Alaa
InstructCV: Instruction-Tuned Text-to-Image Diffusion Models as Vision Generalists
Recent advances in generative diffusion models have enabled text-controlled synthesis of realistic and diverse images with impressive quality. Despite these remarkable advances, the application of text-to-image generative models in computer vision for standard visual recognition tasks remains limited. The current de facto approach for these tasks is to design model architectures and loss functions that are tailored to the task at hand. In this paper, we develop a unified language interface for computer vision tasks that abstracts away task specific design choices and enables task execution by following natural language instructions. Our approach involves casting multiple computer vision tasks as text-to-image generation problems. Here, the text represents an instruction describing the task, and the resulting image is a visually-encoded task output. To train our model, we pool commonly-used computer vision datasets covering a range of tasks, including segmentation, object detection, depth estimation, and classification. We then use a large language model to paraphrase prompt templates that convey the specific tasks to be conducted on each image, and through this process, we create a multi-modal and multi-task training dataset comprising input and output images

along with annotated instructions. Following the InstructPix2Pix architecture, w e apply instruction-tuning to a text-to-image diffusion model using our construc ted dataset, steering its functionality from a generative model to an instructio n-guided multi-task vision learner. Experiments demonstrate that our model, dubb ed InstructCV, performs competitively compared to other generalist and task-spec ific vision models. Moreover, it exhibits compelling generalization capabilities to unseen data, categories, and user instructions.
****
Yang Fu,Shalini De Mello,Xueting Li,Amey Kulkarni,Jan Kautz,Xiaolong Wang,Sifei Liu
3D Reconstruction with Generalizable Neural Fields using Scene Priors
High-fidelity 3D scene reconstruction has been substantially advanced by recent progress in neural fields. However, most existing methods train a separate netwo rk from scratch for each individual scene. This is not scalable, inefficient, an d unable to yield good results given limited views. While learning-based multi-v iew stereo methods alleviate this issue to some extent, their multi-view setting makes it less flexible to scale up and to broad applications. Instead, we intro duce training generalizable Neural Fields incorporating scene Priors (NFPs).
The NFP network maps any single-view RGB-D image into signed distance and radian ce values. A complete scene can be reconstructed by merging individual frames in the volumetric space WITHOUT a fusion module, which provides better flexibility . The scene priors can be trained on large-scale datasets, allowing for fast ad aptation to the reconstruction of a new scene with fewer views. NFP not only dem onstrates SOTA scene reconstruction performance and efficiency, but it also supp orts single-image novel-view synthesis, which is under-explored in neural fields . More qualitative results are available at: https://oasisyang.github.io/neural-prior.
****
Yichen Li,Yilun Du,Chao Liu,Chao Liu,Francis Williams,Michael Foshey,Benjamin Ec kart,Jan Kautz,Joshua B. Tenenbaum,Antonio Torralba,Wojciech Matusik
Learning to Jointly Understand Visual and Tactile Signals
Modeling and analyzing object and shape has been well studied in the past. Howev er, manipulation of these complex tools and articulated objects remains difficul t for autonomous agents. Our human hands, however, are dexterous and adaptive. W e can easily adapt a manipulation skill on one object to all objects in the clas s and to other similar classes. Our intuition comes from that there is a close c onnection between manipulations and topology and articulation of objects. The po ssible articulation of objects indicates the types of manipulation necessary to operate the object. In this work, we aim to take a manipulation perspective to u nderstand everyday objects and tools. We collect a multi-modal visual-tactile da taset that contains paired full-hand force pressure maps and manipulation videos . We also propose a novel method to learn a cross-modal latent manifold that all ow for cross-modal prediction and discovery of latent structure in different dat a modalities. We conduct extensive experiments to demonstrate the effectiveness of our method.
****
Fabricio Arend Torres,Marcello Massimo Negri,Marco Inversi,Jonathan Aellen,Volke r Roth
Lagrangian Flow Networks for Conservation Laws
We introduce Lagrangian Flow Networks (LFlows) for modeling fluid densities and velocities continuously in space and time.
By construction, the proposed LFlows satisfy the continuity equation,
a PDE describing mass conservation in its differential form.
Our model is based on the insight that solutions to the continuity equation can be expressed as
time-dependent density transformations via differentiable and invertible maps.
This follows from classical theory of the existence and uniqueness of Lagrangian flows for smooth vector fields.
Hence, we model fluid densities by transforming a base density with parameterize d diffeomorphisms conditioned on time.

The key benefit compared to methods relying on numerical ODE solvers or PINNs is that the analytic expression of the velocity is always consistent with changes in density.
Furthermore, we require neither expensive numerical solvers, nor additional penalties to enforce the PDE.
LFlows show higher predictive accuracy in density modeling tasks compared to competing models in 2D and 3D,
while being computationally efficient.
As a real-world application, we model bird migration based on sparse weather radar measurements.
****
Yichong Leng,ZHifang Guo,Kai Shen,Zeqian Ju,Xu Tan,Eric Liu,Yufei Liu,Dongchao Yang,leying zhang,Kaitao Song,Lei He,Xiangyang Li,sheng zhao,Tao Qin,Jiang Bian
PromptTTS 2: Describing and Generating Voices with Text Prompt
Speech conveys more information than text, as the same word can be uttered in various voices to convey diverse information. Compared to traditional text-to-speech (TTS) methods relying on speech prompts (reference speech) for voice variability, using text prompts (descriptions) is more user-friendly since speech prompts can be hard to find or may not exist at all. TTS approaches based on the text prompt face two main challenges: 1) the one-to-many problem, where not all details about voice variability can be described in the text prompt, and 2) the limited availability of text prompt datasets, where vendors and large cost of data labeling are required to write text prompts for speech. In this work, we introduce PromptTTS 2 to address these challenges with a variation network to provide variability information of voice not captured by text prompts, and a prompt generation pipeline to utilize the large language models (LLM) to compose high quality text prompts. Specifically, the variation network predicts the representation extracted from the reference speech (which contains full information about voice variability) based on the text prompt representation. For the prompt generation pipeline, it generates text prompts for speech with a speech language understanding model to recognize voice attributes (e.g., gender, speed) from speech and a large language model to formulate text prompts based on the recognition results. Experiments on a large-scale (44K hours) speech dataset demonstrate that compared to the previous works, PromptTTS 2 generates voices more consistent with text prompts and supports the sampling of diverse voice variability, thereby offering users more choices on voice generation. Additionally, the prompt generation pipeline produces high-quality text prompts, eliminating the large labeling cost. The demo page of PromptTTS 2 is available (https://speechresearch.github.io/prompttts2).
****
Darshan Patil,Janarthanan Rajendran,Glen Berseth,Sarath Chandar
Intelligent Switching for Reset-Free RL
In the real world, the strong episode resetting mechanisms that are needed to train
agents in simulation are unavailable. The resetting assumption limits the potential
of reinforcement learning in the real world, as providing resets to an agent usually
requires the creation of additional handcrafted mechanisms or human interventions.
Recent work aims to train agents (forward) with learned resets by constructing a second (backward) agent that returns the forward agent to the initial state. We
find that the termination and timing of the transitions between these two agents are crucial for algorithm success. With this in mind, we create a new algorithm, Reset Free RL with Intelligently Switching Controller (RISC) which intelligently switches between the two agents based on the agent’s confidence in achieving its current goal. Our new method achieves state-of-the-art performance on several challenging environments for reset-free RL.
****

Maarten Buyl,MaryBeth Defrance,Tijl De Bie
fairret: a Framework for Differentiable Fairness Regularization Terms
Current tools for machine learning fairness only admit a limited range of fairness definitions and have seen little integration with automatic differentiation libraries, despite the central role these libraries play in modern machine learning pipelines.

We introduce a framework of fairness regularization terms (fairret) which quantify bias as modular objectives that are easily integrated in automatic differentiation pipelines. By employing a general definition of fairness in terms of linear-fractional statistics, a wide class of fairrets can be computed efficiently. Experiments show the behavior of their gradients and their utility in enforcing fairness with minimal loss of predictive power compared to baselines. Our contribution includes a PyTorch implementation of the fairret framework.
****
Zhiwei Deng,Ting Chen,Yang Li
Perceptual Group Tokenizer: Building Perception with Iterative Grouping
Human visual recognition system shows astonishing capability of compressing visual information into a set of tokens containing rich representations without label supervision. One critical driving principle behind it is perceptual grouping. Despite being widely used in computer vision in the early 2010s, it remains a mystery whether perceptual grouping can be leveraged to derive a neural visual recognition backbone that generates as powerful representations. In this paper, we propose the Perceptual Group Tokenizer, a model that entirely relies on grouping operations to extract visual features and perform self-supervised representation learning, where a series of grouping operations are used to iteratively hypothesize the context for pixels or superpixels to refine feature representations. We show that the proposed model can achieve competitive performance compared to state-of-the-art vision architectures, and inherits desirable properties including adaptive computation without re-training, and interpretability. Specifically, Perceptual Group Tokenizer achieves 79.7% on ImageNet-1K self-supervised learning benchmark with linear probe evaluation, marking a new progress under this paradigm.
****
Sipeng Zheng,jiazheng liu,Yicheng Feng,Zongqing Lu
Steve-Eye: Equipping LLM-based Embodied Agents with Visual Perception in Open Worlds
Recent studies have presented compelling evidence that large language models (LLMs) can equip embodied agents with the self-driven capability to interact with the world, which marks an initial step toward versatile robotics. However, these efforts tend to overlook the visual richness of open worlds, rendering the entire interactive process akin to ``a blindfolded text-based game.'' Consequently, LLM-based agents frequently encounter challenges in intuitively comprehending their surroundings and producing responses that are easy to understand. In this paper, we propose Steve-Eye, an end-to-end trained large multimodal model to address this limitation. Steve-Eye integrates the LLM with a visual encoder to process visual-text inputs and generate multimodal feedback. We adopt a semi-automatic strategy to collect an extensive dataset comprising 850K open-world instruction pairs, enabling our model to encompass three essential functions for an agent: multimodal perception, foundational knowledge base, and skill prediction and planning. Lastly, we develop three open-world evaluation benchmarks and carry out experiments from a wide range of perspectives to validate our model's capability to strategically act and plan. The project’s website and code can be found at https://sites.google.com/view/steve-eye.
****
Michael Gastpar,Ido Nachum,Jonathan Shafer,Thomas Weinberger
Fantastic Generalization Measures are Nowhere to be Found
We study the notion of a generalization bound being _uniformly tight_, meaning that the difference between the bound and the population loss is small for all learning algorithms and all population distributions. Numerous generalization boun

ds have been proposed in the literature as potential explanations for the abilit y of neural networks to generalize in the overparameterized setting.
However, in their paper "Fantastic Generalization Measures and Where to Find The m," Jiang et al. (2020) examine more than a dozen generalization bounds, and sho w empirically that none of them are uniformly tight. This raises the question of whether uniformly-tight generalization bounds are at all possible in the overpa rameterized setting. We consider two types of generalization bounds: (1) bounds that may depend on the training set and the learned hypothesis (e.g., margin bou nds). We prove mathematically that no such bound can be uniformly tight in the o verparameterized setting; (2) bounds that may in addition also depend on the lea rning algorithm (e.g., stability bounds). For these bounds, we show a trade-off between the algorithm's performance and the bound's tightness. Namely, if the al gorithm achieves good accuracy on certain distributions, then no generalization bound can be uniformly tight for it in the overparameterized setting. We explain how these formal results can, in our view, inform research on generalization b ounds for neural networks, while stressing that other interpretations of these r esults are also possible.
****
Carl Hvarfner,Frank Hutter,Luigi Nardi
A General Framework for User-Guided Bayesian Optimization
The optimization of expensive-to-evaluate black-box functions is prevalent in va rious scientific disciplines. Bayesian optimization is an automatic, general and sample-efficient method to solve these problems with minimal knowledge of the t he underlying function dynamics. However, the ability of Bayesian optimization t o incorporate prior knowledge or beliefs about the function at hand in order to accelerate the optimization is limited, which reduces its appeal for knowledgeab le practitioners with tight budgets. To allow domain experts to customize the o ptimization routine, we propose ColaBO, the first Bayesian-principled framework for incorporating prior beliefs beyond the typical kernel structure, such as the likely location of the optimizer or the optimal value. The generality of ColaBO makes it applicable across different Monte Carlo acquisition functions and type s of user beliefs. We empirically demonstrate ColaBO's ability to substantially accelerate optimization when the prior information is accurate, and to retain ap proximately default performance when it is misleading.
****
Dawid Jan Kopiczko,Tijmen Blankevoort,Yuki M Asano
VeRA: Vector-based Random Matrix Adaptation
Low-rank adapation (LoRA) is a popular method that reduces the number of trainab le parameters when finetuning large language models, but still faces acute stora ge challenges when scaling to even larger models or deploying numerous per-user or per-task adapted models. In this work, we present Vector-based Random Matrix Adaptation (VeRA), which significantly reduces the number of trainable parameter s compared to LoRA, yet maintains the same performance. It achieves this by usin g a single pair of low-rank matrices shared across all layers and learning small scaling vectors instead. We demonstrate its effectiveness on the GLUE and E2E b enchmarks, image classification tasks, and show its application in instruction-t uning of 7B and 13B language models. Website: https://dkopi.github.io/vera
****
William Merrill,Ashish Sabharwal
The Expressive Power of Transformers with Chain of Thought
Recent theoretical work has identified surprisingly simple reasoning problems, s uch as checking if two nodes in a graph are connected or simulating finite-state machines, that are provably unsolvable by standard transformers that answer imm ediately after reading their input. However, in practice, transformers' reasonin g can be improved by allowing them to use a "chain of thought" or "scratchpad", i.e., generate and condition on a sequence of intermediate tokens before answeri ng. Motivated by this, we ask: *Does such intermediate generation fundamentally extend the computational power of a decoder-only transformer?* We show that the answer is *yes*, but the amount of increase depends crucially on the amount of i ntermediate generation. For instance, we find that transformer decoders with a l

ogarithmic number of decoding steps (w.r.t. the input length) push the limits of standard transformers only slightly, while a linear number of decoding steps adds a clear new ability (under standard complexity conjectures): recognizing all regular languages. Our results also imply that linear steps keep transformer decoders within context-sensitive languages, and polynomial steps make them recognize exactly the class of polynomial-time solvable problems---the first exact characterization of a type of transformers in terms of standard complexity classes. Together, our results provide a nuanced framework for understanding how the length of a transformer's chain of thought or scratchpad impacts its reasoning power.
****
Ethan Steinberg,Jason Alan Fries,Yizhe Xu,Nigam Shah
MOTOR: A Time-to-Event Foundation Model For Structured Medical Records
We present a self-supervised, time-to-event (TTE) foundation model called MOTOR (Many Outcome Time Oriented Representations) which is pretrained on timestamped sequences of events in electronic health records (EHR) and health insurance claims. TTE models are used for estimating the probability distribution of the time until a specific event occurs, which is an important task in medical settings. TTE models provide many advantages over classification using fixed time horizons, including naturally handling censored observations, but are challenging to train with limited labeled data. MOTOR addresses this challenge by pretraining on up to 55M patient records (9B clinical events). We evaluate MOTOR's transfer learning performance on 19 tasks, across 3 patient databases (a private EHR system, MIMIC-IV, and Merative claims data). Task-specific models adapted from MOTOR improve time-dependent C statistics by 4.6\% over state-of-the-art, improve label efficiency by up to 95\%, and are more robust to temporal distributional shifts. We further evaluate cross-site portability by adapting our MOTOR foundation model for six prediction tasks on the MIMIC-IV dataset, where it outperforms all baselines. MOTOR is the first foundation model for medical TTE predictions and we release a 143M parameter pretrained model for research use at https://huggingface.co/StanfordShahLab/motor-t-base.
****
Hanqi Zhou,Robert Bamler,Charley M Wu,Álvaro Tejero-Cantero
Predictive, scalable and interpretable knowledge tracing on structured domains
Intelligent tutoring systems optimize the selection and timing of learning materials to enhance understanding and long-term retention. This requires estimates of both the learner's progress ("knowledge tracing"; KT), and the prerequisite structure of the learning domain ("knowledge mapping"). While recent deep learning models achieve high KT accuracy, they do so at the expense of the interpretability of psychologically-inspired models. In this work, we present a solution to this trade-off. PSI-KT is a hierarchical generative approach that explicitly models how both individual cognitive traits and the prerequisite structure of knowledge influence learning dynamics, thus achieving interpretability by design. Moreover, by using scalable Bayesian inference, PSI-KT targets the real-world need for efficient personalization even with a growing body of learners and interaction data. Evaluated on three datasets from online learning platforms, PSI-KT achieves superior multi-step **p**redictive accuracy and **s**calable inference in continual-learning settings, all while providing **i**nterpretable representations of learner-specific traits and the prerequisite structure of knowledge that causally supports learning. In sum, predictive, scalable and interpretable knowledge tracing with solid knowledge mapping lays a key foundation for effective personalized learning to make education accessible to a broad, global audience.
****
Dhruva Tirumala,Thomas Lampe,Jose Enrique Chen,Tuomas Haarnoja,Sandy Huang,Guy Lever,Ben Moran,Tim Hertweck,Leonard Hasenclever,Martin Riedmiller,Nicolas Heess,Markus Wulfmeier
Replay across Experiments: A Natural Extension of Off-Policy RL
Replaying data is a principal mechanism underlying the stability and data efficiency of off-policy reinforcement learning (RL).
We present an effective yet simple framework to extend the use of replays across

multiple experiments, minimally adapting the RL workflow for sizeable improvements in controller performance and research iteration times.
At its core, Replay across Experiments (RaE) involves reusing experience from previous experiments to improve exploration and bootstrap learning while reducing required changes to a minimum in comparison to prior work.
We empirically show benefits across a number of RL algorithms and challenging control domains spanning both locomotion and manipulation, including hard exploration tasks from egocentric vision.
Through comprehensive ablations, we demonstrate robustness to the quality and amount of data available and various hyperparameter choices. Finally, we discuss how our approach can be applied more broadly across research life cycles and can increase resilience by reloading data across random seeds or hyperparameter variations.
****
Ruibo Liu,Ruixin Yang,Chenyan Jia,Ge Zhang,Diyi Yang,Soroush Vosoughi
Training Socially Aligned Language Models on Simulated Social Interactions
The goal of social alignment for AI systems is to make sure these models can conduct themselves appropriately following social values. Unlike humans who establish a consensus on value judgments through social interaction, current language models (LMs) are trained to rigidly recite the corpus in social isolation, which causes poor generalization in unfamiliar cases and the lack of robustness under adversarial attacks. In this work, we introduce a new training paradigm that enables LMs to learn from simulated social interactions. Compared with existing methods, our method is much more scalable and efficient, and shows superior performance in alignment benchmarks and human evaluation.
****
Changwen Zhang,Wenli Ouyang,Hao Yuan,Liming Gong,Yong Sun,Ziao Guo,Zhichen Dong,Junchi Yan
Towards Imitation Learning to Branch for MIP: A Hybrid Reinforcement Learning based Sample Augmentation Approach
Branch-and-bound (B\&B) has long been favored for tackling complex Mixed Integer Programming (MIP) problems, where the choice of branching strategy plays a pivotal role. Recently, Imitation Learning (IL)-based policies have emerged as potent alternatives to traditional rule-based approaches. However, it is nontrivial to acquire high-quality training samples, and IL often converges to suboptimal variable choices for branching, restricting the overall performance. In response to these challenges, we propose a novel hybrid online and offline reinforcement learning (RL) approach to enhance the branching policy by cost-effective training sample augmentation. In the online phase, we train an online RL agent to dynamically decide the sample generation processes, drawing from either the learning-based policy or the expert policy. The objective is to strike a balance between exploration and exploitation of the sample generation process. In the offline phase, a value function is trained to fit each decision's cumulative reward and filter the samples with high cumulative returns. This dual-purpose function not only reduces training complexity but also enhances the quality of the samples. To assess the efficacy of our data augmentation mechanism, we conduct comprehensive evaluations across a range of MIP problems. The results consistently show that it excels in making superior branching decisions compared to state-of-the-art learning-based models and the open-source solver SCIP. Notably, it even often outperforms Gurobi.
****
Xuelun Shen,zhipeng cai,Wei Yin,Matthias Müller,Zijun Li,Kaixuan Wang,Xiaozhi Chen,Cheng Wang
GIM: Learning Generalizable Image Matcher From Internet Videos
Image matching is a fundamental computer vision problem. While learning-based methods achieve state-of-the-art performance on existing benchmarks, they generalize poorly to in-the-wild images. Such methods typically need to train separate models for different scene types (e.g., indoor vs. outdoor) and are impractical when the scene type is unknown in advance. One of the underlying problems is the limited scalability of existing data construction pipelines, which limits the di

versity of standard image matching datasets. To address this problem, we propose GIM, a self-training framework for learning a single generalizable model based on any image matching architecture using internet videos, an abundant and diverse data source. Given an architecture, GIM first trains it on standard domain-specific datasets and then combines it with complementary matching methods to create dense labels on nearby frames of novel videos. These labels are filtered by robust fitting, and then enhanced by propagating them to distant frames. The final model is trained on propagated data with strong augmentations. Not relying on complex 3D reconstruction makes GIM much more efficient and less likely to fail than standard SfM-and-MVS based frameworks. We also propose ZEB, the first zero-shot evaluation benchmark for image matching. By mixing data from diverse domains, ZEB can thoroughly assess the cross-domain generalization performance of different methods. Experiments demonstrate the effectiveness and generality of GIM. Applying GIM consistently improves the zero-shot performance of 3 state-of-the-art image matching architectures as the number of downloaded videos increases (Fig. 1 (a)); with 50 hours of YouTube videos, the relative zero-shot performance improves by 6.9% – 18.1%. GIM also enables generalization to extreme cross-domain data such as Bird Eye View (BEV) images of projected 3D point clouds (Fig. 1 (c)). More importantly, our single zero-shot model consistently outperforms domain-specific baselines when evaluated on downstream tasks inherent to their respective domains. The code will be released upon acceptance.
****
Rui Ye,Yaxin Du,Zhenyang Ni,Yanfeng Wang,Siheng Chen
Fake It Till Make It: Federated Learning with Consensus-Oriented Generation
In federated learning (FL), data heterogeneity is one key bottleneck that causes model divergence and limits performance. Addressing this, existing methods often regard data heterogeneity as an inherent property and propose to mitigate its adverse effects by correcting models. In this paper, we seek to break this inherent property by generating data to complement the original dataset to fundamentally mitigate heterogeneity level.
As a novel attempt from the perspective of data, we propose federated learning with consensus-oriented generation (FedCOG). FedCOG consists of two key components at the client side: complementary data generation, which generates data extracted from the shared global model to complement the original dataset, and knowledge-distillation-based model training, which distills knowledge from global model to local model based on the generated data to mitigate over-fitting the original heterogeneous dataset.
FedCOG has two critical advantages: 1) it can be a plug-and-play module to further improve the performance of most existing FL methods, and 2) it is naturally compatible with standard FL protocols such as Secure Aggregation since it makes no modification in communication process.
Extensive experiments on classical and real-world FL datasets show that FedCOG consistently outperforms state-of-the-art methods. Code is available at https://github.com/rui-ye/FedCOG.
****
Chen Qiu,Xingyu Li,Chaithanya Kumar Mummadi,Madan Ravi Ganesh,Zhenzhen Li,Lu Peng,Wan-Yi Lin
Federated Text-driven Prompt Generation for Vision-Language Models
Prompt learning for vision-language models, e.g., CoOp, has shown great success in adapting CLIP to different downstream tasks, making it a promising solution for federated learning due to computational reasons. Existing prompt learning techniques replace hand-crafted text prompts with learned vectors that offer improvements on seen classes, but struggle to generalize to unseen classes. Our work addresses this challenge by proposing Federated Text-driven Prompt Generation (FedTPG), which learns a unified prompt generation network across multiple remote clients in a scalable manner. The prompt generation network is conditioned on task-related text input, thus is context-aware, making it suitable to generalize for both seen and unseen classes. Our comprehensive empirical evaluations on nine diverse image classification datasets show that our method is superior to existing federated prompt learning methods, achieving better overall generalization on

both seen and unseen classes, as well as datasets.
****
Tianjiao Zhang,Huangjie Zheng,Jiangchao Yao,Xiangfeng Wang,Mingyuan Zhou,Ya Zhang,Yanfeng Wang
Long-tailed Diffusion Models with Oriented Calibration
Diffusion models are acclaimed for generating high-quality and diverse images. However, their performance notably degrades when trained on data with a long-tailed distribution. For long tail diffusion model generation, current works focus on the calibration and enhancement of the tail generation with head-tail knowledge transfer. The transfer process relies on the abundant diversity derived from the head class and, more significantly, the condition capacity of the model prediction. However, the dependency on the conditional model prediction to realize the knowledge transfer might exhibit bias during training, leading to unsatisfactory generation results and lack of robustness. Utilizing a Bayesian framework, we develop a weighted denoising score-matching technique for knowledge transfer directly from head to tail classes. Additionally, we incorporate a gating mechanism in the knowledge transfer process. We provide statistical analysis to validate this methodology, revealing that the effectiveness of such knowledge transfer depends on both label distribution and sample similarity, providing the insight to consider sample similarity when re-balancing the label proportion in training. We extensively evaluate our approach with experiments on multiple benchmark datasets, demonstrating its effectiveness and superior performance compared to existing methods. Code: \url{https://github.com/xiaoeyuztj/OC_LT/}.
****
Noga Alon,Dmitrii Avdiukhin,Dor Elboim,Orr Fischer,Grigory Yaroslavtsev
Optimal Sample Complexity of Contrastive Learning
Contrastive learning is a highly successful technique for learning representations of data from labeled tuples, specifying the distance relations within the tuple. We study the sample complexity of contrastive learning, i.e. the minimum number of labeled tuples sufficient for getting high generalization accuracy. We give tight bounds on the sample complexity in a variety of settings, focusing on arbitrary distance functions, ℓ_p-distances, and tree metrics. Our main result is an (almost) optimal bound on the sample complexity of learning ℓ_p-distances for integer p. For any $p \ge 1$, we show that $\tilde \Theta(nd)$ labeled tuples are necessary and sufficient for learning d-dimensional representations of n-point datasets. Our results hold for an arbitrary distribution of the input samples and are based on giving the corresponding bounds on the Vapnik-Chervonenkis/Natarajan dimension of the associated problems. We further show that the theoretical bounds on sample complexity obtained via VC/Natarajan dimension can have strong predictive power for experimental results, in contrast with the folklore belief about a substantial gap between the statistical learning theory and the practice of deep learning.
****
Yuxuan Song,Jingjing Gong,Hao Zhou,Mingyue Zheng,Jingjing Liu,Wei-Ying Ma
Unified Generative Modeling of 3D Molecules with Bayesian Flow Networks
Advanced generative model (\textit{e.g.}, diffusion model) derived from simplified continuity assumptions of data distribution, though showing promising progress, has been difficult to apply directly to geometry generation applications due to the \textit{multi-modality} and \textit{noise-sensitive} nature of molecule geometry.
This work introduces Geometric Bayesian Flow Networks (GeoBFN), which naturally fits molecule geometry by modeling diverse modalities in the differentiable parameter space of distributions. GeoBFN maintains the SE-(3) invariant density modeling property by incorporating equivariant inter-dependency modeling on parameters of distributions and unifying the probabilistic modeling of different modalities.
Through optimized training and sampling techniques, we demonstrate that GeoBFN achieves state-of-the-art performance on multiple 3D molecule generation benchmarks in terms of generation quality (90.87\% molecule stability in QM9 and 85.6\% atom stability in GEOM-DRUG\footnote{The scores are reported at 1k sampling step

s for fair comparison, and our scores could be further improved if sampling suff
iciently longer steps.}). GeoBFN can also conduct sampling with any number of st
eps to reach an optimal trade-off between efficiency and quality (\textit{e.g.},
 20$\times$ speedup without sacrificing performance).
****
Anji Liu,Mathias Niepert,Guy Van den Broeck
Image Inpainting via Tractable Steering of Diffusion Models
Diffusion models are the current state of the art for generating photorealistic
images. Controlling the sampling process for constrained image generation tasks
such as inpainting, however, remains challenging since exact conditioning on su
ch constraints is intractable. While existing methods use various techniques to
approximate the constrained posterior, this paper proposes to exploit the abilit
y of Tractable Probabilistic Models (TPMs) to exactly and efficiently compute th
e constrained posterior, and to leverage this signal to steer the denoising proc
ess of diffusion models. Specifically, this paper adopts a class of expressive T
PMs termed Probabilistic Circuits (PCs). Building upon prior advances, we furthe
r scale up PCs and make them capable of guiding the image generation process of
diffusion models. Empirical results suggest that our approach can consistently i
mprove the overall quality and semantic coherence of inpainted images across thr
ee natural image datasets (i.e., CelebA-HQ, ImageNet, and LSUN) with only ~10% a
dditional computational overhead brought by the TPM. Further, with the help of a
n image encoder and decoder, our method can readily accept semantic constraints
on specific regions of the image, which opens up the potential for more controll
ed image generation tasks. In addition to proposing a new framework for constrai
ned image generation, this paper highlights the benefit of more tractable models
 and motivates the development of expressive TPMs.
****
Yanqiao Zhu,Jeehyun Hwang,Keir Adams,Zhen Liu,Bozhao Nan,Brock Stenfors,Yuanqi D
u,Jatin Chauhan,Olaf Wiest,Olexandr Isayev,Connor W. Coley,Yizhou Sun,Wei Wang
Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks
Molecular Representation Learning (MRL) has proven impactful in numerous biochem
ical applications such as drug discovery and enzyme design. While Graph Neural N
etworks (GNNs) are effective at learning molecular representations from a 2D mol
ecular graph or a single 3D structure, existing works often overlook the flexibl
e nature of molecules, which continuously interconvert across conformations via
chemical bond rotations and minor vibrational perturbations. To better account f
or molecular flexibility, some recent works formulate MRL as an ensemble learnin
g problem, focusing on explicitly learning from a set of conformer structures. H
owever, most of these studies have limited datasets, tasks, and models. In this
work, we introduce the first MoleculAR Conformer Ensemble Learning (MARCEL) benc
hmark to thoroughly evaluate the potential of learning on con- former ensembles
and suggest promising research directions. MARCEL includes four datasets coverin
g diverse molecule- and reaction-level properties of chemically diverse molecule
s including organocatalysts and transition-metal catalysts, extending beyond the
 scope of common GNN benchmarks that are confined to drug-like molecules. In add
ition, we conduct a comprehensive empirical study, which benchmarks representati
ve 1D, 2D, and 3D MRL models, along with two strategies that explicitly incorpor
ate conformer ensembles into 3D models. Our findings reveal that direct learning
 from an accessible conformer space can improve performance on a variety of task
s and models.
****
Mudit Verma,Katherine Metcalf
Hindsight PRIORs for Reward Learning from Human Preferences
Preference based Reinforcement Learning (PbRL) removes the need to hand specify
a reward function by learning one from preference feedback over policy behaviors
. Current approaches to PbRL do not address the credit assignment problem inhere
nt in determining which parts of a behavior most contributed to a preference res
ulting in data intensive approaches and subpar reward models. We address such li
mitations by introducing a credit assignment strategy (PRIOR) that uses a forwar
d dynamics world model to approximate state importance within a trajectory and t

hen guides rewards to be proportional to state importance through an auxiliary predicted return redistribution objective. Incorporating state importance into reward learning improves the speed of policy learning, overall policy performance, and reward recovery on both locomotion and manipulation tasks. For example, PRIOR achieves 80% success rate with half the amount of data compared to baselines. The performance gains and our ablations demonstrate the benefits even a simple credit assignment strategy can have on reward learning and that state importance in forward dynamics prediction is a strong proxy for a state's contribution to a preference decision.
****
Jielong Yan,Yifan Feng,Shihui Ying,Yue Gao
Hypergraph Dynamic System
Recently, hypergraph neural networks (HGNNs) exhibit the potential to tackle tasks with high-order correlations and have achieved success in many tasks. However, existing evolution on the hypergraph has poor controllability and lacks sufficient theoretical support (like dynamic systems), thus yielding sub-optimal performance. One typical scenario is that only one or two layers of HGNNs can achieve good results and more layers lead to degeneration of performance. Under such circumstances, it is important to increase the controllability of HGNNs. In this paper, we first introduce hypergraph dynamic systems (HDS), which bridge hypergraphs and dynamic systems and characterize the continuous dynamics of representations. We then propose a control-diffusion hypergraph dynamic system by an ordinary differential equation (ODE). We design a multi-layer HDS^{ode} as a neural implementation, which contains control steps and diffusion steps. HDS^{ode} has the properties of controllability and stabilization and is allowed to capture long-range correlations among vertices. Experiments on 9 datasets demonstrate HDS^{ode} beat all compared methods. HDS^{ode} achieves stable performance with increased layers and solves the poor controllability of HGNNs. We also provide the feature visualization of the evolutionary process to demonstrate the controllability and stabilization of HDS^{ode}.
****
Youbang Sun,Zitao Li,Yaliang Li,Bolin Ding
Improving LoRA in Privacy-preserving Federated Learning
Low-rank adaptation (LoRA) is one of the most popular task-specific parameter-efficient fine-tuning (PEFT) methods on pre-trained language models for its good performance and computational efficiency.
LoRA injects a product of two trainable rank decomposition matrices over the top of each frozen pre-trained model module.
However, when applied in the setting of privacy-preserving federated learning (FL), LoRA may become unstable due to the following facts: 1) the effects of data heterogeneity and multi-step local updates are non-negligible, 2) additive noise enforced on updating gradients to guarantee differential privacy (DP) can be amplified and 3) the final performance is susceptible to hyper-parameters.
A key factor leading to these phenomena is the discordance between jointly optimizing the two low-rank matrices by local clients and separately aggregating them by the central server.
Thus, this paper proposes an efficient and effective version of LoRA, Federated Freeze A LoRA (FFA-LoRA), to alleviate these challenges and further halve the communication cost of federated fine-tuning LLMs.
The core idea of FFA-LoRA is to fix the randomly initialized non-zero matrices and only fine-tune the zero-initialized matrices.
Compared to LoRA, FFA-LoRA is motivated by practical and theoretical benefits in privacy-preserved FL.
Our experiments demonstrate that FFA-LoRA provides more consistent performance with better computational efficiency over vanilla LoRA in various FL tasks.
****
Tales Henrique Carvalho,Kenneth Tjhia,Levi Lelis
Reclaiming the Source of Programmatic Policies: Programmatic versus Latent Spaces
Recent works have introduced LEAPS and HPRL, systems that learn latent spaces of

domain-specific languages, which are used to define programmatic policies for partially observable Markov decision processes (POMDPs). These systems induce a latent space while optimizing losses such as the behavior loss, which aim to achieve locality in program behavior, meaning that vectors close in the latent space should correspond to similarly behaving programs. In this paper, we show that the programmatic space, induced by the domain-specific language and requiring no training, presents values for the behavior loss similar to those observed in latent spaces presented in previous work. Moreover, algorithms searching in the programmatic space significantly outperform those in LEAPS and HPRL. To explain our results, we measured the ``friendliness'' of the two spaces to local search algorithms. We discovered that algorithms are more likely to stop at local maxima when searching in the latent space than when searching in the programmatic space. This implies that the optimization topology of the programmatic space, induced by the reward function in conjunction with the neighborhood function, is more conducive to search than that of the latent space. This result provides an explanation for the superior performance in the programmatic space.
****
Guangxuan Xiao,Yuandong Tian,Beidi Chen,Song Han,Mike Lewis
Efficient Streaming Language Models with Attention Sinks
Deploying Large Language Models (LLMs) in streaming applications such as multi-round dialogue, where long interactions are expected, is urgently needed but poses two major challenges.
Firstly, during the decoding stage, caching previous tokens' Key and Value states (KV) consumes extensive memory.
Secondly, popular LLMs cannot generalize to longer texts than the training sequence length.
Window attention, where only the most recent KVs are cached, is a natural approach --- but we show that it fails when the text length surpasses the cache size.
We observe an interesting phenomenon, namely attention sink, that keeping the KV of initial tokens will largely recover the performance of window attention. In this paper, we first demonstrate that the emergence of attention sink is due to the strong attention scores towards initial tokens as a ``sink'' even if they are not semantically important.
Based on the above analysis, we introduce StreamingLLM, an efficient framework that enables LLMs trained with a finite length attention window to generalize to infinite sequence length without any fine-tuning.
We show that StreamingLLM can enable Llama-2, MPT, Falcon, and Pythia to perform stable and efficient language modeling with up to 4 million tokens and more.
In addition, we discover that adding a placeholder token as a dedicated attention sink during pre-training can further improve streaming deployment. In streaming settings, StreamingLLM outperforms the sliding window recomputation baseline by up to 22.2$\times$ speedup.
Code and datasets are provided in the anonymous link.
****
Yichao Shen,Zigang Geng,Yuhui Yuan,Yutong Lin,Ze Liu,Chunyu Wang,Han Hu,Nanning Zheng,Baining Guo
V-DETR: DETR with Vertex Relative Position Encoding for 3D Object Detection
We introduce a highly performant 3D object detector for point clouds using the DETR framework. The prior attempts all end up with suboptimal results because they fail to learn accurate inductive biases from the limited scale of training data. In particular, the queries often attend to points that are far away from the target objects, violating the locality principle in object detection. To address the limitation, we introduce a novel 3D Vertex Relative Position Encoding (3DV-RPE) method which computes position encoding for each point based on its relative position to the 3D boxes predicted by the queries in each decoder layer, thus providing clear information to guide the model to focus on points near the objects, in accordance with the principle of locality. Furthermore, we have systematically refined our pipeline, including data normalization, to better align with the task requirements. Our approach demonstrates remarkable performance on the demanding ScanNetV2 benchmark, showcasing substantial enhancements over the prior

state-of-the-art CAGroup3D. Specifically, we achieve an increase in AP_{25} from 75.1% to 77.8% and in ${AP}_{50}$ from 61.3% to 66.0%.
****
Longhui Yu,Weisen Jiang,Han Shi,Jincheng YU,Zhengying Liu,Yu Zhang,James Kwok,Zhenguo Li,Adrian Weller,Weiyang Liu
MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models
Large language models (LLMs) have pushed the limits of natural language understanding and exhibited excellent problem-solving ability. Despite the great success, most existing open-source LLMs (\eg, LLaMA-2) are still far away from satisfactory for solving mathematical problems due to the complex reasoning procedures. To bridge this gap, we propose \emph{MetaMath}, a finetuned language model that specializes in mathematical reasoning. Specifically, we start by bootstrapping mathematical questions by rewriting the question from multiple perspectives, which results in a new dataset called MetaMathQA. Then we finetune the LLaMA-2 models on MetaMathQA. Experimental results on two popular benchmarks (\ie, GSM8K and MATH) for mathematical reasoning demonstrate that MetaMath outperforms a suite of open-source LLMs by a significant margin. Our MetaMath-7B model achieves 66.5% on GSM8K and 19.8% on MATH, exceeding the state-of-the-art models of the same size by 11.5% and 8.7%. Particularly, MetaMath-70B achieves an accuracy of 82.3% on GSM8K, slightly better than GPT-3.5-Turbo. We release the MetaMathQA dataset, the MetaMath models with different model sizes and the training code for public use.
****
Hao Sun,Alihan Hüyük,Mihaela van der Schaar
Query-Dependent Prompt Evaluation and Optimization with Offline Inverse RL
In this study, we aim to enhance the arithmetic reasoning ability of Large Language Models (LLMs) through zero-shot prompt optimization. We identify a previously overlooked objective of query dependency in such optimization and elucidate two ensuing challenges that impede the successful and economical design of prompt optimization techniques. One primary issue is the absence of an effective method to evaluate prompts during inference when the golden answer is unavailable. Concurrently, learning via interactions with the LLMs to navigate the expansive natural language prompting space proves to be resource-intensive.
To address this, we introduce Prompt-OIRL, which harnesses offline inverse reinforcement learning to draw insights from offline prompting demonstration data. Such data exists as by-products when diverse prompts are benchmarked on open-accessible datasets. With Prompt-OIRL, the query-dependent prompt optimization objective is achieved by first learning an offline reward model. This model can evaluate any query-prompt pairs without accessing LLMs. Subsequently, a best-of-N strategy is deployed to recommend the optimal prompt. Our experimental evaluations across various LLM scales and arithmetic reasoning datasets underscore both the efficacy and economic viability of the proposed approach.
****
Yeqi Gao,Lianke Qin,Zhao Song,Yitan Wang
A Sublinear Adversarial Training Algorithm
Adversarial training is a widely used strategy for making neural networks resistant to adversarial perturbations. For a neural network of width m, n input training data in d dimension, it takes $\Omega(mnd)$ time cost per training iteration for the forward and backward computation. In this paper we analyze the convergence guarantee of adversarial training procedure on a two-layer neural network with shifted ReLU activation, and shows that only $o(m)$ neurons will be activated for each input data per iteration. Furthermore, we develop an algorithm for adversarial training with time cost $o(m n d)$ per iteration by applying half-space reporting data structure.
****
Vincent Leroy,Jerome Revaud,Thomas Lucas,Philippe Weinzaepfel
Win-Win: Training High-Resolution Vision Transformers from Two Windows
Transformers have become the standard in state-of-the-art vision architectures, achieving impressive performance on both image-level and dense pixelwise tasks. However, training vision transformers for high-resolution pixelwise tasks has a

prohibitive cost. Typical solutions boil down to hierarchical architectures, fas
t and approximate attention, or training on low-resolution crops. This latter so
lution does not constrain architectural choices, but it leads to a clear perform
ance drop when testing at resolutions significantly higher than that used for tr
aining, thus requiring ad-hoc and slow post-processing schemes. In this paper, w
e propose a novel strategy for efficient training and inference of high-resoluti
on vision transformers. The key principle is to mask out most of the high-resolu
tion inputs during training, keeping only N random windows. This allows the mode
l to learn local interactions between tokens inside each window, and global inte
ractions between tokens from different windows. As a result, the model can direc
tly process the high-resolution input at test time without any special trick. We
 show that this strategy is effective when using relative positional embedding s
uch as rotary embeddings. It is 4 times faster to train than a full-resolution n
etwork, and it is straightforward to use at test time compared to existing appro
aches. We apply this strategy to three dense prediction tasks with high-resoluti
on data. First, we show on the task of semantic segmentation that a simple setti
ng with 2 windows performs best, hence the name of our method: Win-Win. Second,
we confirm this result on the task of monocular depth prediction. Third, to demo
nstrate the generality of our contribution, we further extend it to the binocula
r task of optical flow, reaching state-of-the-art performance on the Spring benc
hmark that contains Full-HD images with an order of magnitude faster inference t
han the best competitor
****
Hyunwook Lee,Sungahn Ko
TESTAM: A Time-Enhanced Spatio-Temporal Attention Model with Mixture of Experts
Accurate traffic forecasting is challenging due to the complex dependency on roa
d networks, various types of roads, and the abrupt speed change due to the event
s. Recent works mainly focus on dynamic spatial modeling with adaptive graph emb
edding or graph attention having less consideration for temporal characteristics
 and in-situ modeling. In this paper, we propose a novel deep learning model nam
ed TESTAM, which individually models recurring and non-recurring traffic pattern
s by a mixture-of-experts model with three experts on temporal modeling, spatio-
temporal modeling with static graph, and dynamic spatio-temporal dependency mode
ling with dynamic graph. By introducing different experts and properly routing t
hem, TESTAM could better model various circumstances, including spatially isolat
ed nodes, highly related nodes, and recurring and non-recurring events. For the
proper routing, we reformulate a gating problem into a classification problem wi
th pseudo labels. Experimental results on three public traffic network datasets,
 METR-LA, PEMS-BAY, and EXPY-TKY, demonstrate that TESTAM achieves a better indi
cation and modeling of recurring and non-recurring traffic.
****
Yuzhou Gu,Zhao Song,Junze Yin,Lichen Zhang
Low Rank Matrix Completion via Robust Alternating Minimization in Nearly Linear
Time
Given a matrix $M\in \mathbb{R}^{m\times n}$, the low rank matrix completion pro
blem asks us to find a rank-k approximation of M as $UV^\top$ for $U\in \mat
hbb{R}^{m\times k}$ and $V\in \mathbb{R}^{n\times k}$ by only observing a few en
tries specified by a set of entries $\Omega\subseteq [m]\times [n]$. In particul
ar, we examine an approach that is widely used in practice --- the alternating m
inimization framework. Jain, Netrapalli and Sanghavi showed that if M has inco
herent rows and columns, then alternating minimization provably recovers the mat
rix M by observing a nearly linear in n number of entries. While the sample
complexity has been subsequently improved, alternating minimization steps are re
quired to be computed exactly. This hinders the development of more efficient al
gorithms and fails to depict the practical implementation of alternating minimiz
ation, where the updates are usually performed approximately in favor of efficie
ncy.

In this paper, we take a major step towards a more efficient and error-robust al
ternating minimization framework. To this end, we develop an analytical framewor

k for alternating minimization that can tolerate a moderate amount of errors caused by approximate updates. Moreover, our algorithm runs in time $\widetilde O(|\Omega| k)$, which is nearly linear in the time to verify the solution while preserving the sample complexity. This improves upon all prior known alternating minimization approaches which require $\widetilde O(|\Omega| k^2)$ time.
****
Juan Rocamonde,Victoriano Montesinos,Elvis Nava,Ethan Perez,David Lindner
Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning
Reinforcement learning (RL) requires either manually specifying a reward function, which is often infeasible, or learning a reward model from a large amount of human feedback, which is often very expensive. We study a more sample-efficient alternative: using pretrained vision-language models (VLMs) as zero-shot reward models (RMs) to specify tasks via natural language. We propose a natural and general approach to using VLMs as reward models, which we call VLM-RMs. We use VLM-RMs based on CLIP to train a MuJoCo humanoid to learn complex tasks without a manually specified reward function, such as kneeling, doing the splits, and sitting in a lotus position. For each of these tasks, we only provide _a single sentence text prompt_ describing the desired task with minimal prompt engineering. We provide videos of the trained agents at: https://sites.google.com/view/vlm-rm. We can improve performance by providing a second "baseline" prompt and projecting out parts of the CLIP embedding space irrelevant to distinguish between goal and baseline. Further, we find a strong scaling effect for VLM-RMs: larger VLMs trained with more compute and data are better reward models. The failure modes of VLM-RMs we encountered are all related to known capability limitations of current VLMs, such as limited spatial reasoning ability or visually unrealistic environments that are far off-distribution for the VLM. We find that VLM-RMs are remarkably robust as long as the VLM is large enough. This suggests that future VLMs will become more and more useful reward models for a wide range of RL applications.
****
Nhu-Thuat Tran,Hady W. Lauw
Learning Multi-Faceted Prototypical User Interests
We seek to uncover the latent interest units from behavioral data to better learn user preferences under the VAE framework. Existing practices tend to ignore the multiple facets of item characteristics, which may not capture it at appropriate granularity. Moreover, current studies equate the granularity of item space to that of user interests, which we postulate is not ideal as user interests would likely map to a small subset of item space. In addition, the compositionality of user interests has received inadequate attention, preventing the modeling of interactions between explanatory factors driving a user's decision.
To resolve this, we propose to align user interests with multi-faceted item characteristics. First, we involve prototype-based representation learning to discover item characteristics along multiple facets. Second, we compose user interests from uncovered item characteristics via binding mechanism, separating the granularity of user preferences from that of item space. Third, we design a dedicated bi-directional binding block, aiding the derivation of compositional user interests.
On real-world datasets, the experimental results demonstrate the strong performance of our proposed method compared to a series of baselines.
****
Ravi Francesco Srinivasan,Francesca Mignacco,Martino Sorbaro,Maria Refinetti,Avi Cooper,Gabriel Kreiman,Giorgia Dellaferrera
Forward Learning with Top-Down Feedback: Empirical and Analytical Characterization
"Forward-only" algorithms, which train neural networks while avoiding a backward pass, have recently gained attention as a way of solving the biologically unrealistic aspects of backpropagation. Here, we first address compelling challenges related to the "forward-only" rules, which include reducing the performance gap with backpropagation and providing an analytical understanding of their dynamics. To this end, we show that the forward-only algorithm with top-down feedback is

well-approximated by an "adaptive-feedback-alignment" algorithm, and we analytically track its performance during learning in a prototype high-dimensional setting. Then, we compare different versions of forward-only algorithms, focusing on the Forward-Forward and PEPITA frameworks, and we show that they share the same learning principles. Overall, our work unveils the connections between three key neuro-inspired learning rules, providing a link between "forward-only" algorithms, i.e., Forward-Forward and PEPITA, and an approximation of backpropagation, i.e., Feedback Alignment.
****
Xiu-Chuan Li,Kun Zhang,Tongliang Liu
Causal Structure Recovery with Latent Variables under Milder Distributional and Graphical Assumptions
Traditional causal discovery approaches typically assume the absence of latent variables, a simplification that often does not align with real-world situations. Recently, there has been a surge of causal discovery methods that explicitly consider latent variables. While some works aim to reveal causal relations between observed variables in the presence of latent variables, others seek to identify latent variables and recover the causal structure over them. The latter typically entail strong distributional and graphical assumptions, such as the non-Gaussianity, purity, and two-pure-children assumption. In this paper, we endeavor to recover the whole causal structure involving both latent and observed variables under milder assumptions. We formulate two cases, one allows entirely arbitrary distribution and requires only one pure child per latent variable, and the other requires no pure child and imposes the non-Gaussianity requirement on only a subset of variables, and they both avoid the purity assumption. We prove the identifiability of linear latent variable models in both cases, and our constructive proof leads to theoretically sound and computationally efficient algorithms.
****
Jonathan Scott,Hossein Zakerinia,Christoph H Lampert
PeFLL: Personalized Federated Learning by Learning to Learn
We present PeFLL, a new personalized federated learning algorithm that improves over the state-of-the-art in three aspects: 1) it produces more accurate models, especially in the low-data regime, and not only for clients present during its training phase, but also for any that may emerge in the future; 2) it reduces the amount of on-client computation and client-server communication by providing future clients with ready-to-use personalized models that require no additional finetuning or optimization; 3) it comes with theoretical guarantees that establish generalization from the observed clients to future ones.
At the core of PeFLL lies a learning-to-learn approach that jointly trains an embedding network and a hypernetwork. The embedding network is used to represent clients in a latent descriptor space in a way that reflects their similarity to each other. The hypernetwork takes as input such descriptors and outputs the parameters of fully personalized client models. In combination, both networks constitute a learning algorithm that achieves state-of-the-art performance in several personalized federated learning benchmarks.
****
Sadegh Mahdavi,Renjie Liao,Christos Thrampoulidis
Memorization Capacity of Multi-Head Attention in Transformers
Transformers have become the go-to architecture for language and vision tasks, yet their theoretical properties, especially memorization capacity, remain elusive. This paper investigates the memorization abilities of multi-head attention mechanisms, examining how many example sequences they can memorize, as a function of the number of heads and sequence length. Motivated by experimental findings on vision transformers, we introduce novel assumptions about the linear independence of input data, distinct from the commonly used general-position assumption. Under these assumptions, we demonstrate that an attention layer with H heads, dimension d, and context size $n < d,$ featuring $\Theta(Hd^2)$ parameters, can memorize $\Omega(Hn)$ examples. Our analysis sheds light on how different attention heads handle various example sequences, aided by the softmax operator’s saturation property. We validate our findings through experiments on synthetic dat

a.
****
Suning Huang,Boyuan Chen,Huazhe Xu,Vincent Sitzmann
DittoGym: Learning to Control Soft Shape-Shifting Robots
Robot co-design, where the morphology of a robot is optimized jointly with a learned policy to solve a specific task, is an emerging area of research. It holds particular promise for soft robots, which are amenable to novel manufacturing techniques that can realize learned morphologies and actuators. Inspired by nature and recent novel robot designs, we propose to go a step further and explore the novel reconfigurable robots, defined as robots that can change their morphology within their lifetime. We formalize control of reconfigurable soft robots as a high-dimensional reinforcement learning (RL) problem. We unify morphology change, locomotion, and environment interaction in the same action space, and introduce an appropriate, coarse-to-fine curriculum that enables us to discover policies that accomplish fine-grained control of the resulting robots. We also introduce DittoGym, a comprehensive RL benchmark for reconfigurable soft robots that require fine-grained morphology changes to accomplish the tasks. Finally, we evaluate our proposed coarse-to-fine algorithm on DittoGym, and demonstrate robots that learn to change their morphology several times within a sequence, uniquely enabled by our RL algorithm. More results are available at https://dittogym.github.io.
****
Maximilian Seitzer,Sjoerd van Steenkiste,Thomas Kipf,Klaus Greff,Mehdi S. M. Sajjadi
DyST: Towards Dynamic Neural Scene Representations on Real-World Videos
Visual understanding of the world goes beyond the semantics and flat structure of individual images. In this work, we aim to capture both the 3D structure and dynamics of real-world scenes from monocular real-world videos. Our Dynamic Scene Transformer (DyST) model leverages recent work in neural scene representation to learn a latent decomposition of monocular real-world videos into scene content, per-view scene dynamics, and camera pose. This separation is achieved through a novel co-training scheme on monocular videos and our new synthetic dataset DySO. DyST learns tangible latent representations for dynamic scenes that enable view generation with separate control over the camera and the content of the scene.
****
Stefano B. Blumberg,Paddy J. Slator,Daniel C. Alexander
Experimental Design for Multi-Channel Imaging via Task-Driven Feature Selection
This paper presents a data-driven, task-specific paradigm for experimental design, to shorten acquisition time, reduce costs, and accelerate the deployment of imaging devices. Current approaches in experimental design focus on model-parameter estimation and require specification of a particular model, whereas in imaging, other tasks may drive the design. Furthermore, such approaches often lead to intractable optimization problems in real-world imaging applications. Here we present a new paradigm for experimental design that simultaneously optimizes the design (set of image channels) and trains a machine-learning model to execute a user-specified image-analysis task. The approach obtains data densely-sampled over the measurement space (many image channels) for a small number of acquisitions, then identifies a subset of channels of prespecified size that best supports the task. We propose a method: TADRED for TAsk-DRiven Experimental Design in imaging, to identify the most informative channel-subset whilst simultaneously training a network to execute the task given the subset. Experiments demonstrate the potential of TADRED in diverse imaging applications: several clinically-relevant tasks in magnetic resonance imaging; and remote sensing and physiological applications of hyperspectral imaging. Results show substantial improvement over classical experimental design, two recent application-specific methods within the new paradigm, and state-of-the-art approaches in supervised feature selection. We anticipate further applications of our approach. Code is available: https://github.com/sbb-gh/experimental-design-multichannel
****

Anne Harrington,Vasha DuTell,Mark Hamilton,Ayush Tewari,Simon Stent,William T. Freeman,Ruth Rosenholtz
COCO-Periph: Bridging the Gap Between Human and Machine Perception in the Periphery
Evaluating deep neural networks (DNNs) as models of human perception has given rich insights into both human visual processing and representational properties of DNNs. We extend this work by analyzing how well DNNs perform compared to humans when constrained by peripheral vision -- which limits human performance on a variety of tasks, but also benefits the visual system significantly. We evaluate this by (1) modifying the Texture Tiling Model (TTM), a well tested model of peripheral vision to be more flexibly used with DNNs, (2) generating a large dataset which we call COCO-Periph that contains images transformed to capture the information available in human peripheral vision, and (3) comparing DNNs to humans at peripheral object detection using a psychophysics experiment. Our results show that common DNNs underperform at object detection compared to humans when simulating peripheral vision with TTM. Training on COCO-Periph begins to reduce the gap between human and DNN performance and leads to small increases in corruption robustness, but DNNs still struggle to capture human-like sensitivity to peripheral clutter. Our work brings us closer to accurately modeling human vision, and paves the way for DNNs to mimic and sometimes benefit from properties of human visual processing.
****
Po-Chen Ko,Jiayuan Mao,Yilun Du,Shao-Hua Sun,Joshua B. Tenenbaum
Learning to Act from Actionless Videos through Dense Correspondences
In this work, we present an approach to construct a video-based robot policy capable of reliably executing diverse tasks across different robots and environments from few video demonstrations without using any action annotations. Our method leverages images as a task-agnostic representation, encoding both the state and action information, and text as a general representation for specifying robot goals. By synthesizing videos that "hallucinate" robot executing actions and in combination with dense correspondences between frames, our approach can infer the closed-formed action to execute to an environment without the need of any explicit action labels. This unique capability allows us to train the policy solely based on RGB videos and deploy learned policies to various robotic tasks. We demonstrate the efficacy of our approach in learning policies on table-top manipulation and navigation tasks. Additionally, we contribute an open-source framework for efficient video modeling, enabling the training of high-fidelity policy models with four GPUs within a single day.
****
Hyunsu Kim,Jongmin Yoon,Juho Lee
Fast Ensembling with Diffusion Schrödinger Bridge
Deep Ensemble approach is a straightforward technique used to enhance the performance of deep neural networks by training them from different initial points, converging towards various local optima. However, a limitation of this methodology lies in its high computational overhead for inference, arising from the necessity to store numerous learned parameters and execute individual forward passes for each parameter during the inference stage.We propose a novel approach called Diffusion Bridge Network to address this challenge. Based on the theory of Schr\"odinger bridge, this method directly learns to simulate an Stochastic Differential Equation (SDE) that connects the output distribution of a single ensemble member to the output distribution of the ensembled model, allowing us to obtain ensemble prediction without having to invoke forward pass through all the ensemble models. By substituting the heavy ensembles with this lightweight neural network constructing DBN, we achieved inference with reduced computational cost while maintaining accuracy and uncertainty scores on benchmark datasets such as CIFAR-10, CIFAR-100, and TinyImageNet.
****
Lingfeng Shen,Sihao Chen,Linfeng Song,Lifeng Jin,Baolin Peng,Haitao Mi,Daniel Khashabi,Dong Yu
The Trickle-down Impact of Reward Inconsistency on RLHF

Standard practice within Reinforcement Learning from Human Feedback (RLHF) involves optimizing against a Reward Model (RM), which itself is trained to reflect human preferences for desirable generations. A notable subject that is understudied is the (in-)consistency of RMs --- whether they can recognize the semantic changes to different prompts and
appropriately adapt their reward assignments

--- and their impact on the downstream RLHF model.

In this paper, we visit a series of research questions relevant to RM inconsistency:
(1) How can we measure the consistency of reward models?
(2) How consistent are the existing RMs and how can we improve them?
(3) In what ways does reward inconsistency influence the chatbots resulting from the RLHF model training?

We propose **Contrast Instruction** -- a benchmarking strategy for the consistency of RM.
Each example in **Contrast Instruction** features a pair of lexically similar instructions with different ground truth responses. A consistent RM is expected to rank the corresponding instruction and response higher than other combinations. We observe that current RMs trained with the standard ranking objective fail miserably on \contrast{} compared to average humans. To show that RM consistency can be improved efficiently without using extra training budget, we propose two techniques **ConvexDA** and **RewardFusion**, which enhance reward consistency through extrapolation during the RM training and inference stage, respectively.
We show that RLHF models trained with a more consistent RM yield more useful responses, suggesting that reward inconsistency exhibits a trickle-down effect on the downstream RLHF process.
****
Mingqing Xiao,Qingyan Meng,Zongpeng Zhang,Di He,Zhouchen Lin
Hebbian Learning based Orthogonal Projection for Continual Learning of Spiking Neural Networks
Neuromorphic computing with spiking neural networks is promising for energy-efficient artificial intelligence (AI) applications. However, different from humans who continually learn different tasks in a lifetime, neural network models suffer from catastrophic forgetting. How could neuronal operations solve this problem is an important question for AI and neuroscience. Many previous studies draw inspiration from observed neuroscience phenomena and propose episodic replay or synaptic metaplasticity, but they are not guaranteed to explicitly preserve knowledge for neuron populations. Other works focus on machine learning methods with more mathematical grounding, e.g., orthogonal projection on high dimensional spaces, but there is no neural correspondence for neuromorphic computing. In this work, we develop a new method with neuronal operations based on lateral connections and Hebbian learning, which can protect knowledge by projecting activity traces of neurons into an orthogonal subspace so that synaptic weight update will not interfere with old tasks. We show that Hebbian and anti-Hebbian learning on recurrent lateral connections can effectively extract the principal subspace of neural activities and enable orthogonal projection. This provides new insights into how neural circuits and Hebbian learning can help continual learning, and also how the concept of orthogonal projection can be realized in neuronal systems. Our method is also flexible to utilize arbitrary training methods based on presynaptic activities/traces. Experiments show that our method consistently solves forgetting for spiking neural networks with nearly zero forgetting under various supervised training methods with different error propagation approaches, and outperforms previous approaches under various settings. Our method can pave a solid path for building continual neuromorphic computing systems. The code is available at https://github.com/pkuxmq/HLOP-SNN.
****

Chang Liu,Zhichen Dong,Haobo Ma,Weilin Luo,Xijun Li,Bowen Pang,Jia Zeng,Junchi Y
an
L2P-MIP: Learning to Presolve for Mixed Integer Programming
Modern solvers for solving mixed integer programming (MIP) often rely on the bra
nch-and-bound (B&B) algorithm which could be of high time complexity, and presol
ving techniques are well designed to simplify the instance as pre-processing bef
ore B&B. However, such presolvers in existing literature or open-source solvers
are mostly set by default agnostic to specific input instances, and few studies
have been reported on tailoring presolving settings. In this paper, we aim to di
ve into this open question and show that the MIP solver can be indeed largely im
proved when switching the default instance-agnostic presolving into instance-spe
cific presolving. Specifically, we propose a combination of supervised learning
and classic heuristics to achieve efficient presolving adjusting, avoiding tedio
us reinforcement learning. Notably, our approach is orthogonal from many recent
efforts in incorporating learning modules into the B&B framework after the preso
lving stage, and to our best knowledge, this is the first work for introducing l
earning to presolve in MIP solvers. Experiments on multiple real-world datasets
show that well-trained neural networks can infer proper presolving for arbitrary
 incoming MIP instances in less than 0.5s, which is neglectable compared with th
e solving time often hours or days.
****
Youliang Yuan,Wenxiang Jiao,Wenxuan Wang,Jen-tse Huang,Pinjia He,Shuming Shi,Zha
openg Tu
GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher
Safety lies at the core of the development of Large Language Models (LLMs). Ther
e is ample work on aligning LLMs with human ethics and preferences, including da
ta filtering in pretraining, supervised fine-tuning, reinforcement learning from
 human feedback, red teaming, etc. In this study, we discover that chat in ciphe
r can bypass the safety alignment techniques of LLMs, which are mainly conducted
 in natural languages. We propose a novel framework CipherChat to systematically
 examine the generalizability of safety alignment to non-natural languages -- ci
phers. CipherChat enables humans to chat with LLMs through cipher prompts topped
 with system role descriptions and few-shot enciphered demonstrations. We use Ci
pherChat to assess state-of-the-art LLMs, including ChatGPT and GPT-4 for differ
ent representative human ciphers across 11 safety domains in both English and Ch
inese. Experimental results show that certain ciphers succeed almost 100% of the
 time in bypassing the safety alignment of GPT-4 in several safety domains, demo
nstrating the necessity of developing safety alignment for non-natural languages
. Notably, we identify that LLMs seem to have a ''secret cipher'', and propose a
 novel SelfCipher that uses only role play and several unsafe demonstrations in
natural language to evoke this capability. SelfCipher surprisingly outperforms e
xisting human ciphers in almost all cases.
****
Byeongjun Park,Sangmin Woo,Hyojun Go,Jin-Young Kim,Changick Kim
Denoising Task Routing for Diffusion Models
Diffusion models generate highly realistic images by learning a multi-step denoi
sing process, naturally embodying the principles of multi-task learning (MTL). D
espite the inherent connection between diffusion models and MTL, there remains a
n unexplored area in designing neural architectures that explicitly incorporate
MTL into the framework of diffusion models. In this paper, we present Denoising
Task Routing (DTR), a simple add-on strategy for existing diffusion model archit
ectures to establish distinct information pathways for individual tasks within a
 single architecture by selectively activating subsets of channels in the model.
 What makes DTR particularly compelling is its seamless integration of prior kno
wledge of denoising tasks into the framework: (1) Task Affinity: DTR activates s
imilar channels for tasks at adjacent timesteps and shifts activated channels as
 sliding windows through timesteps, capitalizing on the inherent strong affinity
 between tasks at adjacent timesteps. (2) Task Weights: During the early stages
(higher timesteps) of the denoising process, DTR assigns a greater number of tas
k-specific channels, leveraging the insight that diffusion models prioritize rec

onstructing global structure and perceptually rich contents in earlier stages, and focus on simple noise removal in later stages. Our experiments reveal that DTR not only consistently boosts diffusion models' performance across different evaluation protocols without adding extra parameters but also accelerates training convergence. Finally, we show the complementarity between our architectural approach and existing MTL optimization techniques, providing a more complete view of MTL in the context of diffusion training. Significantly, by leveraging this complementarity, we attain matched performance of DiT-XL using the smaller DiT-L with a reduction in training iterations from 7M to 2M. Our project page is available at https://byeongjun-park.github.io/DTR/
****
Samir Khaki,Konstantinos N Plataniotis
The Need for Speed: Pruning Transformers with One Recipe
We introduce the $\textbf{O}$ne-shot $\textbf{P}$runing $\textbf{T}$echnique for $\textbf{I}$nterchangeable $\textbf{N}$etworks ($\textbf{OPTIN}$) framework as a tool to increase the efficiency of pre-trained transformer architectures $\textit{without requiring re-training}$. Recent works have explored improving transformer efficiency, however often incur computationally expensive re-training procedures or depend on architecture-specific characteristics, thus impeding practical wide-scale adoption.
To address these shortcomings, the OPTIN framework leverages intermediate feature distillation, capturing the long-range dependencies of model parameters (coined $\textit{trajectory}$), to produce state-of-the-art results on natural language, image classification, transfer learning, and semantic segmentation tasks $\textit{without re-training}$. Given a FLOP constraint, the OPTIN framework will compress the network while maintaining competitive accuracy performance and improved throughput. Particularly, we show a ≤ 2% accuracy degradation from NLP baselines and a 0.5% improvement from state-of-the-art methods on image classification at competitive FLOPs reductions. We further demonstrate the generalization of tasks and architecture with comparative performance using Mask2Former for semantic segmentation and cnn-style networks. OPTIN presents one of the first one-shot efficient frameworks for compressing transformer architectures that generalizes well across different class domains, in particular: natural language and image-related tasks, without $\textit{re-training}$.
****
Lazar Valkov,Akash Srivastava,Swarat Chaudhuri,Charles Sutton
A Probabilistic Framework for Modular Continual Learning
Modular approaches that use a different composition of modules for each problem are a promising direction in continual learning (CL). However, searching through the large, discrete space of module compositions is challenging, especially because evaluating a composition’s performance requires a round of neural network training. We address this challenge through a modular CL framework, PICLE, that uses a probabilistic model to cheaply compute the fitness of each composition, allowing PICLE to achieve both perceptual, few-shot and latent transfer. The model combines prior knowledge about good module compositions with dataset-specific information. We evaluate PICLE using two benchmark suites designed to assess different desiderata of CL techniques. Comparing to a wide range of approaches, we show that PICLE is the first modular CL algorithm to achieve perceptual, few-shot and latent transfer while scaling well to large search spaces, outperforming previous state-of-the-art modular CL approaches on long problem sequences.
****
Guowei Xu,Ruijie Zheng,Yongyuan Liang,Xiyao Wang,Zhecheng Yuan,Tianying Ji,Yu Luo,Xiaoyu Liu,Jiaxin Yuan,Pu Hua,Shuzhen Li,Yanjie Ze,Hal Daumé III,Furong Huang,Huazhe Xu
DrM: Mastering Visual Reinforcement Learning through Dormant Ratio Minimization
Visual reinforcement learning (RL) has shown promise in continuous control tasks.
Despite its progress, current algorithms are still unsatisfactory in virtually every aspect of the performance such as sample efficiency, asymptotic performance, and their robustness to the choice of random seeds.

In this paper, we identify a major shortcoming in existing visual RL methods tha
t is the agents often exhibit sustained inactivity during early training, thereb
y limiting their ability to explore effectively.
Expanding upon this crucial observation, we additionally unveil a significant co
rrelation between the agents' inclination towards motorically inactive explorati
on and the absence of neuronal activity within their policy networks.
To quantify this inactivity, we adopt dormant ratio as a metric to measure inact
ivity in the RL agent's network.
Empirically, we also recognize that the dormant ratio can act as a standalone in
dicator of an agent's activity level, regardless of the received reward signals.
Leveraging the aforementioned insights, we introduce DrM, a method that uses thr
ee core mechanisms to guide agents' exploration-exploitation trade-offs by activ
ely minimizing the dormant ratio.
Experiments demonstrate that DrM achieves significant improvements in sample ef
ficiency and asymptotic performance with no broken seeds (76 seeds in total) acr
oss three continuous control benchmark environments, including DeepMind Control
Suite, MetaWorld, and Adroit.
Most importantly, DrM is the first model-free algorithm that consistently solves
 tasks in both the Dog and Manipulator domains from the DeepMind Control Suite a
s well as three dexterous hand manipulation tasks without demonstrations in Adro
it, all based on pixel observations.
****
Chengrui Li,Soon Ho Kim,Chris Rodgers,Hannah Choi,Anqi Wu
One-hot Generalized Linear Model for Switching Brain State Discovery
Exposing meaningful and interpretable neural interactions is critical to underst
anding neural circuits. Inferred neural interactions from neural signals primari
ly reflect functional connectivity. In a long experiment, subject animals may ex
perience different stages defined by the experiment, stimuli, or behavioral stat
es, and hence functional connectivity can change over time. To model dynamically
 changing functional connectivity, prior work employs state-switching generalize
d linear models with hidden Markov models (i.e., HMM-GLMs). However, we argue th
ey lack biological plausibility, as functional connectivities are shaped and con
fined by the underlying anatomical connectome. Here, we propose two novel prior-
informed state-switching GLMs, called Gaussian HMM-GLM (Gaussian prior) and one-
hot HMM-GLM (Gumbel-Softmax one-hot prior). We show that the learned prior shoul
d capture the state-invariant interaction, shedding light on the underlying anat
omical connectome and revealing more likely physical neuron interactions. The st
ate-dependent interaction modeled by each GLM offers traceability to capture fun
ctional variations across multiple brain states. Our methods effectively recover
 true interaction structures in simulated data, achieve the highest predictive l
ikelihood, and enhance the interpretability of interaction patterns and hidden s
tates when applied to real neural data. The code is available at \url{https://gi
thub.com/JerrySoybean/onehot-hmmglm}.
****
Yining Li,Peizhong Ju,Ness Shroff
Achieving Sample and Computational Efficient Reinforcement Learning by Action Sp
ace Reduction via Grouping
Reinforcement learning often needs to deal with the exponential growth of states
 and actions when exploring optimal control in high-dimensional spaces (often kn
own as the curse of dimensionality). In this work, we address this issue by lear
ning the inherent structure of action-wise similar MDP to appropriately balance
the performance degradation versus sample/computational complexity. In particul
ar, we partition the action spaces into multiple groups based on the similarity
in transition distribution and reward function, and build a linear decomposition
 model to capture the difference between the intra-group transition kernel and t
he intra-group rewards. Both our theoretical analysis and experiments reveal a *
surprising and counter-intuitive result*: while a more refined grouping strategy
 can reduce the approximation error caused by treating actions in the same group
 as identical, it also leads to increased estimation error when the size of samp
les or the computation resources is limited. This finding highlights the groupin

g strategy as a new degree of freedom that can be optimized to minimize the overall performance loss. To address this issue, we formulate a general optimization problem for determining the optimal grouping strategy, which strikes a balance between performance loss and sample/computational complexity. We further propose a computationally efficient method for selecting a nearly-optimal grouping strategy, which maintains its computational complexity independent of the size of the action space.
****
Ernst Röell,Bastian Rieck
Differentiable Euler Characteristic Transforms for Shape Classification
The _Euler Characteristic Transform_ (ECT) is a powerful invariant, combining geometrical and topological characteristics of shapes and graphs. However, the ECT was hitherto unable to learn task-specific representations. We overcome this issue and develop a novel computational layer that enables learning the ECT in an end-to-end fashion. Our method, the _Differentiable Euler Characteristic Transform_ (DECT) is fast and computationally efficient, while exhibiting performance on a par with more complex models in both graph and point cloud classification tasks. Moreover, we show that this seemingly simple statistic provides the same topological expressivity as more complex topological deep learning layers.
****
Angelica Chen,Ravid Shwartz-Ziv,Kyunghyun Cho,Matthew L Leavitt,Naomi Saphra
Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs
Most interpretability research in NLP focuses on understanding the behavior and features of a fully trained model. However, certain insights into model behavior may only be accessible by observing the trajectory of the training process. We present a case study of syntax acquisition in masked language models (MLMs) that demonstrates how analyzing the evolution of interpretable artifacts throughout training deepens our understanding of emergent behavior. In particular, we study Syntactic Attention Structure (SAS), a naturally emerging property of MLMs wherein specific Transformer heads tend to focus on specific syntactic relations. We identify a brief window in pretraining when models abruptly acquire SAS, concurrent with a steep drop in loss. This breakthrough precipitates the subsequent acquisition of linguistic capabilities. We then examine the causal role of SAS by manipulating SAS during training, and demonstrate that SAS is necessary for the development of grammatical capabilities. We further find that SAS competes with other beneficial traits during training, and that briefly suppressing SAS improves model quality. These findings offer an interpretation of a real-world example of both simplicity bias and breakthrough training dynamics.
****
Jiafei Lyu,Xiaoteng Ma,Le Wan,Runze Liu,Xiu Li,Zongqing Lu
SEABO: A Simple Search-Based Method for Offline Imitation Learning
Offline reinforcement learning (RL) has attracted much attention due to its ability in learning from static offline datasets and eliminating the need of interacting with the environment. Nevertheless, the success of offline RL relies heavily on the offline transitions annotated with reward labels. In practice, we often need to hand-craft the reward function, which is sometimes difficult, labor-intensive, or inefficient. To tackle this challenge, we set our focus on the offline imitation learning (IL) setting, and aim at getting a reward function based on the expert data and unlabeled data. To that end, we propose a simple yet effective search-based offline IL method, tagged SEABO. SEABO allocates a larger reward to the transition that is close to its closest neighbor in the expert demonstration, and a smaller reward otherwise, all in an unsupervised learning manner. Experimental results on a variety of D4RL datasets indicate that SEABO can achieve competitive performance to offline RL algorithms with ground-truth rewards, gi

ven only a single expert trajectory, and can outperform prior reward learning an
d offline IL methods across many tasks. Moreover, we demonstrate that SEABO also
 works well if the expert demonstrations contain only observations. Our code is
publicly available at https://github.com/dmksjfl/SEABO.
****
Zhenfang Chen,Rui Sun,Wenjun Liu,Yining Hong,Chuang Gan
Generative Neuro-Symbolic Visual Reasoning by Growing and Reusing Modules
Recent works have shown that Large Language Models (LLMs) could empower traditio
nal neuro-symbolic models via programming capabilities to translate lan- guages
into module descriptions, thus achieving strong visual reasoning results while m
aintaining the model's transparency and efficiency. However, these mod- els usua
lly exhaustively generate the entire code snippet given each new instance of a t
ask, which is extremely ineffective. On the contrary, human beings grad- ually a
cquire knowledge that can be reused and grow into more profound skills for fast
generalization to new tasks since we are an infant. Inspired by this, we propose
 generative neuro-symbolic visual reasoning by growing and reusing mod- ules. Sp
ecifically, our model consists of three unique stages, module initialization, mo
dule generation, and module execution. First, given a vision-language task, we a
dopt LLMs to examine whether we could reuse and grow over established mod- ules
to handle this new task. If not, we initialize a new module needed by the task a
nd specify the inputs and outputs of this new module. After that, the new module
 is created by querying LLMs to generate corresponding code snippets that match
the requirements. In order to get a better sense of the new module's ability, we
 treat few-shot training examples as test cases to see if our new module could p
ass these cases. If yes, the new module is added to the module library for futur
e reuse. Finally, we evaluate the performance of our model on the testing set by
 executing the parsed programs with the newly made visual modules to get the res
ults. We find the proposed GNSVR model possesses several advantages. First, it p
erforms competitively on standard tasks like visual question answering and refer
ring ex- pression comprehension; Second, the visual modules learned from one tas
k can be seamlessly transferred to new tasks; Last but not least, it is able to
adapt to new visual reasoning tasks by observing a few training examples and reu
sing modules.
****
Yuan Liu,Cheng Lin,Zijiao Zeng,Xiaoxiao Long,Lingjie Liu,Taku Komura,Wenping Wan
g
SyncDreamer: Generating Multiview-consistent Images from a Single-view Image
In this paper, we present a novel diffusion model called SyncDreamer that genera
tes multiview-consistent images from a single-view image. Using pretrained large
-scale 2D diffusion models, recent work Zero123 demonstrates the ability to gene
rate plausible novel views from a single-view image of an object. However, maint
aining consistency in geometry and colors for the generated images remains a cha
llenge. To address this issue, we propose a synchronized multiview diffusion mod
el that models the joint probability distribution of multiview images, enabling
the generation of multiview-consistent images in a single reverse process. SyncD
reamer synchronizes the intermediate states of all the generated images at every
 step of the reverse process through a 3D-aware feature attention mechanism that
 correlates the corresponding features across different views. Experiments show
that SyncDreamer generates images with high consistency across different views,
thus making it well-suited for various 3D generation tasks such as novel-view-sy
nthesis, text-to-3D, and image-to-3D. Project page: https://liuyuan-pal.github.i
o/SyncDreamer/.
****
Yi Wang,Yinan He,Yizhuo Li,Kunchang Li,Jiashuo Yu,Xin Ma,Xinhao Li,Guo Chen,Xiny
uan Chen,Yaohui Wang,Ping Luo,Ziwei Liu,Yali Wang,Limin Wang,Yu Qiao
InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Gen
eration
This paper introduces InternVid, a large-scale video-centric multimodal dataset
that enables learning powerful and transferable video-text representations for m
ultimodal understanding and generation. InternVid contains over 7 million videos

lasting nearly 760K hours, yielding 234M video clips accompanied by detailed de scriptions of total 4.1B words. Our core contribution is to develop a scalable a pproach to autonomously build a high-quality video-text dataset with large langu age models (LLM), thereby showcasing its efficacy in learning video-language rep resentation at scale. Specifically, we utilize a multi-scale approach to generat e video-related descriptions. Furthermore, we introduce ViCLIP, a video-text rep resentation learning model based on ViT-L. Learned on InternVid via contrastive learning, this model demonstrates leading zero-shot action recognition and compe titive video retrieval performance. Beyond basic video understanding tasks like recognition and retrieval, our dataset and model have broad applications. They a re particularly beneficial for generating interleaved video-text data for learni ng a video-centric dialogue system, advancing video-to-text and text-to-video ge neration research. These proposed resources provide a tool for researchers and p ractitioners interested in multimodal video understanding and generation.
****

Xue Wang,Tian Zhou,Qingsong Wen,Jinyang Gao,Bolin Ding,Rong Jin
CARD: Channel Aligned Robust Blend Transformer for Time Series Forecasting
Recent studies have demonstrated the great power of Transformer models for time series forecasting. One of the key elements that lead to the transformer's succe ss is the channel-independent (CI) strategy to improve the training robustness. However, the ignorance of the correlation among different channels in CI would limit the model's forecasting capacity. In this work, we design a special Transf ormer, i.e., **C**hannel **A**ligned **R**obust Blen**d** Transformer (CARD for short), that addresses key shortcomings of CI type Transformer in time series f orecasting. First, CARD introduces a channel-aligned attention structure that al lows it to capture both temporal correlations among signals and dynamical depend ence among multiple variables over time. Second, in order to efficiently utilize the multi-scale knowledge, we design a token blend module to generate tokens wi th different resolutions. Third, we introduce a robust loss function for time se ries forecasting to alleviate the potential overfitting issue. This new loss fun ction weights the importance of forecasting over a finite horizon based on predi ction uncertainties. Our evaluation of multiple long-term and short-term forecas ting datasets demonstrates that CARD significantly outperforms state-of-the-art time series forecasting methods. The code is available at the following anonymou s repository: https://anonymous.4open.science/r/CARD-6EEC
****

Matthew Thomas Jackson,Chris Lu,Louis Kirsch,Robert Tjarko Lange,Shimon Whiteson ,Jakob Nicolaus Foerster
Discovering Temporally-Aware Reinforcement Learning Algorithms
Recent advancements in meta-learning have enabled the automatic discovery of nov el reinforcement learning algorithms parameterized by surrogate objective functi ons. To improve upon manually designed algorithms, the parameterization of this learned objective function must be expressive enough to represent novel principl es of learning (instead of merely recovering already established ones) while sti ll generalizing to a wide range of settings outside of its meta-training distrib ution. However, existing methods focus on discovering objective functions that, like many widely used objective functions in reinforcement learning, do not take into account the total number of steps allowed for training, or “training horiz on”. In contrast, humans use a plethora of different learning objectives across the course of acquiring a new ability. For instance, students may alter their st udying techniques based on the proximity to exam deadlines and their self-assess ed capabilities. This paper contends that ignoring the optimization time horizon significantly restricts the expressive potential of discovered learning algorit hms. We propose a simple augmentation to two existing objective discovery approa ches that allows the discovered algorithm to dynamically update its objective fu nction throughout the agent’s training procedure, resulting in expressive schedu les and increased generalization across different training horizons. In the proc ess, we find that commonly used meta-gradient approaches fail to discover such a daptive objective functions while evolution strategies discover highly dynamic l earning rules. We demonstrate the effectiveness of our approach on a wide range

of tasks and analyze the resulting learned algorithms, which we find effectively balance exploration and exploitation by modifying the structure of their learning rules throughout the agent’s lifetime.
****

Kaixiang Zheng,EN-HUI YANG
Knowledge Distillation Based on Transformed Teacher Matching
As a technique to bridge logit matching and probability distribution matching, temperature scaling plays a pivotal role in knowledge distillation (KD). Conventionally, temperature scaling is applied to both teacher's logits and student's logits in KD. Motivated by some recent works, in this paper, we drop instead temperature scaling on the student side, and systematically study the resulting variant of KD, dubbed transformed teacher matching (TTM). By reinterpreting temperature scaling as a power transform of probability distribution, we show that in comparison with the original KD, TTM has an inherent Rényi entropy term in its objective function, which serves as an extra regularization term. Extensive experiment results demonstrate that thanks to this inherent regularization, TTM leads to trained students with better generalization than the original KD. To further enhance student's capability to match teacher's power transformed probability distribution, we introduce a sample-adaptive weighting coefficient into TTM, yielding a novel distillation approach dubbed weighted TTM (WTTM). It is shown, by comprehensive experiments, that although WTTM is simple, it is effective, improves upon TTM, and achieves state-of-the-art accuracy performance. Our source code is available at https://github.com/zkxufo/TTM.
****

Ameya Daigavane,Song Eun Kim,Mario Geiger,Tess Smidt
Symphony: Symmetry-Equivariant Point-Centered Spherical Harmonics for Molecule Generation
We present Symphony, an $E(3)$ equivariant autoregressive generative model for 3D molecular geometries
that iteratively builds a molecule from molecular fragments.
Existing autoregressive models such as G-SchNet and G-SphereNet for molecules utilize rotationally invariant features to respect the 3D symmetries of molecules.
In contrast, Symphony uses message-passing with higher-degree $E(3)$-equivariant features.
This allows a novel representation of probability distributions via spherical harmonic signals to efficiently model the 3D geometry of
molecules. We show that Symphony is able to accurately generate small molecules from the QM9 dataset, outperforming existing
autoregressive models and approaching the performance of diffusion models.
****

Gerard Ben Arous,Reza Gheissari,Jiaoyang Huang,Aukosh Jagannath
High-dimensional SGD aligns with emerging outlier eigenspaces
We rigorously study the joint evolution of training dynamics via stochastic gradient descent (SGD) and the spectra of empirical Hessian and gradient matrices. We prove that in two canonical classification tasks for multi-class high-dimensional mixtures and either 1 or 2-layer neural networks, the SGD trajectory rapidly aligns with emerging low-rank outlier eigenspaces of the Hessian and gradient matrices. Moreover, in multi-layer settings this alignment occurs per layer, with the final layer's outlier eigenspace evolving over the course of training, and exhibiting rank deficiency when the SGD converges to sub-optimal classifiers. This establishes some of the rich predictions that have arisen from extensive numerical studies in the last decade about the spectra of Hessian and information matrices over the course of training in overparametrized networks.
****

Thomas Soares Mullen,Marine Schimel,Guillaume Hennequin,Christian K. Machens,Michael Orger,Adrien Jouary
Learning interpretable control inputs and dynamics underlying animal locomotion
A central objective in neuroscience is to understand how the brain orchestrates movement. Recent advances in automated tracking technologies have made it possib

le to document behavior with unprecedented temporal resolution and scale, genera
ting rich datasets which can be exploited to gain insights into the neural contr
ol of movement. One common approach is to identify stereotypical motor primitive
s using cluster analysis. However, this categorical description can limit our ab
ility to model the effect of more continuous control schemes. Here we take a con
trol theoretic approach to behavioral modeling and argue that movements can be u
nderstood as the output of a controlled dynamical system. Previously, models of
movement dynamics, trained solely on behavioral data, have been effective in rep
roducing observed features of neural activity. These models addressed specific s
cenarios where animals were trained to execute particular movements upon receivi
ng a prompt. In this study, we extend this approach to analyze the full natural
locomotor repertoire of an animal: the zebrafish larva. Our findings demonstrate
 that this repertoire can be effectively generated through a sparse control sign
al driving a latent Recurrent Neural Network (RNN). Our model's learned latent s
pace preserves key kinematic features and disentangles different categories of m
ovements. To further interpret the latent dynamics, we used balanced model reduc
tion to yield a simplified model. Collectively, our methods serve as a case stud
y for interpretable system identification, and offer a novel framework for under
standing neural activity in relation to movement.
****
Vijaya Raghavan T Ramkumar,Bahram Zonooz,Elahe Arani
The Effectiveness of Random Forgetting for Robust Generalization
Deep neural networks are susceptible to adversarial attacks, which can compromis
e their performance and accuracy. Adversarial Training (AT) has emerged as a pop
ular approach for protecting neural networks against such attacks. However, a ke
y challenge of AT is robust overfitting, where the network's robust performance
on test data deteriorates with further training, thus hindering generalization.
Motivated by the concept of active forgetting in the brain, we introduce a novel
 learning paradigm called "Forget to Mitigate Overfitting (FOMO)". FOMO alternat
es between the forgetting phase, which randomly forgets a subset of weights and
regulates the model's information through weight reinitialization, and the relea
rning phase, which emphasizes learning generalizable features. Our experiments o
n benchmark datasets and adversarial attacks show that FOMO alleviates robust ov
erfitting by significantly reducing the gap between the best and last robust tes
t accuracy while improving the state-of-the-art robustness. Furthermore, FOMO pr
ovides a better trade-off between the standard and robust accuracy outperforming
 baseline adversarial methods. Finally, our framework is robust to AutoAttacks a
nd increases generalization in many real-world scenarios.
****
Sinong Geng,Aldo Pacchiano,Andrey Kolobov,Ching-An Cheng
Improving Offline RL by Blending Heuristics
We propose **H**e**u**ristic **Bl**ending (HUBL), a simple performance-improving
 technique for a broad class of offline RL algorithms based on value bootstrappi
ng. HUBL modifies the Bellman operators used in these algorithms, partially repl
acing the bootstrapped values with heuristic ones that are estimated with Monte-
Carlo returns. For trajectories with higher returns, HUBL relies more on the heu
ristic values and less on bootstrapping; otherwise, it leans more heavily on boo
tstrapping. HUBL is very easy to combine with many existing offline RL implement
ations by relabeling the offline datasets with adjusted rewards and discount fac
tors. We derive a theory that explains HUBL's effect on offline RL as reducing o
ffline RL's complexity and thus increasing its finite-sample performance. Furth
ermore, we empirically demonstrate that HUBL consistently improves the policy qu
ality of four state-of-the-art bootstrapping-based offline RL algorithms (ATAC,
CQL, TD3+BC, and IQL), by 9% on average over 27 datasets of the D4RL and Meta-Wo
rld benchmarks.
****
Yang Deng,Wenxuan Zhang,Wai Lam,See-Kiong Ng,Tat-Seng Chua
Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents
Proactive dialogues serve as a practical yet challenging dialogue problem in the
 era of large language models (LLMs), where the dialogue policy planning is the

key to improving the proactivity of LLMs. Most existing studies enable the dialo
gue policy planning of LLMs using various prompting schemes or iteratively enhan
ce this capability in handling the given case with verbal AI feedback. However,
these approaches are either bounded by the policy planning capability of the fro
zen LLMs or hard to be transferred to new cases. In this work, we introduce a ne
w dialogue policy planning paradigm to strategize LLMs for proactive dialogue pr
oblems with a tunable language model plug-in as a plug-and-play dialogue policy
planner, named PPDPP. Specifically, we develop a novel training framework to fac
ilitate supervised fine-tuning over available human-annotated data as well as re
inforcement learning from goal-oriented AI feedback with dynamic interaction dat
a collected by the LLM-based self-play simulation. In this manner, the LLM-power
ed dialogue agent can not only be generalized to different cases after the train
ing, but also be applicable to different applications by just substituting the l
earned plug-in. In addition, we propose to evaluate the policy planning capabili
ty of dialogue systems under the interactive setting. Experimental results demon
strate that PPDPP consistently and substantially outperforms existing approaches
 on three different proactive dialogue applications, including negotiation, emot
ional support, and tutoring dialogues.
****
Jake Grigsby,Linxi Fan,Yuke Zhu
AMAGO: Scalable In-Context Reinforcement Learning for Adaptive Agents
We introduce AMAGO, an in-context Reinforcement Learning (RL) agent that uses se
quence models to tackle the challenges of generalization, long-term memory, and
meta-learning. Recent works have shown that off-policy learning can make in-cont
ext RL with recurrent policies viable. Nonetheless, these approaches require ext
ensive tuning and limit scalability by creating key bottlenecks in agents' memor
y capacity, planning horizon, and model size. AMAGO revisits and redesigns the o
ff-policy in-context approach to successfully train long-sequence Transformers o
ver entire rollouts in parallel with end-to-end RL. Our agent is scalable and ap
plicable to a wide range of problems, and we demonstrate its strong performance
empirically in meta-RL and long-term memory domains. AMAGO's focus on sparse rew
ards and off-policy data also allows in-context learning to extend to goal-condi
tioned problems with challenging exploration. When combined with a multi-goal hi
ndsight relabeling scheme, AMAGO can solve a previously difficult category of op
en-world domains, where agents complete many possible instructions in procedural
ly generated environments.
****
Zhixuan Lin,Pierluca D'Oro,Evgenii Nikishin,Aaron Courville
The Curse of Diversity in Ensemble-Based Exploration
We uncover a surprising phenomenon in deep reinforcement learning: training a di
verse ensemble of data-sharing agents -- a well-established exploration strategy
 -- can significantly impair the performance of the individual ensemble members
when compared to standard single-agent training. Through careful analysis, we at
tribute the degradation in performance to the low proportion of self-generated d
ata in the shared training data for each ensemble member, as well as the ineffic
iency of the individual ensemble members to learn from such highly off-policy da
ta. We thus name this phenomenon *the curse of diversity*. We find that several
intuitive solutions -- such as a larger replay buffer or a smaller ensemble size
 -- either fail to consistently mitigate the performance loss or undermine the a
dvantages of ensembling. Finally, we demonstrate the potential of representation
 learning to counteract the curse of diversity with a novel method named Cross-E
nsemble Representation Learning (CERL) in both discrete and continuous control d
omains. Our work offers valuable insights into an unexpected pitfall in ensemble
-based exploration and raises important caveats for future applications of simil
ar approaches.
****
Mirco Mutti,Riccardo De Santi,Marcello Restelli,Alexander Marx,Giorgia Ramponi
Exploiting Causal Graph Priors with Posterior Sampling for Reinforcement Learnin
g
Posterior sampling allows exploitation of prior knowledge on the environment's t

ransition dynamics to improve the sample efficiency of reinforcement learning. The prior is typically specified as a class of parametric distributions, the design of which can be cumbersome in practice, often resulting in the choice of uninformative priors. In this work, we propose a novel posterior sampling approach in which the prior is given as a (partial) causal graph over the environment's variables. The latter is often more natural to design, such as listing known causal dependencies between biometric features in a medical treatment study. Specifically, we propose a hierarchical Bayesian procedure, called C-PSRL, simultaneously learning the full causal graph at the higher level and the parameters of the resulting factored dynamics at the lower level. We provide an analysis of the Bayesian regret of C-PSRL that explicitly connects the regret rate with the degree of prior knowledge. Our numerical evaluation conducted in illustrative domains confirms that C-PSRL strongly improves the efficiency of posterior sampling with an uninformative prior while performing close to posterior sampling with the full causal graph.
****
Renjie Pi,Lewei Yao,Jianhua Han,Xiaodan Liang,Wei Zhang,Hang Xu
Ins-DetCLIP: Aligning Detection Model to Follow Human-Language Instruction
This paper introduces Instruction-oriented Object Detection (IOD), a new task that enhances human-computer interaction by enabling object detectors to understand user instructions and locate relevant objects. Unlike traditional open-vocabulary object detection tasks that rely on users providing a list of required category names, IOD requires models to comprehend natural-language instructions, contextual reasoning, and output the name and location of the desired categories. This poses fresh challenges for modern object detection systems. To develop an IOD system, we create a dataset called IOD-Bench, which consists of instruction-guided detections, along with specialized evaluation metrics. We leverage large-scale language models (LLMs) to generate a diverse set of instructions (8k+) based on existing public object detection datasets, covering a wide range of real-world scenarios. As an initial approach to the IOD task, we propose a model called Ins-DetCLIP. It harnesses the extensive knowledge within LLMs to empower the detector with instruction-following capabilities. Specifically, our Ins-DetCLIP employs a visual encoder (i.e., DetCLIP, an open-vocabulary detector) to extract object-level features. These features are then aligned with the input instructions using a cross-modal fusion module integrated into a pre-trained LLM. Experimental results conducted on IOD-Bench demonstrate that our model consistently outperforms baseline methods that directly combine LLMs with detection models. This research aims to pave the way for a more adaptable and versatile interaction paradigm in modern object detection systems, making a significant contribution to the field.
****
Yixiao Li,Yifan Yu,Chen Liang,Nikos Karampatziakis,Pengcheng He,Weizhu Chen,Tuo Zhao
LoftQ: LoRA-Fine-Tuning-aware Quantization for Large Language Models
Quantization is an indispensable technique for serving Large Language Models (LLMs) and has recently found its way into LoRA fine-tuning (Dettmers et al., 2023). In this work we focus on the scenario where quantization and LoRA fine- tuning are applied together on a pre-trained model. In such cases it is common to observe a consistent gap in the performance on downstream tasks between full fine-tuning and quantization plus LoRA fine-tuning approach. In response, we propose LoftQ (LoRA-Fine-Tuning-aware Quantization), a novel quantization framework that simultaneously quantizes an LLM and finds a proper low-rank initialization for LoRA fine-tuning. Such an initialization alleviates the discrep- ancy between the quantized and full-precision model and significantly improves the generalization in downstream tasks. We evaluate our method on natural lan- guage understanding, question answering, summarization, and natural language generation tasks. Experiments show that our method is highly effective and out- performs existing quantization methods, especially in the challenging 2-bit and 2/4-bit mixed precision regimes. We will release our code.
****

Harold Luc Benoit,Liangze Jiang,Andrei Atanov,Oguzhan Fatih Kar,Mattia Rigotti,Amir Zamir
Unraveling the Key Components of OOD Generalization via Diversification
Supervised learning datasets may contain multiple cues that explain the training set equally well, i.e., learning any of them would lead to the correct predictions on the training data. However, many of them can be spurious, i.e., lose their predictive power under a distribution shift and consequently fail to generalize to out-of-distribution (OOD) data. Recently developed "diversification" methods (Lee et al., 2023; Pagliardini et al., 2023) approach this problem by finding multiple diverse hypotheses that rely on different features. This paper aims to study this class of methods and identify the key components contributing to their OOD generalization abilities.

We show that (1) diversification methods are highly sensitive to the distribution of the unlabeled data used for diversification and can underperform significantly when away from a method-specific sweet spot. (2) Diversification alone is insufficient for OOD generalization. The choice of the used learning algorithm, e.g., the model's architecture and pretraining, is crucial. In standard experiments (classification on Waterbirds and Office-Home datasets), using the second-best choice leads to an up to 20\% absolute drop in accuracy. (3) The optimal choice of learning algorithm depends on the unlabeled data and vice versa i.e. they are co-dependent. (4) Finally, we show that, in practice, the above pitfalls cannot be alleviated by increasing the number of diverse hypotheses, the major feature of diversification methods.

These findings provide a clearer understanding of the critical design factors influencing the OOD generalization abilities of diversification methods. They can guide practitioners in how to use the existing methods best and guide researchers in developing new, better ones.
****
Chenguo Lin,Yadong MU
InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior
Comprehending natural language instructions is a charming property for 3D indoor scene synthesis systems. Existing methods directly model object joint distributions and express object relations implicitly within a scene, thereby hindering the controllability of generation. We introduce InstructScene, a novel generative framework that integrates a semantic graph prior and a layout decoder to improve controllability and fidelity for 3D scene synthesis. The proposed semantic graph prior jointly learns scene appearances and layout distributions, exhibiting versatility across various downstream tasks in a zero-shot manner. To facilitate the benchmarking for text-driven 3D scene synthesis, we curate a high-quality dataset of scene-instruction pairs with large language and multimodal models. Extensive experimental results reveal that the proposed method surpasses existing state-of-the-art approaches by a large margin. Thorough ablation studies confirm the efficacy of crucial design components. Project page: https://chenguolin.github.io/projects/InstructScene.
****
Tianyu Li,Peijin Jia,Bangjun Wang,Li Chen,KUN JIANG,Junchi Yan,Hongyang Li
LaneSegNet: Map Learning with Lane Segment Perception for Autonomous Driving
A map, as crucial information for downstream applications of an autonomous driving system, is usually represented in lanelines or centerlines. However, existing literature on map learning primarily focuses on either detecting geometry-based lanelines or perceiving topology relationships of centerlines. Both of these methods ignore the intrinsic relationship of lanelines and centerlines, that lanelines bind centerlines. While simply predicting both types of lane in one model is mutually excluded in learning objective, we advocate lane segment as a new representation that seamlessly incorporates both geometry and topology information. Thus, we introduce LaneSegNet, the first end-to-end mapping network generating lane segments to obtain a complete representation of the road structure. Our alg

orithm features two key modifications. One is a lane attention module to capture pivotal region details within the long-range feature space. Another is an identical initialization strategy for reference points, which enhances the learning of positional priors for lane attention. On the OpenLane-V2 dataset, LaneSegNet outperforms previous counterparts by a substantial gain across three tasks, i.e., map element detection (+4.8 mAP), centerline perception (+6.9 DET$_l$), and the newly defined one, lane segment perception (+5.6 mAP). Furthermore, it obtains a real-time inference speed of 14.7 FPS. Code is accessible at https://github.com/OpenDriveLab/LaneSegNet.
****
Ryan Wong,Necati Cihan Camgoz,Richard Bowden
Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation
Automatic Sign Language Translation requires the integration of both computer vision and natural language processing to effectively bridge the communication gap between sign and spoken languages. However, the deficiency in large-scale training data to support sign language translation means we need to leverage resources from spoken language. We introduce, Sign2GPT, a novel framework for sign language translation that utilizes large-scale pretrained vision and language models via lightweight adapters for gloss-free sign language translation. The lightweight adapters are crucial for sign language translation, due to the constraints imposed by limited dataset sizes and the computational requirements when training with long sign videos.
We also propose a novel pretraining strategy that directs our encoder to learn sign representations from automatically extracted pseudo-glosses without requiring gloss order information or annotations.
We evaluate our approach on two public benchmark sign language translation datasets, namely RWTH-PHOENIX-Weather 2014T and CSL-Daily, and improve on state-of-the-art gloss-free translation performance with a significant margin.
****
Jie Hao,Xiaochuan Gong,Mingrui Liu
Bilevel Optimization under Unbounded Smoothness: A New Algorithm and Convergence Analysis
Bilevel optimization is an important formulation for many machine learning problems, such as meta-learning and hyperparameter optimization. Current bilevel optimization algorithms assume that the gradient of the upper-level function is Lipschitz (i.e., the upper-level function has a bounded smoothness parameter). However, recent studies reveal that certain neural networks such as recurrent neural networks (RNNs) and long-short-term memory networks (LSTMs) exhibit potential unbounded smoothness, rendering conventional bilevel optimization algorithms unsuitable for these neural networks. In this paper, we design a new bilevel optimization algorithm, namely BO-REP, to address this challenge. This algorithm updates the upper-level variable using normalized momentum and incorporates two novel techniques for updating the lower-level variable: \textit{initialization refinement} and \textit{periodic updates}. Specifically, once the upper-level variable is initialized, a subroutine is invoked to obtain a refined estimate of the corresponding optimal lower-level variable, and the lower-level variable is updated only after every specific period instead of each iteration. When the upper-level problem is nonconvex and unbounded smooth, and the lower-level problem is strongly convex, we prove that our algorithm requires $\widetilde{\mathcal{O}}(1/\epsilon^4)$ \footnote{Here $\widetilde{\mathcal{O}}(\cdot)$ compresses logarithmic factors of $1/\epsilon$ and $1/\delta$, where $\delta\in(0,1)$ denotes the failure probability.} iterations to find an ϵ-stationary point in the stochastic setting, where each iteration involves calling a stochastic gradient or Hessian-vector product oracle. Notably, this result matches the state-of-the-art complexity results under the bounded smoothness setting and without mean-squared smoothness of the stochastic gradient, up to logarithmic factors. Our proof relies on novel technical lemmas for the periodically updated lower-level variable, which are of independent interest. Our experiments on hyper-representation learning, hyperparameter optimization, and data hyper-cleaning for text classification

tasks demonstrate the effectiveness of our proposed algorithm. The code is avail
able at [https://github.com/MingruiLiu-ML-Lab/Bilevel-Optimization-under-Unbound
ed-Smoothness](https://github.com/MingruiLiu-ML-Lab/Bilevel-Optimization-under-U
nbounded-Smoothness).
****
Siming Yan,Yuqi Yang,Yu-Xiao Guo,Hao Pan,Peng-Shuai Wang,Xin Tong,Yang Liu,Qixin
g Huang
3D Feature Prediction for Masked-AutoEncoder-Based Point Cloud Pretraining
Masked autoencoders (MAE) have recently been introduced to 3D self-supervised pr
etraining for point clouds due to their great success in NLP and computer vision
. Unlike MAEs used in the image domain, where the pretext task is to restore fea
tures at the masked pixels, such as colors, the existing 3D MAE works reconstruc
t the missing geometry only, i.e, the location of the masked points. In contrast
 to previous studies, we advocate that point location recovery is inessential an
d restoring intrinsic point features is much superior. To this end, we propose t
o ignore point position reconstruction and recover high-order features at masked
 points including surface normals and surface variations, through a novel attent
ion-based decoder which is independent of the encoder design. We validate the ef
fectiveness of our pretext task and decoder design using different encoder struc
tures for 3D training and demonstrate the advantages of our pretrained networks
on various point cloud analysis tasks.
****
Jintang Li,Huizhe Zhang,Ruofan Wu,Zulun Zhu,Baokun Wang,Changhua Meng,Zibin Zhen
g,Liang Chen
A Graph is Worth 1-bit Spikes: When Graph Contrastive Learning Meets Spiking Neu
ral Networks
While contrastive self-supervised learning has become the de-facto learning para
digm for graph neural networks, the pursuit of higher task accuracy requires a l
arger hidden dimensionality to learn informative and discriminative full-precisi
on representations, raising concerns about computation, memory footprint, and en
ergy consumption burden (largely overlooked) for real-world applications. This w
ork explores a promising direction for graph contrastive learning (GCL) with spi
king neural networks (SNNs), which leverage sparse and binary characteristics to
 learn more biologically plausible and compact representations. We propose Spike
GCL, a novel GCL framework to learn binarized 1-bit representations for graphs,
making balanced trade-offs between efficiency and performance. We provide theore
tical guarantees to demonstrate that SpikeGCL has comparable expressiveness with
 its full-precision counterparts. Experimental results demonstrate that, with ne
arly 32x representation storage compression, SpikeGCL is either comparable to or
 outperforms many fancy state-of-the-art supervised and self-supervised methods
across several graph benchmarks.
****
Hyungho Na,Yunkyeong Seo,Il-chul Moon
Efficient Episodic Memory Utilization of Cooperative Multi-Agent Reinforcement L
earning
In cooperative multi-agent reinforcement learning (MARL), agents aim to achieve
a common goal, such as defeating enemies or scoring a goal. Existing MARL algori
thms are effective but still require significant learning time and often get tra
pped in local optima by complex tasks, subsequently failing to discover a goal-r
eaching policy. To address this, we introduce Efficient episodic Memory Utilizat
ion (EMU) for MARL, with two primary objectives: (a) accelerating reinforcement
learning by leveraging semantically coherent memory from an episodic buffer and
(b) selectively promoting desirable transitions to prevent local convergence. To
 achieve (a), EMU incorporates a trainable encoder/decoder structure alongside M
ARL, creating coherent memory embeddings that facilitate exploratory memory reca
ll. To achieve (b), EMU introduces a novel reward structure called episodic ince
ntive based on the desirability of states. This reward improves the TD target in
 Q-learning and acts as an additional incentive for desirable transitions. We pr
ovide theoretical support for the proposed incentive and demonstrate the effecti
veness of EMU compared to conventional episodic control. The proposed method is

evaluated in StarCraft II and Google Research Football, and empirical results in dicate further performance improvement over state-of-the-art methods.
****
Moritz Imfeld,Jacopo Graldi,Marco Giordano,Thomas Hofmann,Sotiris Anagnostidis,Sidak Pal Singh
Transformer Fusion with Optimal Transport
Fusion is a technique for merging multiple independently-trained neural networks in order to combine their capabilities. Past attempts have been restricted to the case of fully-connected, convolutional, and residual networks. This paper presents a systematic approach for fusing two or more transformer-based networks exploiting Optimal Transport to (soft-)align the various architectural components. We flesh out an abstraction for layer alignment, that can generalize to arbitrary architectures -- in principle -- and we apply this to the key ingredients of Transformers such as multi-head self-attention, layer-normalization, and residual connections, and we discuss how to handle them via various ablation studies. Furthermore, our method allows the fusion of models of different sizes (heterogeneous fusion), providing a new and efficient way to compress Transformers. The proposed approach is evaluated on both image classification tasks via Vision Transformer and natural language modeling tasks using BERT. Our approach consistently outperforms vanilla fusion, and, after a surprisingly short finetuning, also outperforms the individual converged parent models.
In our analysis, we uncover intriguing insights about the significant role of soft alignment in the case of Transformers. Our results showcase the potential of fusing multiple Transformers, thus compounding their expertise, in the budding paradigm of model fusion and recombination. Code is available at https://github.com/graldij/transformer-fusion.
****
Linara Adilova,Maksym Andriushchenko,Michael Kamp,Asja Fischer,Martin Jaggi
Layer-wise linear mode connectivity
Averaging neural network parameters is an intuitive method for fusing the knowledge of two independent models. It is most prominently used in federated learning. If models are averaged at the end of training, this can only lead to a good performing model if the loss surface of interest is very particular, i.e., the loss in the midpoint between the two models needs to be sufficiently low. This is impossible to guarantee for the non-convex losses of state-of-the-art networks. For averaging models trained on vastly different datasets, it was proposed to average only the parameters of particular layers or combinations of layers, resulting in better performing models. To get a better understanding of the effect of layer-wise averaging, we analyse the performance of the models that result from averaging single layers, or groups of layers. Based on our empirical and theoretical investigation, we introduce a novel notion of the layer-wise linear connectivity, and show that deep networks do not have layer-wise barriers between them.
****
Kyuyoung Kim,Jongheon Jeong,Minyong An,Mohammad Ghavamzadeh,Krishnamurthy Dj Dvijotham,Jinwoo Shin,Kimin Lee
Confidence-aware Reward Optimization for Fine-tuning Text-to-Image Models
Fine-tuning text-to-image models with reward functions trained on human feedback data has proven effective for aligning model behavior with human intent. However, excessive optimization with such reward models, which serve as mere proxy objectives, can compromise the performance of fine-tuned models, a phenomenon known as reward overoptimization. To investigate this issue in depth, we introduce the Text-Image Alignment Assessment (TIA2) benchmark, which comprises a diverse collection of text prompts, images, and human annotations. Our evaluation of several state-of-the-art reward models on this benchmark reveals their frequent misalignment with human assessment. We empirically demonstrate that overoptimization occurs notably when a poorly aligned reward model is used as the fine-tuning objective. To address this, we propose TextNorm, a simple method that enhances alignment based on a measure of reward model confidence estimated across a set of semantically contrastive text prompts. We demonstrate that incorporating the confidence-calibrated rewards in fine-tuning effectively reduces overoptimization, re

sulting in twice as many wins in human evaluation for text-image alignment compared against the baseline reward models.
****
Jean-Pierre René Falet,Hae Beom Lee,Nikolay Malkin,Chen Sun,Dragos Secrieru,Dinghuai Zhang,Guillaume Lajoie,Yoshua Bengio
Delta-AI: Local objectives for amortized inference in sparse graphical models
We present a new algorithm for amortized inference in sparse probabilistic graphical models (PGMs), which we call Δ-amortized inference (Δ-AI). Our approach is based on the observation that when the sampling of variables in a PGM is seen as a sequence of actions taken by an agent, sparsity of the PGM enables local credit assignment in the agent's policy learning objective. This yields a local constraint that can be turned into a local loss in the style of generative flow networks (GFlowNets) that enables off-policy training but avoids the need to instantiate all the random variables for each parameter update, thus speeding up training considerably. The Δ-AI objective matches the conditional distribution of a variable given its Markov blanket in a tractable learned sampler, which has the structure of a Bayesian network, with the same conditional distribution under the target PGM. As such, the trained sampler recovers marginals and conditional distributions of interest and enables inference of partial subsets of variables. We illustrate Δ-AI's effectiveness for sampling from synthetic PGMs and training latent variable models with sparse factor structure. Code: https://github.com/GFNOrg/Delta-AI.
****
Taehyeon Kim,Joonkee Kim,Gihun Lee,Se-Young Yun
Instructive Decoding: Instruction-Tuned Large Language Models are Self-Refiner from Noisy Instructions
While instruction-tuned language models have demonstrated impressive zero-shot generalization, these models often struggle to generate accurate responses when faced with instructions that fall outside their training set. This paper presents Instructive Decoding (ID), a simple yet effective approach that augments the efficacy of instruction-tuned models. Specifically, ID adjusts the logits for next-token prediction in a contrastive manner, utilizing predictions generated from a manipulated version of the original instruction, referred to as a noisy instruction. This noisy instruction aims to elicit responses that could diverge from the intended instruction yet remain plausible. We conduct experiments across a spectrum of such noisy instructions, ranging from those that insert semantic noise via random words to others like 'opposite' that elicit the deviated responses. Our approach achieves considerable performance gains across various instruction-tuned models and tasks without necessitating any additional parameter updates. Notably, utilizing 'opposite' as the noisy instruction in ID, which shows the maximum divergence from the original instruction, consistently produces the most significant performance gains across multiple models and tasks.
****
Yuandong Tian,Yiping Wang,Zhenyu Zhang,Beidi Chen,Simon Shaolei Du
JoMA: Demystifying Multilayer Transformers via Joint Dynamics of MLP and Attention
We propose Joint MLP/Attention (JoMA) dynamics, a novel mathematical framework to understand the training procedure of multilayer Transformer architectures. This is achieved by integrating out the self-attention layer in Transformers, producing a modified dynamics of MLP layers only. JoMA removes unrealistic assumptions in previous analysis (e.g., lack of residual connection), and predicts that the attention first becomes sparse (to learn salient tokens), then dense (to learn less salient tokens) in the presence of nonlinear activations, while in the linear case, it is consistent with existing works. We leverage JoMA to qualitatively explains how tokens are combined to form hierarchies in multilayer Transformers, when the input tokens are generated by a latent hierarchical generative model. Experiments on models trained from real-world dataset (Wikitext2/Wikitext103) and various pre- trained models (OPT, Pythia) verify our theoretical findings. The code is at https://github.com/facebookresearch/luckmatters/tree/yuandong3.
****

Frank Shih,Faming Liang
Fast Value Tracking for Deep Reinforcement Learning
Reinforcement learning (RL) tackles sequential decision-making problems by creat
ing
agents that interacts with their environment. However, existing algorithms often
 view these problem as
static, focusing on point estimates for model parameters to maximize expected re
wards, neglecting the stochastic dynamics of agent-environment interactions and
the critical role of uncertainty quantification.
Our research leverages the Kalman filtering paradigm to introduce a novel and sc
alable sampling algorithm called Langevinized Kalman Temporal-Difference (LKTD)
for deep reinforcement learning. This algorithm, grounded in Stochastic Gradient
 Markov Chain Monte Carlo (SGMCMC), efficiently draws samples from the posterior
 distribution of deep neural network parameters. Under mild conditions, we prove
 that the posterior samples generated by the LKTD algorithm converge to a statio
nary distribution. This convergence not only enables us to quantify uncertaintie
s associated with the value function and model parameters but also allows us to
monitor these uncertainties during policy updates throughout the training phase.
 The LKTD algorithm paves the way for more robust and adaptable reinforcement le
arning approaches.
****
zhengyao jiang,Yingchen Xu,Nolan Wagener,Yicheng Luo,Michael Janner,Edward Grefe
nstette,Tim Rocktäschel,Yuandong Tian
H-GAP: Humanoid Control with a Generalist Planner
Humanoid control is an important research challenge offering avenues for integra
tion into human-centric infrastructures and enabling physics-driven humanoid ani
mations.
The daunting challenges in this field stem from the difficulty of optimizing in
high-dimensional action spaces and the instability introduced by the bipedal mor
phology of humanoids.
However, the extensive collection of human motion-captured data and the derived
datasets of humanoid trajectories, such as MoCapAct, paves the way to tackle the
se challenges. In this context, we present Humanoid Generalist Autoencoding Plan
ner (H-GAP), a state-action trajectory generative model trained on humanoid traj
ectories derived from human motion-captured data, capable of adeptly handling do
wnstream control tasks with Model Predictive Control (MPC).
For 56 degrees of freedom humanoid, we empirically demonstrate that H-GAP learns
 to represent and generate a wide range of motor behaviors. Further, without any
 learning from online interactions, it can also flexibly transfer these behaviou
rs to solve novel downstream control tasks via planning. Notably, H-GAP excels e
stablished MPC baselines with access to the ground truth model, and is superior
or comparable to offline RL methods trained for individual tasks.
Finally, we do a series of empirical studies on the scaling properties of H-GAP,
 showing the potential for performance gains via additional data but not computi
ng.
****
Simran Arora,Sabri Eyuboglu,Aman Timalsina,Isys Johnson,Michael Poli,James Zou,A
tri Rudra,Christopher Re
Zoology: Measuring and Improving Recall in Efficient Language Models
Attention-free language models that combine gating and convolutions are growing
in popularity due to their efficiency and increasingly competitive performance.
To better understand these architectures, we pretrain a suite of 17 attention an
d gated-convolution language models, finding that SoTA gated-convolution archite
ctures still underperform attention by up to 2.1 perplexity points on the Pile.
In fine-grained analysis, we find 82% of the gap is explained by each model's ab
ility to recall information that is previously mentioned in-context, e.g. "Hakun
a Matata means no worries Hakuna Matata it means no" -> ??. On this task, termed
 "associative recall", we find that attention outperforms gated-convolutions by
a large margin: a 70M parameter attention model outperforms a 1.4 billion parame
ter gated-convolution model on associative recall. This is surprising because pr

ior work shows gated convolutions can perfectly solve synthetic tests for AR capability. To close the gap between synthetics and real language, we develop a new formalization of the task called multi-query associative recall (MQAR) that better reflects actual language. We perform an empirical and theoretical study of MQAR that elucidates differences in the parameter-efficiency of attention and gated-convolution recall. Informed by our analysis, we evaluate simple convolution-attention hybrids and show that hybrids with input-dependent sparse attention patterns can close 97.4% of the gap to attention, while maintaining sub-quadratic scaling. Code is at: https://github.com/HazyResearch/zoology.

****

Weijia Shi,Sewon Min,Maria Lomeli,Chunting Zhou,Margaret Li,Xi Victoria Lin,Noah A. Smith,Luke Zettlemoyer,Wen-tau Yih,Mike Lewis

In-Context Pretraining: Language Modeling Beyond Document Boundaries

Language models are currently trained to predict tokens given document prefixes, enabling them to zero shot long form generation and prompting-style tasks which can be reduced to document completion. We instead present IN-CONTEXT PRETRAINING, a new approach where language models are trained on a sequence of related documents, thereby explicitly encouraging them to read and reason across document boundaries. Our approach builds on the fact that current pipelines train by concatenating random sets of shorter documents to create longer context windows; this improves efficiency even though the prior documents provide no signal for predicting the next document. Given this fact, we can do IN-CONTEXT PRETRAINING by simply changing the document ordering so that each context contains related documents, and directly applying existing pretraining pipelines. However, this document sorting problem is challenging. There are billions of documents and we would like the sort to maximize contextual similarity for every document without repeating any data. To do this, we introduce approximate algorithms for finding related documents with efficient nearest neighbor search and constructing coherent batches with a graph cover algorithm. Our experiments show IN-CONTEXT PRETRAINING offers a scalable and simple approach to significantly enhance LM performance: we see notable improvements in tasks that require more complex contextual reasoning, including in-context learning (+8%), reading comprehension (+15%), faithfulness to previous contexts (+16%), long-context reasoning (+5%), and retrieval augmentation (+9%).

****

Haozhe Jiang,Qiwen Cui,Zhihan Xiong,Maryam Fazel,Simon Shaolei Du

A Black-box Approach for Non-stationary Multi-agent Reinforcement Learning

We investigate learning the equilibria in non-stationary multi-agent systems and address the challenges that differentiate multi-agent learning from single-agent learning. Specifically, we focus on games with bandit feedback, where testing an equilibrium can result in substantial regret even when the gap to be tested is small, and the existence of multiple optimal solutions (equilibria) in stationary games poses extra challenges. To overcome these obstacles, we propose a versatile black-box approach applicable to a broad spectrum of problems, such as general-sum games, potential games, and Markov games, when equipped with appropriate learning and testing oracles for stationary environments. Our algorithms can achieve $\widetilde{O}\left(\Delta^{1/4}T^{3/4}\right)$ regret when the degree of nonstationarity, as measured by total variation Δ, is known, and $\widetilde{O}\left(\Delta^{1/5}T^{4/5}\right)$ regret when Δ is unknown, where T is the number of rounds. Meanwhile, our algorithm inherits the favorable dependence on number of agents from the oracles. As a side contribution that may be independent of interest, we show how to test for various types of equilibria by a black-box reduction to single-agent learning, which includes Nash equilibria, correlated equilibria, and coarse correlated equilibria.

****

Zican Hu,Zongzhang Zhang,Huaxiong Li,Chunlin Chen,Hongyu Ding,Zhi Wang

Attention-Guided Contrastive Role Representations for Multi-agent Reinforcement Learning

Real-world multi-agent tasks usually involve dynamic team composition with the emergence of roles, which should also be a key to efficient cooperation in multi-

agent reinforcement learning (MARL). Drawing inspiration from the correlation be tween roles and agent's behavior patterns, we propose a novel framework of **A** ttention-guided **CO**ntrastive **R**ole representation learning for **M**ARL (* *ACORM**) to promote behavior heterogeneity, knowledge transfer, and skillful co ordination across agents. First, we introduce mutual information maximization to formalize role representation learning, derive a contrastive learning objective , and concisely approximate the distribution of negative pairs. Second, we lever age an attention mechanism to prompt the global state to attend to learned role representations in value decomposition, implicitly guiding agent coordination in a skillful role space to yield more expressive credit assignment. Experiments o n challenging StarCraft II micromanagement and Google research football tasks de monstrate the state-of-the-art performance of our method and its advantages over existing approaches. Our code is available at https://github.com/NJU-RL/ACORM.
****
Caixia Yan,Xiaojun Chang,Zhihui Li,Lina Yao,Minnan Luo,Qinghua Zheng
Masked Distillation Advances Self-Supervised Transformer Architecture Search
Transformer architecture search (TAS) has achieved remarkable progress in automa ting the neural architecture design process of vision transformers. Recent TAS a dvancements have discovered outstanding transformer architectures while saving t remendous labor from human experts. However, it is still cumbersome to deploy th ese methods in real-world applications due to the expensive costs of data labeli ng under the supervised learning paradigm. To this end, this paper proposes a ma sked image modelling (MIM) based self-supervised neural architecture search meth od specifically designed for vision transformers, termed as MaskTAS, which compl etely avoids the expensive costs of data labeling inherited from supervised lear ning. Based on the one-shot NAS framework, MaskTAS requires to train various wei ght-sharing subnets, which can easily diverged without strong supervision in MIM -based self-supervised learning. For this issue, we design the search space of M askTAS as a siamesed teacher-student architecture to distill knowledge from pre-trained networks, allowing for efficient training of the transformer supernet. T o achieve self-supervised transformer architecture search, we further design a n ovel unsupervised evaluation metric for the evolutionary search algorithm, where each candidate of the student branch is rated by measuring its consistency with the larger teacher network. Extensive experiments demonstrate that the searched architectures can achieve state-of-the-art accuracy on CIFAR-10, CIFAR-100, and ImageNet datasets even without using manual labels. Moreover, the proposed Mask TAS can generalize well to various data domains and tasks by searching specializ ed transformer architectures in self-supervised manner.
****
Xuan Zhang,Jacob Helwig,Yuchao Lin,Yaochen Xie,Cong Fu,Stephan Wojtowytsch,Shuiw ang Ji
SineNet: Learning Temporal Dynamics in Time-Dependent Partial Differential Equat ions
We consider using deep neural networks to solve time-dependent partial different ial equations (PDEs), where multi-scale processing is crucial for modeling compl ex, time-evolving dynamics. While the U-Net architecture with skip connections i s commonly used by prior studies to enable multi-scale processing, our analysis shows that the need for features to evolve across layers results in temporally m isaligned features in skip connections, which limits the model’s performance. To address this limitation, we propose SineNet, consisting of multiple sequentiall y connected U-shaped network blocks, referred to as waves. In SineNet, high-reso lution features are evolved progressively through multiple stages, thereby reduc ing the amount of misalignment within each stage. We furthermore analyze the rol e of skip connections in enabling both parallel and sequential processing of mul ti-scale information. Our method is rigorously tested on multiple PDE datasets, including the Navier-Stokes equations and shallow water equations, showcasing th e advantages of our proposed approach over conventional U-Nets with a comparable parameter budget. We further demonstrate that increasing the number of waves in SineNet while maintaining the same number of parameters leads to a monotonicall

y improved performance. The results highlight the effectiveness of SineNet and the potential of our approach in advancing the state-of-the-art in neural PDE solver design. Our code is available as part of AIRS (https://github.com/divelab/AIRS).
****
Liyiming Ke,Yunchu Zhang,Abhay Deshpande,Siddhartha Srinivasa,Abhishek Gupta
CCIL: Continuity-Based Data Augmentation for Corrective Imitation Learning
We present a new technique to enhance the robustness of imitation learning methods by generating corrective data to account for compounding error and disturbances. While existing methods rely on interactive expert labeling, additional offline datasets, or domain-specific invariances, our approach requires minimal additional assumptions beyond expert data. The key insight is to leverage local continuity in the environment dynamics. Our method first constructs a dynamics model from the expert demonstration, enforcing local Lipschitz continuity while skipping the discontinuous regions. In the locally continuous regions, this model allows us to generate corrective labels within the neighborhood of the demonstrations but beyond the actual set of states and actions in the dataset. Training on this augmented data enhances the agent's ability to recover from perturbations and deal with compounding error. We demonstrate the effectiveness of our generated labels through experiments in a variety of robotics domains that have distinct forms of continuity and discontinuity, including classic control, drone flying, high-dimensional navigation, locomotion, and tabletop manipulation.
****
Geyang Guo,Ranchi Zhao,Tianyi Tang,Xin Zhao,Ji-Rong Wen
Beyond Imitation: Leveraging Fine-grained Quality Signals for Alignment
Alignment with human preference is a desired property of large language models (LLMs). Currently, the main alignment approach is based on reinforcement learning from human feedback (RLHF). Despite the effectiveness of RLHF, it is intricate to implement and train, thus recent studies explore how to develop alternative alignment approaches based on supervised fine-tuning (SFT). A major limitation of SFT is that it essentially does imitation learning, which can't fully understand what are the expected behaviors. To address this issue, we propose an improved alignment approach named $\textbf{FIGA}$. Different from prior methods, we incorporate fine-grained (i.e., token or phrase level) quality signals that are derived by contrasting good and bad responses. Our approach has made two major contributions. Firstly, we curate a refined alignment dataset that pairs initial responses and the corresponding revised ones. Secondly, we devise a new loss function can leverage fine-grained quailty signals to instruct the learning of LLMs for alignment. Extensive experiments have demonstrated the effectiveness of our approaches by comparing a number of competitive baselines.
****
Lei You,Hei Victor Cheng
SWAP: Sparse Entropic Wasserstein Regression for Robust Network Pruning
This study addresses the challenge of inaccurate gradients in computing the empirical Fisher Information Matrix during network pruning. We introduce SWAP, a formulation of Entropic Wasserstein regression (EWR) for network pruning, capitalizing on the geometric properties of the optimal transport problem. The “swap” of the commonly used linear regression with the EWR in optimization is analytically demonstrated to offer noise mitigation effects by incorporating neighborhood interpolation across data points with only marginal additional computational cost. The unique strength of SWAP is its intrinsic ability to balance noise reduction and covariance information preservation effectively. Extensive experiments performed on various networks and datasets show comparable performance of SWAP with state-of-the-art (SoTA) network pruning algorithms. Our proposed method outperforms the SoTA when the network size or the target sparsity is large, the gain is even larger with the existence of noisy gradients, possibly from noisy data, analog memory, or adversarial attacks. Notably, our proposed method achieves a gain of 6% improvement in accuracy and 8% improvement in testing loss for MobileNetV1 with less than one-fourth of the network parameters remaining.
****

George Stoica,Daniel Bolya,Jakob Brandt Bjorner,Pratik Ramesh,Taylor Hearn,Judy Hoffman
ZipIt! Merging Models from Different Tasks without Training
Typical deep visual recognition models are capable of performing the one task they were trained on. In this paper, we tackle the extremely difficult problem of combining distinct models with different initializations, each solving a separate task, into one multi-task model without any additional training. Prior work in model merging permutes one model to the space of the other then averages them together. While this works for models trained on the same task, we find that this fails to account for the differences in models trained on disjoint tasks. Thus, we introduce "ZipIt!", a general method for merging two arbitrary models of the same architecture that incorporates two simple strategies. First, in order to account for features that aren't shared between models, we expand the model merging problem to allow for merging features within each model by defining a general "zip" operation. Second, we add support for partially zipping the models up until a specified layer, naturally creating a multi-head model. We find that these two changes combined account for 20-60% improvement over prior work, making it more feasible to merge models trained on disjoint tasks without retraining.
****
Aya Abdelsalam Ismail,Julius Adebayo,Hector Corrada Bravo,Stephen Ra,Kyunghyun Cho
Concept Bottleneck Generative Models
We introduce a generative model with an intrinsically interpretable layer---a concept bottleneck layer---that constrains the model to encode human-understandable concepts. The concept bottleneck layer partitions the generative model into three parts: the pre-concept bottleneck portion, the CB layer, and the post-concept bottleneck portion. To train CB generative models, we complement the traditional task-based loss function for training generative models with a concept loss and an orthogonality loss. The CB layer and these loss terms are model agnostic, which we demonstrate by applying the CB layer to three different families of generative models: generative adversarial networks, variational autoencoders, and diffusion models. On multiple datasets across different types of generative models, steering a generative model, with the CB layer, outperforms all baselines---in some cases, it is \textit{10 times} more effective. In addition, we show how the CB layer can be used to interpret the output of the generative model and debug the model during or post training.
****
Liyuan Mao,Haoran Xu,Weinan Zhang,Xianyuan Zhan
ODICE: Revealing the Mystery of Distribution Correction Estimation via Orthogonal-gradient Update
In this study, we investigate the DIstribution Correction Estimation (DICE) methods, an important line of work in offline reinforcement learning (RL) and imitation learning (IL). DICE-based methods impose state-action-level behavior constraint, which is an ideal choice for offline learning. However, they typically perform much worse than current state-of-the-art (SOTA) methods that solely use action-level behavior constraint. After revisiting DICE-based methods, we find there exist two gradient terms when learning the value function using true-gradient update: forward gradient (taken on the current state) and backward gradient (taken on the next state). Using forward gradient bears a large similarity to many offline RL methods, and thus can be regarded as applying action-level constraint. However, directly adding the backward gradient may degenerate or cancel out its effect if these two gradients have conflicting directions. To resolve this issue, we propose a simple yet effective modification that projects the backward gradient onto the normal plane of the forward gradient, resulting in an orthogonal-gradient update, a new learning rule for DICE-based methods. We conduct thorough theoretical analyses and find that the projected backward gradient brings state-level behavior regularization, which reveals the mystery of DICE-based methods: the value learning objective does try to impose state-action-level constraint, but needs to be used in a corrected way. Through toy examples and extensive experiments on complex offline RL and IL tasks, we demonstrate that DICE-based method

s using orthogonal-gradient updates achieve SOTA performance and great robustnes
s.
****
Tim Franzmeyer,Edith Elkind,Philip Torr,Jakob Nicolaus Foerster,Joao F. Henrique
s
Select to Perfect: Imitating desired behavior from large multi-agent data
AI agents are commonly trained with large datasets of demonstrations of human be
havior.
However, not all behaviors are equally safe or desirable.
Desired characteristics for an AI agent can be expressed by assigning desirabili
ty scores, which we assume are not assigned to individual behaviors but to colle
ctive trajectories.
For example, in a dataset of vehicle interactions, these scores might relate to
the number of incidents that occurred.
We first assess the effect of each individual agent's behavior on the collective
 desirability score, e.g., assessing how likely an agent is to cause incidents.
This allows us to selectively imitate agents with a positive effect, e.g., only
imitating agents that are unlikely to cause incidents.
To enable this, we propose the concept of an agent's \textit{Exchange Value}, wh
ich quantifies an individual agent's contribution to the collective desirability
 score.
The Exchange Value is the expected change in desirability score when substitutin
g the agent for a randomly selected agent.
We propose additional methods for estimating Exchange Values from real-world dat
asets, enabling us to learn desired imitation policies that outperform relevant
baselines. The project website can be found at https://tinyurl.com/select-to-per
fect.
****
Chenmien Tan,Ge Zhang,Jie Fu
Massive Editing for Large Language Models via Meta Learning
While large language models (LLMs) have enabled learning knowledge from the pre-
training corpora, the acquired knowledge may be fundamentally incorrect or outda
ted over time, which necessitates rectifying the knowledge of the language model
 (LM) after the training. A promising approach involves employing a hyper-networ
k to generate parameter shift, whereas existing hyper-networks suffer from infer
ior scalability in synchronous editing operation amount (Hase et al., 2023b; Hua
ng et al., 2023). For instance, Mitchell et al. (2022) mimics gradient accumulat
ion to sum the parameter shifts together, which lacks statistical significance a
nd is prone to cancellation effect. To mitigate the problem, we propose the MAss
ive Language Model Editing Network (MALMEN), which formulates the parameter shif
t aggregation as the least square problem, subsequently updating the LM paramete
r using the normal equation. To accommodate editing multiple facts simultaneousl
y with limited memory budgets, we separate the computation on the hyper-network
and LM, enabling arbitrary batch size on both neural networks. Our method is eva
luated by editing up to thousands of facts on LMs with different architectures,
i.e., BERT-base, GPT-2, and GPT-J (6B), across various knowledge-intensive NLP t
asks, i.e., closed book fact-checking and question answering. Remarkably, MALMEN
 is capable of editing hundreds of times more facts than MEND (Mitchell et al.,
2022) with the identical hyper-network architecture and outperforms editor speci
fically designed for GPT, i.e., MEMIT (Meng et al., 2023).
****
Archiki Prasad,Elias Stengel-Eskin,Mohit Bansal
Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Mod
els
An increasing number of vision-language tasks can be handled with little to no t
raining, i.e., in a zero and few-shot manner, by marrying large language models
(LLMs) to vision encoders, resulting in large vision-language models (LVLMs). Wh
ile this has huge upsides, such as not requiring training data or custom archite
ctures, how an input is presented to an LVLM can have a major impact on zero-sho
t model performance. In particular, inputs phrased in an underspecified way can

result in incorrect answers due to factors like missing visual information, complex implicit reasoning, or linguistic ambiguity. Therefore, adding visually-grounded information to the input as a preemptive clarification should improve model performance by reducing underspecification, e.g., by localizing objects and disambiguating references. Similarly, in the VQA setting, changing the way questions are framed can make them easier for models to answer. To this end, we present **Rep**hrase, **A**ugment and **Re**ason (RepARe), a gradient-free framework that extracts salient details about the image using the underlying LVLM as a captioner and reasoner, in order to propose modifications to the original question. We then use the LVLM’s confidence over a generated answer as an unsupervised scoring function to select the rephrased question most likely to improve zero-shot performance. Focusing on three visual question answering tasks, we show that RepARe can result in a 3.85% (absolute) increase in zero-shot accuracy on VQAv2, 6.41%, and 7.94% points increase on A-OKVQA, and VizWiz respectively. Additionally, we find that using gold answers for oracle question candidate selection achieves a substantial gain in VQA accuracy by up to 14.41%. Through extensive analysis, we demonstrate that outputs from RepARe increase syntactic complexity, and effectively utilize vision-language interaction and the frozen LLM.

****

Han Zhou,Xingchen Wan,Lev Proleev,Diana Mincu,Jilin Chen,Katherine A Heller,Subhrajit Roy

Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering

Prompting and in-context learning (ICL) have become efficient learning paradigms for large language models (LLMs). However, LLMs suffer from prompt brittleness and various bias factors in the prompt, including but not limited to the formatting, the choice verbalizers, and the ICL examples. To address this problem that results in unexpected performance degradation, calibration methods have been developed to mitigate the effects of these biases while recovering LLM performance. In this work, we first conduct a systematic analysis of the existing calibration methods, where we both provide a unified view and reveal the failure cases. Inspired by these analyses, we propose Batch Calibration (BC), a simple yet intuitive method that controls the contextual bias from the batched input, unifies various prior approaches and effectively addresses the aforementioned issues. BC is zero-shot, inference-only, and incurs negligible additional costs. In the few-shot setup, we further extend BC to allow it to learn the contextual bias from labeled data. We validate the effectiveness of BC with PaLM 2-(S, M, L) and CLIP models and demonstrate state-of-the-art performance over previous calibration baselines across more than 10 natural language understanding and image classification tasks.

****

Bo Zhao,Robert M. Gower,Robin Walters,Rose Yu

Improving Convergence and Generalization Using Parameter Symmetries

In many neural networks, different values of the parameters may result in the same loss value. Parameter space symmetries are loss-invariant transformations that change the model parameters. Teleportation applies such transformations to accelerate optimization. However, the exact mechanism behind this algorithm's success is not well understood. In this paper, we show that teleportation not only speeds up optimization in the short-term, but gives overall faster time to convergence. Additionally, teleporting to minima with different curvatures improves generalization, which suggests a connection between the curvature of the minimum and generalization ability. Finally, we show that integrating teleportation into a wide range of optimization algorithms and optimization-based meta-learning improves convergence. Our results showcase the versatility of teleportation and demonstrate the potential of incorporating symmetry in optimization.

****

Arnav Gudibande,Eric Wallace,Charlie Victor Snell,Xinyang Geng,Hao Liu,Pieter Abbeel,Sergey Levine,Dawn Song

The False Promise of Imitating Proprietary Language Models

An emerging method to cheaply improve a weaker language model is to finetune it

on outputs from a stronger model, such as a proprietary system like ChatGPT (e.g
., Alpaca, Self-Instruct, and others). In this work, we critically analyze this
approach of imitating language models. We first finetune a series of LMs that im
itate ChatGPT using varying base model sizes (1.5B--13B), data sources, and imit
ation data amounts (0.3M--150M tokens). We then evaluate the models using crowd
raters and canonical NLP benchmarks. Initially, we were surprised by the output
quality of our imitation models---they appear far better at following instructio
ns, and crowd workers rate their outputs as competitive with ChatGPT. However, w
hen conducting more targeted automatic evaluations, we find that imitation model
s close little to none of the gap from the base LM to ChatGPT on tasks that are
not heavily supported in the imitation data. We show that these performance disc
repancies may slip past human raters because imitation models are adept at mimic
king ChatGPT’s style but not its factuality. Overall, we conclude that while mod
el imitation can be useful for training models to follow instructions and avoid
toxic outputs, it falls short its full promise in many ways. In particular, ther
e exists a substantial capabilities gap between open and closed LMs that we find
 cannot be bridged merely by adding more imitation data. Instead, we find that f
ine-tuning more capable base LMs has a significantly more substantial effect on
closing this gap. In turn, we argue that the higher leverage action for improvin
g open-source models is to tackle the difficult challenge of developing better b
ase LMs, rather than taking the shortcut of imitating proprietary systems.
****
Xun Tang,Michael Shavlovsky,Holakou Rahmanian,Elisa Tardini,Kiran Koshy Thekumpa
rampil,Tesi Xiao,Lexing Ying
Accelerating Sinkhorn algorithm with sparse Newton iterations
Computing the optimal transport distance between statistical distributions is a
fundamental task in machine learning. One remarkable recent advancement is entro
pic regularization and the Sinkhorn algorithm, which utilizes only matrix scalin
g and guarantees an approximated solution with near-linear runtime. Despite the
success of the Sinkhorn algorithm, its runtime may still be slow due to the pote
ntially large number of iterations needed for convergence. To achieve possibly s
uper-exponential convergence, we introduce Sinkhorn-Newton-Sparse (SNS), an exte
nsion to the Sinkhorn algorithm, by introducing early stopping for the matrix sc
aling steps and a second stage featuring a Newton-type subroutine. Adopting the
variational viewpoint that the Sinkhorn algorithm maximizes a concave Lyapunov p
otential, we offer the insight that the Hessian matrix of the potential function
 is approximately sparse. Sparsification of the Hessian results in a fast $O(n^2
)$ per-iteration complexity, the same as the Sinkhorn algorithm. In terms of to
tal iteration count, we observe that the SNS algorithm converges orders of magni
tude faster across a wide range of practical cases, including optimal transporta
tion between empirical distributions and calculating the Wasserstein W_1, W_2
distance of discretized continuous densities. The empirical performance is corro
borated by a rigorous bound on the approximate sparsity of the Hessian matrix.
****
Jinyi Hu,Yuan Yao,Chongyi Wang,SHAN WANG,Yinxu Pan,Qianyu Chen,Tianyu Yu,Hanghao
 Wu,Yue Zhao,Haoye Zhang,Xu Han,Yankai Lin,Jiao Xue,dahai li,Zhiyuan Liu,Maosong
 Sun
Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages
Recently there has been a significant surge in multimodal learning in terms of b
oth image-to-text and text-to-image generation. However, the success is typicall
y limited to English, leaving other languages largely behind. Building a competi
tive counterpart in other languages is highly challenging due to the low-resourc
e nature of non-English multimodal data (i.e., lack of large-scale, high-quality
 image-text data). In this work, we propose MPM, an effective training paradigm
for training large multimodal models in low-resource languages. MPM demonstrates
 that Multilingual language models can Pivot zero-shot Multimodal learning acros
s languages. Specifically, based on a strong multilingual large language model,
multimodal models pretrained on English-only image-text data can well generalize
 to other languages in a (quasi)-zero-shot manner, even surpassing models traine
d on image-text data in native languages. Taking Chinese as a practice of MPM, w

e build large multimodal models VisCPM in image-to-text and text-to-image genera
tion, which achieve state-of-the-art (open-source) performance in Chinese. To fa
cilitate future research, we open-source codes and model weights at https://gith
ub.com/OpenBMB/VisCPM.
****
Xinyun Chen,Maxwell Lin,Nathanael Schärli,Denny Zhou
Teaching Large Language Models to Self-Debug
Large language models (LLMs) have achieved impressive performance on code genera
tion. However, for complex programming tasks, generating the correct solution in
 one go becomes challenging, thus some prior works have designed program repair
approaches to improve code generation performance. In this work, we propose self
-debugging, which teaches a large language model to debug its predicted program.
 In particular, we demonstrate that self-debugging can teach the large language
model to perform rubber duck debugging; i.e., without any human feedback on the
code correctness or error messages, the model is able to identify its mistakes b
y leveraging code execution and explaining the generated code in natural languag
e. Self-debugging achieves the state-of-the-art performance on several code gene
ration benchmarks, including the Spider dataset for text-to-SQL generation, Tran
sCoder for C++-to-Python translation, and MBPP for text-to-Python generation. On
 the Spider benchmark where there are no unit tests to verify the correctness of
 predictions, self-debugging with code explanation consistently improves the bas
eline by 2-3%, and improves the prediction accuracy on problems of the hardest l
evel by 9%. On TransCoder and MBPP where unit tests are available, self-debuggin
g improves the baseline accuracy by up to 12%. Meanwhile, by leveraging feedback
 messages and reusing failed predictions, self-debugging notably improves sample
 efficiency, and can match or outperform baseline models that generate more than
 10$\times$ candidate programs.
****
Yian Wang,Juntian Zheng,Zhehuan Chen,Zhou Xian,Gu Zhang,Chao Liu,Chuang Gan
Thin-Shell Object Manipulations With Differentiable Physics Simulations
In this work, we aim to teach robots to manipulate various thin-shell materials.

Prior works studying thin-shell object manipulation mostly rely on heuristic pol
icies or learn policies from real-world video demonstrations, and only focus on
limited material types and tasks (e.g., cloth unfolding). However, these approac
hes face significant challenges when extended to a wider variety of thin-shell m
aterials and a diverse range of tasks.
On the other hand, while virtual simulations are shown to be effective in divers
e robot skill learning and evaluation, prior thin-shell simulation environments
only support a subset of thin-shell materials, which also limits their supported
 range of tasks.
To fill in this gap, we introduce ThinShellLab - a fully differentiable simulati
on platform tailored for robotic interactions with diverse thin-shell materials
possessing varying material properties, enabling flexible thin-shell manipulatio
n skill learning and evaluation. Building on top of our developed simulation eng
ine, we design a diverse set of manipulation tasks centered around different thi
n-shell objects. Our experiments suggest that manipulating thin-shell objects pr
esents several unique challenges: 1) thin-shell manipulation relies heavily on f
rictional forces due to the objects' co-dimensional nature, 2) the materials bei
ng manipulated are highly sensitive to minimal variations in interaction actions
, and 3) the constant and frequent alteration in contact pairs makes trajectory
optimization methods susceptible to local optima, and neither standard reinforce
ment learning algorithms nor trajectory optimization methods (either gradient-ba
sed or gradient-free) are able to solve the tasks alone. To overcome these chall
enges, we present an optimization scheme that couples sampling-based trajectory
optimization and gradient-based optimization, boosting both learning efficiency
and converged performance across various proposed tasks. In addition, the differ
entiable nature of our platform facilitates a smooth sim-to-real transition. By
tuning simulation parameters with a minimal set of real-world data, we demonstra
te successful deployment of the learned skills to real-robot settings. ThinShel

lLab will be publicly available. Video demonstration and more information can be found on the project website https://thinshelllab.github.io.
****
Raphaël Avalos,Florent Delgrange,Ann Nowe,Guillermo Perez,Diederik M Roijers
The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models
Partially Observable Markov Decision Processes (POMDPs) are used to model environments where the state cannot be perceived, necessitating reasoning based on past observations and actions. However, remembering the full history is generally intractable due to the exponential growth in the history space. Maintaining a probability distribution that models the belief over the current state can be used as a sufficient statistic of the history, but its computation requires access to the model of the environment and is often intractable. While SOTA algorithms use Recurrent Neural Networks to compress the observation-action history aiming to learn a sufficient statistic, they lack guarantees of success and can lead to sub-optimal policies. To overcome this, we propose the Wasserstein Belief Updater, an RL algorithm that learns a latent model of the POMDP and an approximation of the belief update under the assumption that the state is observable during training. Our approach comes with theoretical guarantees on the quality of our approximation ensuring that our latent beliefs allow for learning the optimal value function.
****
S Chandra Mouli,Muhammad Alam,Bruno Ribeiro
MetaPhysiCa: Improving OOD Robustness in Physics-informed Machine Learning
A fundamental challenge in physics-informed machine learning (PIML) is the design of robust PIML methods for out-of-distribution (OOD) forecasting tasks. These OOD tasks require learning-to-learn from observations of the same (ODE) dynamical system with different unknown ODE parameters, and demand accurate forecasts even under out-of-support initial conditions and out-of-support ODE parameters. In this work we propose to improve the OOD robustness of PIML via a meta-learning procedure for causal structure discovery. Using three different OOD tasks, we empirically observe that the proposed approach significantly outperforms existing state-of-the-art PIML and deep learning methods (with $2\times$ to $28\times$ lower OOD errors).
****
Gabriele Corso,Yilun Xu,Valentin De Bortoli,Regina Barzilay,Tommi S. Jaakkola
Particle Guidance: non-I.I.D. Diverse Sampling with Diffusion Models
In light of the widespread success of generative models, a significant amount of research has gone into speeding up their sampling time. However, generative models are often sampled multiple times to obtain a diverse set incurring a cost that is orthogonal to sampling time. We tackle the question of how to improve diversity and sample efficiency by moving beyond the common assumption of independent samples. We propose particle guidance, an extension of diffusion-based generative sampling where a joint-particle time-evolving potential enforces diversity. We analyze theoretically the joint distribution that particle guidance generates, how to learn a potential that achieves optimal diversity, and the connections with methods in other disciplines. Empirically, we test the framework both in the setting of conditional image generation, where we are able to increase diversity without affecting quality, and molecular conformer generation, where we reduce the state-of-the-art median error by 13% on average.
****
Alexander H. Liu,Matthew Le,Apoorv Vyas,Bowen Shi,Andros Tjandra,Wei-Ning Hsu
Generative Pre-training for Speech with Flow Matching
Generative models have gained more and more attention in recent years for their remarkable success in tasks that required estimating and sampling data distribution to generate high-fidelity synthetic data. In speech, text-to-speech synthesis and neural vocoder are good examples where generative models have shined. While generative models have been applied to different applications in speech, there exists no general-purpose generative model that models speech directly. In this work, we take a step toward this direction by showing a single pre-trained gene

rative model can be adapted to different downstream tasks with strong performanc
e. Specifically, we pre-trained a generative model, named SpeechFlow, on 60k hou
rs of untranscribed speech with Flow Matching and masked conditions. Experiment
results show the pre-trained generative model can be fine-tuned with task-specif
ic data to match or surpass existing expert models on speech enhancement, separa
tion, and synthesis. Our work suggested a foundational model for generation task
s in speech can be built with generative pre-training.
****
Xiangchi Yuan,Chunhui Zhang,Yijun Tian,Yanfang Ye,Chuxu Zhang
Mitigating Emergent Robustness Degradation while Scaling Graph Learning
Although graph neural networks have exhibited remarkable performance in various
graph tasks, a significant concern is their vulnerability to adversarial attacks
. Consequently, many defense methods have been proposed to alleviate the deleter
ious effects of adversarial attacks and learn robust graph representations. Howe
ver, most of them are difficult to *simultaneously* avoid two major limitations:
 (i) an emergent and severe degradation in robustness when exposed to very inten
se attacks, and (ii) heavy computation complexity hinders them from scaling to l
arge graphs. In response to these challenges, we introduce an innovative graph d
efense method for unpredictable real-world scenarios by *designing a graph robus
t learning framework that is resistant to robustness degradation* and *refrainin
g from unscalable designs with heavy computation*: specifically, our method empl
oys a denoising module, which eliminates edges that are associated with attacked
 nodes to reconstruct a cleaner graph; Then, it applies Mixture-of-Experts to se
lect differentially private noises with varying magnitudes to counteract the hid
den features attacked at different intensities toward robust predictions; Moreov
er, our overall design avoids the reliance on heavy adjacency matrix computation
s, such as SVD, thus facilitating its applicability even on large graphs. Compre
hensive experiments have been conducted to demonstrate the anti-degraded robustn
ess and scalability of our method, as compared to popular graph adversarial lear
ning methods, under diverse attack intensities and various datasets of different
 sizes.
****
Montgomery Bohde,Meng Liu,Alexandra Saxton,Shuiwang Ji
On the Markov Property of Neural Algorithmic Reasoning: Analyses and Methods
Neural algorithmic reasoning is an emerging research direction that endows neura
l networks with the ability to mimic algorithmic executions step-by-step. A comm
on paradigm in existing designs involves the use of historical embeddings in pre
dicting the results of future execution steps. Our observation in this work is t
hat such historical dependence intrinsically contradicts the Markov nature of al
gorithmic reasoning tasks. Based on this motivation, we present our ForgetNet, w
hich does not use historical embeddings and thus is consistent with the Markov n
ature of the tasks. To address challenges in training ForgetNet at early stages,
 we further introduce G-ForgetNet, which uses a gating mechanism to allow for th
e selective integration of historical embeddings. Such an enhanced capability pr
ovides valuable computational pathways during the model's early training phase.
Our extensive experiments, based on the CLRS-30 algorithmic reasoning benchmark,
 demonstrate that both ForgetNet and G-ForgetNet achieve better generalization c
apability than existing methods. Furthermore, we investigate the behavior of the
 gating mechanism, highlighting its degree of alignment with our intuitions and
its effectiveness for robust performance. Our code is publicly available at http
s://github.com/divelab/ForgetNet.
****
Ziyao Wang,Jianyu Wang,Ang Li
FedHyper: A Universal and Robust Learning Rate Scheduler for Federated Learning
with Hypergradient Descent
The theoretical landscape of federated learning (FL) undergoes rapid evolution,
but its practical application encounters a series of intricate challenges, and h
yperparameter optimization is one of these critical challenges. Amongst the dive
rse adjustments in hyperparameters, the adaptation of the learning rate emerges
as a crucial component, holding the promise of significantly enhancing the effic

acy of FL systems. In response to this critical need, this paper presents FedHyper, a novel hypergradient-based learning rate adaptation algorithm specifically designed for FL. FedHyper serves as a universal learning rate scheduler that can adapt both global and local rates as the training progresses. In addition, FedHyper not only showcases unparalleled robustness to a spectrum of initial learning rate configurations but also significantly alleviates the necessity for laborious empirical learning rate adjustments. We provide a comprehensive theoretical analysis of FedHyper’s convergence rate and conduct extensive experiments on vision and language benchmark datasets. The results demonstrate that FEDHYPER consistently converges 1.1-3× faster than FedAvg and the competing baselines while achieving superior final accuracy. Moreover, FEDHYPER catalyzes a remarkable surge in accuracy, augmenting it by up to 15% compared to FedAvg under suboptimal initial learning rate settings.
****
Yifan Lu,Yue Hu,Yiqi Zhong,Dequan Wang,Siheng Chen,Yanfeng Wang
An Extensible Framework for Open Heterogeneous Collaborative Perception
Collaborative perception aims to mitigate the limitations of single-agent perception, such as occlusions, by facilitating data exchange among multiple agents. However, most current works consider a homogeneous scenario where all agents use identity sensors and perception models. In reality, heterogeneous agent types may continually emerge and inevitably face a domain gap when collaborating with existing agents. In this paper, we introduce a new open heterogeneous problem: how to accommodate continually emerging new heterogeneous agent types into collaborative perception, while ensuring high perception performance and low integration cost? To address this problem, we propose HEterogeneous ALliance (HEAL), a novel extensible collaborative perception framework. HEAL first establishes a unified feature space with initial agents via a novel multi-scale foreground-aware Pyramid Fusion network. When heterogeneous new agents emerge with previously unseen modalities or models, we align them to the established unified space with an innovative backward alignment. This step only involves individual training on the new agent type, thus presenting extremely low training costs and high extensibility. To enrich agents' data heterogeneity, we bring OPV2V-H, a new large-scale dataset with more diverse sensor types. Extensive experiments on OPV2V-H and DAIR-V2X datasets show that HEAL surpasses SOTA methods in performance while reducing the training parameters by 91.5\% when integrating 3 new agent types. We further implement a comprehensive codebase at: https://github.com/yifanlu0227/HEAL
****
Zhengxiang Shi,Aldo Lipani
DePT: Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning
Prompt tuning (PT), where a small amount of trainable soft (continuous) prompt vectors is affixed to the input of language models (LM), has shown promising results across various tasks and models for parameter-efficient fine-tuning (PEFT). PT stands out from other PEFT approaches because it maintains competitive performance with fewer trainable parameters and does not drastically scale up its parameters as the model size expands. However, PT introduces additional soft prompt tokens, leading to longer input sequences, which significantly impacts training and inference time and memory usage due to the Transformer's quadratic complexity. Particularly concerning for Large Language Models (LLMs) that face heavy daily querying. To address this issue, we propose Decomposed Prompt Tuning (DePT), which decomposes the soft prompt into a shorter soft prompt and a pair of low-rank matrices that are then optimised with two different learning rates. This allows DePT to achieve better performance while saving substantial memory and time costs compared to vanilla PT and its variants, without changing trainable parameter sizes. Through extensive experiments on 23 natural language processing (NLP) and vision-language (VL) tasks, we demonstrate that DePT outperforms state-of-the-art PEFT approaches, including the full fine-tuning baseline, in some scenarios. Additionally, we empirically show that DEPT grows more efficient as the model size increases. Our further study reveals that DePT integrates seamlessly with parameter-efficient transfer learning in the few-shot learning setting and highlights its adaptability to various model architectures and sizes.

****
Yihan Du,R. Srikant,Wei Chen
Cascading Reinforcement Learning
Cascading bandits have gained popularity in recent years due to their applicability to recommendation systems and online advertising. In the cascading bandit model, at each timestep, an agent recommends an ordered subset of items (called an item list) from a pool of items, each associated with an unknown attraction probability. Then, the user examines the list, and clicks the first attractive item (if any), and after that, the agent receives a reward. The goal of the agent is to maximize the expected cumulative reward. However, the prior literature on cascading bandits ignores the influences of user states (e.g., historical behaviors) on recommendations and the change of states as the session proceeds. Motivated by this fact, we propose a generalized cascading RL framework, which considers the impact of user states and state transition into decisions. In cascading RL, we need to select items not only with large attraction probabilities but also leading to good successor states. This imposes a huge computational challenge due to the combinatorial action space. To tackle this challenge, we delve into the properties of value functions, and design an oracle BestPerm to efficiently find the optimal item list. Equipped with BestPerm, we develop two algorithms CascadingVI and CascadingBPI, which are both computationally-efficient and sample-efficient, and provide near-optimal regret and sample complexity guarantees. Furthermore, we present experiments to show the improved computational and sample efficiencies of our algorithms compared to straightforward adaptations of existing RL algorithms in practice.
****
Sameera Ramasinghe,Violetta Shevchenko,Gil Avraham,Hisham Husain,Anton van den Hengel
Improving the Convergence of Dynamic NeRFs via Optimal Transport
Synthesizing novel views for dynamic scenes from a collection of RGB inputs poses significant challenges due to the inherent under-constrained nature of the problem. To mitigate this ill-posedness, practitioners in the field of neural radiance fields (NeRF) often resort to the adoption of intricate geometric regularization techniques, including scene flow, depth estimation, or learned perceptual similarity. While these geometric cues have demonstrated their effectiveness, their incorporation leads to evaluation of computationally expensive off-the-shelf models, introducing substantial computational overhead into the pipeline. Moreover, seamlessly integrating such modules into diverse dynamic NeRF models can be a non-trivial task, hindering their utilization in an architecture-agnostic manner. In this paper, we propose a theoretically grounded, lightweight regularizer by treating the dynamics of a time-varying scene as a low-frequency change of a probability distribution of the light intensity. We constrain the dynamics of this distribution using optimal transport (OT) and provide error bounds under reasonable assumptions. Our regularization is learning-free, architecture agnostic, and can be implemented with just a few lines of code. Finally, we demonstrate the practical efficacy of our regularizer across state-of-the-art architectures.
****
Neha Gupta,Harikrishna Narasimhan,Wittawat Jitkrittum,Ankit Singh Rawat,Aditya Krishna Menon,Sanjiv Kumar
Language Model Cascades: Token-Level Uncertainty And Beyond
Recent advances in language models (LMs) have led to significant improvements in quality on complex NLP tasks, but at the expense of increased inference costs. A simple strategy to achieve more favorable cost-quality tradeoffs is cascading: here, a small model is invoked for most “easy” instances, while a few “hard” instances are deferred to the large model. While the principles underpinning effective cascading are well-studied for classification tasks — with deferral based on predicted class uncertainty favored theoretically and practically — a similar understanding is lacking for generative LM tasks. In this work, we initiate a systematic study of deferral rules for LM cascades. We begin by examining the natural extension of predicted class uncertainty to generative LM tasks, namely, the predicted sequence uncertainty. We show that this measure suffers from the leng

th bias problem, either over- or under-emphasizing outputs based on their lengths. This is because LMs produce a sequence of uncertainty values, one for each output token; and moreover, the number of output tokens is variable across different examples. To mitigate the length bias, we propose to exploit the richer token-level uncertainty information implicit in generative LMs. We argue that naive predicted sequence uncertainty corresponds to a simple aggregation of these uncertainties. By contrast, we show that incorporating token-level uncertainty through learned post-hoc deferral rules can significantly outperform such simple aggregation strategies, via experiments on a range of natural language benchmarks with FLAN-T5 models. We further show that incorporating embeddings from the smaller model and intermediate layers of the larger model can give an additional boost in the overall cost-quality tradeoff.
****
Hwanwoo Kim,Xin Zhang,Jiwei Zhao,Qinglong Tian
ReTaSA: A Nonparametric Functional Estimation Approach for Addressing Continuous Target Shift
The presence of distribution shifts poses a significant challenge for deploying modern machine learning models in real-world applications. This work focuses on the target shift problem in a regression setting (Zhang et al., 2013; Nguyen et al., 2016). More specifically, the target variable y (also known as the response variable), which is continuous, has different marginal distributions in the training source and testing domain, while the conditional distribution of features $\boldsymbol{x}$ given y remains the same. While most literature focuses on classification tasks with finite target space, the regression problem has an *infinite dimensional* target space, which makes many of the existing methods inapplicable. In this work, we show that the continuous target shift problem can be addressed by estimating the importance weight function from an ill-posed integral equation. We propose a nonparametric regularized approach named *ReTaSA* to solve the ill-posed integral equation and provide theoretical justification for the estimated importance weight function. The effectiveness of the proposed method has been demonstrated with extensive numerical studies on synthetic and real-world datasets.
****
Xin Zheng,Dongjin Song,Qingsong Wen,Bo Du,Shirui Pan
Online GNN Evaluation Under Test-time Graph Distribution Shifts
Evaluating the performance of a well-trained GNN model on real-world graphs is a pivotal step for reliable GNN online deployment and serving.
Due to a lack of test node labels and unknown potential training-test graph data distribution shifts, conventional model evaluation encounters limitations in calculating performance metrics (e.g., test error) and measuring graph data-level discrepancies, particularly when the training graph used for developing GNNs remains unobserved during test time.
In this paper, we study a new research problem, online GNN evaluation, which aims to provide valuable insights into the well-trained GNNs's ability to effectively generalize to real-world unlabeled graphs under the test-time graph distribution shifts.
Concretely, we develop an effective learning behavior discrepancy score, dubbed LeBeD, to estimate the test-time generalization errors of well-trained GNN models.
Through a novel GNN re-training strategy with a parameter-free optimality criterion, the proposed LeBeD comprehensively integrates learning behavior discrepancies from both node prediction and structure reconstruction perspectives.
This enables the effective evaluation of the well-trained GNNs' ability to capture test node semantics and structural representations, making it an expressive metric for estimating the generalization error in online GNN evaluation.
Extensive experiments on real-world test graphs under diverse graph distribution shifts could verify the effectiveness of the proposed method, revealing its strong correlation with ground-truth test errors on various well-trained GNN models.
****

Kazem Meidani,Parshin Shojaee,Chandan K. Reddy,Amir Barati Farimani
SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-trainin
g
In an era where symbolic mathematical equations are indispensable for modeling c
omplex natural phenomena, scientific inquiry often involves collecting observati
ons and translating them into mathematical expressions. Recently, deep learning
has emerged as a powerful tool for extracting insights from data. However, exist
ing models typically specialize in either numeric or symbolic domains, and are u
sually trained in a supervised manner tailored to specific tasks. This approach
neglects the substantial benefits that could arise from a task-agnostic multi-mo
dal understanding between symbolic equations and their numeric counterparts. To
bridge the gap, we introduce SNIP, a Symbolic-Numeric Integrated Pre-training mo
del, which employs contrastive learning between symbolic and numeric domains, en
hancing their mutual similarities in the embeddings. By performing latent space
analysis, we observe that SNIP provides cross-domain insights into the represent
ations, revealing that symbolic supervision enhances the embeddings of numeric d
ata and vice versa. We evaluate SNIP across diverse tasks, including symbolic-to
-numeric mathematical property prediction and numeric-to-symbolic equation disco
very, commonly known as symbolic regression. Results show that SNIP effectively
transfers to various tasks, consistently outperforming fully supervised baseline
s and competing strongly with established task-specific methods, especially in t
he low data regime scenarios where available data is limited.
****
Blake Bordelon,Lorenzo Noci,Mufan Bill Li,Boris Hanin,Cengiz Pehlevan
Depthwise Hyperparameter Transfer in Residual Networks: Dynamics and Scaling Lim
it
The cost of hyperparameter tuning in deep learning has been rising with model si
zes, prompting practitioners to find new tuning methods using a proxy of smaller
 networks. One such proposal uses μP parameterized networks, where the optim
al hyperparameters for small width networks *transfer* to networks with arbitrar
ily large width. However, in this scheme, hyperparameters do not transfer across
 depths. As a remedy, we study residual networks with a residual branch scale of
 $1/\sqrt{\text{depth}}$ in combination with the μP parameterization. We pro
vide experiments demonstrating that residual architectures including convolution
al ResNets and vision transformers trained with this parameterization exhibit tr
ansfer of optimal hyperparameters across width and depth on CIFAR-10 and ImageNe
t. Furthermore, our empirical findings are supported and motivated by theory. U
sing recent developments in the dynamical mean field theory (DMFT) description o
f neural network learning dynamics, we show that this parameterization of ResNet
s admits a well-defined feature learning joint infinite-width and infinite-depth
 limit and show convergence of finite-size network dynamics towards this limit.
****
Meng-Chieh Lee,Haiyang Yu,Jian Zhang,Vassilis N. Ioannidis,Xiang song,Soji Adesh
ina,Da Zheng,Christos Faloutsos
NetInfoF Framework: Measuring and Exploiting Network Usable Information
Given a node-attributed graph, and a graph task (link prediction or node classif
ication), can we tell if a graph neural network (GNN) will perform well? More sp
ecifically, do the graph structure and the node features carry enough usable inf
ormation for the task? Our goals are
(1) to develop a fast tool to measure how much information is in the graph struc
ture and in the node features, and
(2) to exploit the information to solve the task, if there is enough.
We propose NetInfoF, a framework including NetInfoF_Probe and NetInfoF_Act, for
the measurement and the exploitation of network usable information (NUI), respec
tively. Given a graph data, NetInfoF_Probe measures NUI without any model traini
ng, and NetInfoF_Act solves link prediction and node classification, while two m
odules share the same backbone.
In summary, NetInfoF has following notable advantages:
(a) General, handling both link prediction and node classification;
(b) Principled, with theoretical guarantee and closed-form solution;

(c) Effective, thanks to the proposed adjustment to node similarity;
(d) Scalable, scaling linearly with the input size.
In our carefully designed synthetic datasets, NetInfoF correctly identifies the ground truth of NUI and is the only method being robust to all graph scenarios. Applied on real-world datasets, NetInfoF wins in 11 out of 12 times on link prediction compared to general GNN baselines.
****
Riccardo Massidda,Francesco Landolfi,Martina Cinquini,Davide Bacciu
Constraint-Free Structure Learning with Smooth Acyclic Orientations
The structure learning problem consists of fitting data generated by a Directed Acyclic Graph (DAG) to correctly reconstruct its arcs. In this context, differentiable approaches constrain or regularize an optimization problem with a continuous relaxation of the acyclicity property. The computational cost of evaluating graph acyclicity is cubic on the number of nodes and significantly affects scalability. In this paper, we introduce COSMO, a constraint-free continuous optimization scheme for acyclic structure learning. At the core of our method lies a novel differentiable approximation of an orientation matrix parameterized by a single priority vector. Differently from previous works, our parameterization fits a smooth orientation matrix and the resulting acyclic adjacency matrix without evaluating acyclicity at any step. Despite this absence, we prove that COSMO always converges to an acyclic solution. In addition to being asymptotically faster, our empirical analysis highlights how COSMO performance on graph reconstruction compares favorably with competing structure learning methods.
****
Pan Lu,Hritik Bansal,Tony Xia,Jiacheng Liu,Chunyuan Li,Hannaneh Hajishirzi,Hao Cheng,Kai-Wei Chang,Michel Galley,Jianfeng Gao
MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts
Large Language Models (LLMs) and Large Multimodal Models (LMMs) exhibit impressive problem-solving skills in many tasks and domains, but their ability in mathematical reasoning in visual contexts has not been systematically studied. To bridge this gap, we present MathVista, a benchmark designed to combine challenges from diverse mathematical and visual tasks. It consists of 6,141 examples, derived from 28 existing multimodal datasets involving mathematics and 3 newly created datasets (i.e., IQTest, FunctionQA, and PaperQA). Completing these tasks requires fine-grained, deep visual understanding and compositional reasoning, which all state-of-the-art foundation models find challenging. With MathVista, we have conducted a comprehensive, quantitative evaluation of 12 prominent foundation models. The best-performing GPT-4V model achieves an overall accuracy of 49.9%, substantially outperforming Bard, the second-best performer, by 15.1%. Our in-depth analysis reveals that the superiority of GPT-4V is mainly attributed to its enhanced visual perception and mathematical reasoning. However, GPT-4V still falls short of human performance by 10.4%, as it often struggles to understand complex figures and perform rigorous reasoning. This significant gap underscores the critical role that MathVista will play in the development of general-purpose AI agents capable of tackling mathematically intensive and visually rich real-world tasks. We further explore the new ability of self-verification, the application of self-consistency, and the interactive chatbot capabilities of GPT-4V, highlighting its promising potential for future research. The project is available at https://mathvista.github.io/.
****
Yoonyoung Cho,Junhyek Han,Yoontae Cho,Beomjoon Kim
CORN: Contact-based Object Representation for Nonprehensile Manipulation of General Unseen Objects
Nonprehensile manipulation is essential for manipulating objects that are too thin, large, or otherwise ungraspable in the wild. To sidestep the difficulty of contact modeling in conventional modeling-based approaches, reinforcement learning (RL) has recently emerged as a promising alternative. However, previous RL approaches either lack the ability to generalize over diverse object shapes, or use simple action primitives that limit the diversity of robot motions. Furthermore

, using RL over diverse object geometry is challenging due to the high cost of t
raining a policy that takes in high-dimensional sensory inputs. We propose a nov
el contact-based object representation and pretraining pipeline to tackle this.
To enable massively parallel training, we leverage a lightweight patch-based tra
nsformer architecture for our encoder that processes point clouds, thus scaling
our training across thousands of environments. Compared to learning from scratch
, or other shape representation baselines, our representation facilitates both t
ime- and data-efficient learning. We validate the efficacy of our overall system
 by zero-shot transferring the trained policy to novel real-world objects. We hi
ghly recommend the video attached in the supplementary material. Code and videos
 are available at \url{https://sites.google.com/view/contact-non-prehensile}.
****
Xiangyan Liu,Rongxue LI,Wei Ji,Tao Lin
Towards Robust Multi-Modal Reasoning via Model Selection
The reasoning capabilities of LLM (Large Language Model) are widely acknowledged
 in recent research, inspiring studies on tool learning and autonomous agents. L
LM serves as the ``brain'' of the agent, orchestrating multiple tools for collab
orative multi-step task solving. Unlike methods invoking tools like calculators
or weather APIs for straightforward tasks, multi-modal agents excel by integrati
ng diverse AI models for complex challenges. However, current multi-modal agents
 neglect the significance of model selection: they primarily focus on the planni
ng and execution phases, and will only invoke predefined task-specific models fo
r each subtask, making the execution fragile. Meanwhile, other traditional model
 selection methods are either incompatible with or suboptimal for the multi-moda
l agent scenarios, due to ignorance of dependencies among subtasks arising by mu
lti-step reasoning.

To this end, we identify the key challenges therein and propose the $\textbf{\te
xtit{M}}^\textbf{\textit{3}}$ framework as a plug-in with negligible runtime ove
rhead at test-time. This framework improves model selection and bolsters the rob
ustness of multi-modal agents in multi-step reasoning. In the absence of suitabl
e benchmarks, we create MS-GQA, a new dataset specifically designed to investiga
te the model selection challenge in multi-modal agents. Our experiments reveal t
hat our framework enables dynamic model selection, considering both user inputs
and subtask dependencies, thereby robustifying the overall reasoning process. Ou
r code and benchmark: https://github.com/LINs-lab/M3.
****
Wenhao Wang,Muhammad Ahmad Kaleem,Adam Dziedzic,Michael Backes,Nicolas Papernot,
Franziska Boenisch
Memorization in Self-Supervised Learning Improves Downstream Generalization
Self-supervised learning (SSL) has recently received significant attention due t
o its ability to train high-performance encoders purely on unlabeled data---ofte
n scraped from the internet. This data can still be sensitive and empirical evid
ence suggests that SSL encoders memorize private information of their training d
ata and can disclose them at inference time. Since existing theoretical definiti
ons of memorization from supervised learning rely on labels, they do not transfe
r to SSL. To address this gap, we propose a framework for defining memorization
within the context of SSL. Our definition compares the difference in alignment o
f representations for data points and their augmented views returned by both enc
oders that were trained on these data points and encoders that were not. Through
 comprehensive empirical analysis on diverse encoder architectures and datasets
we highlight that even though SSL relies on large datasets and strong augmentati
ons---both known in supervised learning as regularization techniques that reduce
 overfitting---still significant fractions of training data points experience hi
gh memorization. Through our empirical results, we show that this memorization i
s essential for encoders to achieve higher generalization performance on differe
nt downstream tasks.
****
Yonatan Oren,Nicole Meister,Niladri S. Chatterji,Faisal Ladhak,Tatsunori Hashimo
to

Proving Test Set Contamination in Black-Box Language Models
Large language models are trained on vast amounts of internet data, prompting concerns that they have memorized public benchmarks. Detecting this type of contamination is challenging because the pretraining data used by proprietary models are often not publicly accessible.

We propose a procedure for detecting test set contamination of language models with exact false positive guarantees and without access to pretraining data or model weights. Our approach leverages the fact that when there is no data contamination, all orderings of an exchangeable benchmark should be equally likely. In contrast, the tendency for language models to memorize example order means that a contaminated language model will find certain canonical orderings to be much more likely than others. Our test flags potential contamination whenever the likelihood of a canonically ordered benchmark dataset is significantly higher than the likelihood after shuffling the examples.

We demonstrate that our procedure is sensitive enough to reliably detect contamination in challenging situations, including models as small as 1.4 billion parameters, on small test sets only 1000 examples, and datasets that appear only a few times in the pretraining corpus. Finally, we evaluate LLaMA-2 to apply our test in a realistic setting and find our results to be consistent with existing contamination evaluations.
****
Shreyas Havaldar,Navodita Sharma,Shubhi Sareen,Karthikeyan Shanmugam,Aravindan Raghuveer
Learning from Label Proportions: Bootstrapping Supervised Learners via Belief Propagation
Learning from Label Proportions (LLP) is a learning problem where only aggregate level labels are available for groups of instances, called bags, during training, and the aim is to get the best performance at the instance-level on the test data. This setting arises in domains like advertising and medicine due to privacy considerations. We propose a novel algorithmic framework for this problem that iteratively performs two main steps. For the first step (Pseudo Labeling) in every iteration, we define a Gibbs distribution over binary instance labels that incorporates a) covariate information through the constraint that instances with similar covariates should have similar labels and b) the bag level aggregated label. We then use Belief Propagation (BP) to marginalize the Gibbs distribution to obtain pseudo labels. In the second step (Embedding Refinement), we use the pseudo labels to provide supervision for a learner that yields a better embedding. Further, we iterate on the two steps again by using the second step's embeddings as new covariates for the next iteration. In the final iteration, a classifier is trained using the pseudo labels. Our algorithm displays strong gains against several SOTA baselines (upto **15%**) for the LLP Binary Classification problem on various dataset types - tabular and Image. We achieve these improvements with minimal computational overhead above standard supervised learning due to Belief Propagation, for large bag sizes, even for a million samples.
****
Hao Zhang,Fang Li,Samyak Rawlekar,Narendra Ahuja
Learning Implicit Representation for Reconstructing Articulated Objects
3D Reconstruction of moving articulated objects without additional information about object structure is a challenging problem. Current methods overcome such challenges by employing category-specific skeletal models. Consequently, they do not generalize well to articulated objects in the wild. We treat an articulated object as an unknown, semi-rigid skeletal structure surrounded by nonrigid material (e.g., skin). Our method simultaneously estimates the visible (explicit) representation (3D shapes, colors, camera parameters) and the underlying (implicit) skeletal representation, from motion cues in the object video without 3D supervision. Our implicit representation consists of four parts. (1) skeleton, which specifies which semi-rigid parts are connected. (2) Semi-rigid Part Assignment, which associates each surface vertex with a semi-rigid part. (3) Rigidity Coeffici

ents, specifying the articulation of the local surface. (4) Time-Varying Transformations, which specify the skeletal motion and surface deformation parameters. We introduce an algorithm that uses these constraints as regularization terms and iteratively estimates both implicit and explicit representations. Our method is category-agnostic, thus eliminating the need for category-specific skeletons, we show that our method outperforms state-of-the-art across standard video datasets.
****
Hanwen Jiang,Zhenyu Jiang,Yue Zhao,Qixing Huang
LEAP: Liberate Sparse-View 3D Modeling from Camera Poses
Are camera poses necessary for multi-view 3D modeling? Existing approaches predominantly assume access to accurate camera poses. While this assumption might hold for dense views, accurately estimating camera poses for sparse views is often elusive. Our analysis reveals that noisy estimated poses lead to degraded performance for existing sparse-view 3D modeling methods. To address this issue, we present LEAP, a novel pose-free approach, therefore challenging the prevailing notion that camera poses are indispensable. LEAP discards pose-based operations and learns geometric knowledge from data. LEAP is equipped with a neural volume, which is shared across scenes and is parameterized to encode geometry and texture priors. For each incoming scene, we update the neural volume by aggregating 2D image features in a feature-similarity-driven manner. The updated neural volume is decoded into the radiance field, enabling novel view synthesis from any viewpoint. On both object-centric and scene-level datasets, we show that LEAP significantly outperforms prior methods when they employ predicted poses from state-of-the-art pose estimators. Notably, LEAP performs on par with prior approaches that use ground-truth poses while running $400\times$ faster than PixelNeRF. We show LEAP generalizes to novel object categories and scenes, and learns knowledge closely resembles epipolar geometry.
****
Weiran Yao,Shelby Heinecke,Juan Carlos Niebles,Zhiwei Liu,Yihao Feng,Le Xue,Rithesh R N,Zeyuan Chen,Jianguo Zhang,Devansh Arpit,Ran Xu,Phil L Mui,Huan Wang,Caiming Xiong,Silvio Savarese
Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization
Recent months have seen the emergence of a powerful new trend in which large language models (LLMs) are augmented to become autonomous language agents capable of performing objective oriented multi-step tasks on their own, rather than merely responding to queries from human users. Most existing language agents, however, are not optimized using environment-specific rewards. Although some agents enable iterative refinement through verbal feedback, they do not reason and plan in ways that are compatible with gradient-based learning from rewards. This paper introduces a principled framework for reinforcing large language agents by learning a retrospective model, which automatically tunes the language agent prompts from environment feedback through policy gradient. Specifically, our proposed agent architecture learns from rewards across multiple environments and tasks, for fine-tuning a pre-trained language model which refines the language agent prompt by summarizing the root cause of prior failed attempts and proposing action plans. Experimental results on various tasks demonstrate that the language agents improve over time and that our approach considerably outperforms baselines that do not properly leverage gradients from the environment.
****
Mahdi Karami
HiGen: Hierarchical Graph Generative Networks
Most real-world graphs exhibit a hierarchical structure, which is often overlooked by existing graph generation methods. To address this limitation, we propose a novel graph generative network that captures the hierarchical nature of graphs and successively generates the graph sub-structures in a coarse-to-fine fashion. At each level of hierarchy, this model generates communities in parallel, followed by the prediction of cross-edges between communities using separate neural networks. This modular approach enables scalable graph generation for large and

complex graphs. Moreover, we model the output distribution of edges in the hier
archical graph with a multinomial distribution and derive a recursive factorizat
ion for this distribution. This enables us to generate community graphs with in
teger-valued edge weights in an autoregressive manner. Empirical studies demonst
rate the effectiveness and scalability of our proposed generative model, achievi
ng state-of-the-art performance in terms of graph quality across various benchma
rk datasets.
Code available at https://github.com/Karami-m/HiGen_main.
****
YINGWEI MA,Yue Liu,Yue Yu,Yuanliang Zhang,Yu Jiang,Changjian Wang,Shanshan Li
At Which Training Stage Does Code Data Help LLMs Reasoning?
Large Language models (LLMs) have exhibited remarkable reasoning capabilities an
d become the foundation of language technologies. Inspired by the great success
of code data in training LLMs, we naturally wonder at which training stage intro
ducing code data can really help LLMs reasoning. To this end, this paper systema
tically explores the impact of code data on LLMs at different stages. Concretely
, we introduce the code data at the pre-training stage, instruction-tuning stage
, and both of them, respectively. Then, the reasoning capability of LLMs is comp
rehensively and fairly evaluated via six reasoning tasks. We critically analyze
the experimental results and provide conclusions with insights. First, pre-train
ing LLMs with the mixture of code and text can significantly enhance LLMs' gener
al reasoning capability almost without negative transfer on other tasks. Besides
, at the instruction-tuning stage, code data endows LLMs the task-specific reaso
ning capability. Moreover, the dynamic mixing strategy of code and text data ass
ists LLMs to learn reasoning capability step-by-step during training. These insi
ghts deepen the understanding of LLMs regarding reasoning ability for their appl
ication, such as scientific question answering, legal support, etc.
****
Tianyu Du,Luca Melis,Ting Wang
ReMasker: Imputing Tabular Data with Masked Autoencoding
We present ReMasker, a new method of imputing missing values in tabular data by
extending the masked autoencoding framework. Compared with prior work, ReMasker
is extremely simple -- besides the missing values (i.e., naturally masked), we r
andomly "re-mask" another set of values, optimize the autoencoder by reconstruct
ing this re-masked set, and apply the trained model to predict the missing value
s; and yet highly effective -- with extensive evaluation on benchmark datasets,
we show that ReMasker performs on par with or outperforms state-of-the-art metho
ds in terms of both imputation fidelity and utility under various missingness se
ttings, while its performance advantage often increases with the ratio of missin
g data. We further explore theoretical justification for its effectiveness, show
ing that ReMasker tends to learn missingness-invariant representations of tabula
r data. Our findings indicate that masked modeling represents a promising direct
ion for further research on tabular data imputation. The code is publicly availa
ble.
****
Haruo Hosoya
A Cognitive Model for Learning Abstract Relational Structures from Memory-based
Decision-Making Tasks
Motivated by a recent neuroscientific hypothesis, some theoretical studies have
accounted for neural cognitive maps in the rodent hippocampal formation as a rep
resentation of the general relational structure across task environments. Howev
er, despite their remarkable results, it is unclear whether their account can be
 extended to more general settings beyond spatial random-walk tasks in 2D enviro
nments. To address this question, we construct a novel cognitive model that per
forms memory-based relational decision-making tasks, inspired by previous human
studies, for learning abstract structures in non-spatial relations. Building on
 previous approaches of modular architecture, we develop a learning algorithm th
at performs reward-guided search for representation of abstract relations, while
 dynamically maintaining their binding to concrete entities using our specific m
emory mechanism enabling content replacement. Our experiments show (i) the capa

bility of our model to capture relational structures that can generalize over new domains with unseen entities, (ii) the difficulty of our task that leads previous models, including Neural Turing Machine and vanilla Transformer, to complete failure, and (iii) the similarity of performance and internal representations of our model to recent human behavioral and fMRI experimental data in the human hippocampal formation.
****
Chen Gan,Zihao Yin,Kelei He,Yang Gao,Junfeng Zhang
Diving Segmentation Model into Pixels
More distinguishable and consistent pixel features for each category will benefit the semantic segmentation under various settings.
Existing efforts to mine better pixel-level features attempt to explicitly model the categorical distribution, which fails to achieve optimal due to the significant pixel feature variance.
Moreover, prior research endeavors have scarcely delved into the thorough analysis and meticulous handling of pixel-level variance, leaving semantic segmentation at a coarse granularity.
In this work, we analyze the causes of pixel-level variance and introduce the concept of $\textbf{pixel learning}$ to concentrate on the tailored learning process of pixels to handle the pixel-level variance, enhancing the per-pixel recognition capability of segmentation models.
Under the context of the pixel learning scheme, each image is viewed as a distribution of pixels, and pixel learning aims to pursue consistent pixel representation inside an image, continuously align pixels from different images (distributions), and eventually achieve consistent pixel representation for each category, even cross-domains.
We proposed a pure pixel-level learning framework, namely PiXL, which consists of a pixel partition module to divide pixels into sub-domains, a prototype generation, a selection module to prepare targets for subsequent alignment, and a pixel alignment module to guarantee pixel feature consistency intra-/inter-images, and inter-domains.
Extensive evaluations of multiple learning paradigms, including unsupervised domain adaptation and semi-/fully-supervised segmentation, show that PiXL outperforms state-of-the-art performances, especially when annotated images are scarce.
Visualization of the embedding space further demonstrates that pixel learning attains a superior representation of pixel features.
The code is available at https://github.com/ChenGan-JS/PiXL.
****
Alex Fang,Albin Madappally Jose,Amit Jain,Ludwig Schmidt,Alexander T Toshev,Vaishaal Shankar
Data Filtering Networks
Large training sets have become a cornerstone of machine learning and are the foundation for recent advances in language modeling and multimodal learning. While data curation for pre-training is often still ad-hoc, one common paradigm is to first collect a massive pool of data from the Web and then filter this candidate pool down to an actual training set via various heuristics. In this work, we study the problem of learning a *data filtering network* (DFN) for this second step of filtering a large uncurated dataset. Our key finding is that the quality of a network for filtering is distinct from its performance on downstream tasks: for instance, a model that performs well on ImageNet can yield worse training sets than a model with low ImageNet accuracy that is trained on a small amount of high-quality data. Based on our insights, we construct new data filtering networks that induce state-of-the-art image-text datasets. Specifically, our best performing dataset DFN-5B enables us to train state-of-the-art models for their compute budgets: among other improvements on a variety of tasks, a ViT-H trained on our dataset achieves 83.0% zero-shot transfer accuracy on ImageNet, out-performing larger models trained on other datasets such as LAION-2B, DataComp-1B, or OpenAI’s WIT. In order to facilitate further research in dataset design, we also release a new 2 billion example dataset DFN-2B and show that high performance data filtering networks can be trained from scratch using only publicly available d

ata.
****
Jinxi Xiang,Ricong Huang,Jun Zhang,Guanbin Li,Xiao Han,Yang Wei
VersVideo: Leveraging Enhanced Temporal Diffusion Models for Versatile Video Generation
Creating stable, controllable videos is a complex task due to the need for significant variation in temporal dynamics and cross-frame temporal consistency. To address this, we enhance the spatial-temporal capability and introduce a versatile video generation model, VersVideo, which leverages textual, visual, and stylistic conditions. Current video diffusion models typically extend image diffusion architectures by supplementing 2D operations (such as convolutions and attentions) with temporal operations. While this approach is efficient, it often restricts spatial-temporal performance due to the oversimplification of standard 3D operations. To counter this, we incorporate two key elements: (1) multi-excitation paths for spatial-temporal convolutions with dimension pooling across different axes, and (2) multi-expert spatial-temporal attention blocks. These enhancements boost the model's spatial-temporal performance without significantly escalating training and inference costs. We also tackle the issue of information loss that arises when a variational autoencoder is used to transform pixel space into latent features and then back into pixel frames. To mitigate this, we incorporate temporal modules into the decoder to maintain inter-frame consistency. Lastly, by utilizing the innovative denoising UNet and decoder, we develop a unified ControlNet model suitable for various conditions, including image, Canny, HED, depth, and style. Examples of the videos generated by our model can be found at https://anonymous-pages.github.io/video_demos/.
****
Shuo Chen,Gang Niu,Chen Gong,Okan Koc,Jian Yang,Masashi Sugiyama
Robust Similarity Learning with Difference Alignment Regularization
Similarity-based representation learning has shown impressive capabilities in both supervised (e.g., metric learning) and unsupervised (e.g., contrastive learning) scenarios. Existing approaches effectively constrained the representation difference (i.e., the disagreement between the embeddings of two instances) to fit the corresponding (pseudo) similarity supervision. However, most of them can hardly restrict the variation of representation difference, sometimes leading to overfitting results where the clusters are disordered by drastically changed differences. In this paper, we thus propose a novel difference alignment regularization (DAR) to encourage all representation differences between inter-class instances to be as close as possible, so that the learning algorithm can produce consistent differences to distinguish data points from each other. To this end, we construct a new cross-total-variation (CTV) norm to measure the divergence among representation differences, and we convert it into an equivalent stochastic form for easy optimization. Then, we integrate the proposed regularizer into the empirical loss for difference-aligned similarity learning (DASL), shrinking the hypothesis space and alleviating overfitting. Theoretically, we prove that our regularizer tightens the error bound of the traditional similarity learning. Experiments on multi-domain data demonstrate the superiority of DASL over existing approaches in both supervised metric learning and unsupervised contrastive learning tasks.
****
Jiacheng Lin,Meng XU,Zhihua Xiong,Huangang Wang
CAMBranch: Contrastive Learning with Augmented MILPs for Branching
Recent advancements have introduced machine learning frameworks to enhance the Branch and Bound (B\&B) branching policies for solving Mixed Integer Linear Programming (MILP). These methods, primarily relying on imitation learning of Strong Branching, have shown superior performance. However, collecting expert samples for imitation learning, particularly for Strong Branching, is a time-consuming endeavor. To address this challenge, we propose \textbf{C}ontrastive Learning with \textbf{A}ugmented \textbf{M}ILPs for \textbf{Branch}ing (CAMBranch), a framework that generates Augmented MILPs (AMILPs) by applying variable shifting to limited expert data from their original MILPs. This approach enables the acquisition

of a considerable number of labeled expert samples. CAMBranch leverages both MILPs and AMILPs for imitation learning and employs contrastive learning to enhance the model's ability to capture MILP features, thereby improving the quality of branching decisions. Experimental results demonstrate that CAMBranch, trained with only 10\% of the complete dataset, exhibits superior performance. Ablation studies further validate the effectiveness of our method.
****
Xiang Lan,Hanshu Yan,Shenda Hong,Mengling Feng
Towards Enhancing Time Series Contrastive Learning: A Dynamic Bad Pair Mining Approach
Not all positive pairs are beneficial to time series contrastive learning. In this paper, we study two types of bad positive pairs that can impair the quality of time series representation learned through contrastive learning: the noisy positive pair and the faulty positive pair. We observe that, with the presence of noisy positive pairs, the model tends to simply learn the pattern of noise (Noisy Alignment). Meanwhile, when faulty positive pairs arise, the model wastes considerable amount of effort aligning non-representative patterns (Faulty Alignment). To address this problem, we propose a Dynamic Bad Pair Mining (DBPM) algorithm, which reliably identifies and suppresses bad positive pairs in time series contrastive learning. Specifically, DBPM utilizes a memory module to dynamically track the training behavior of each positive pair along training process. This allows us to identify potential bad positive pairs at each epoch based on their historical training behaviors. The identified bad pairs are subsequently down-weighted through a transformation module, thereby mitigating their negative impact on the representation learning process. DBPM is a simple algorithm designed as a lightweight **plug-in** without learnable parameters to enhance the performance of existing state-of-the-art methods. Through extensive experiments conducted on four large-scale, real-world time series datasets, we demonstrate DBPM's efficacy in mitigating the adverse effects of bad positive pairs.
****
MinGyu Choi,Changhee Lee
Conditional Information Bottleneck Approach for Time Series Imputation
Time series imputation presents a significant challenge because it requires capturing the underlying temporal dynamics from partially observed time series data. Among the recent successes of imputation methods based on generative models, the information bottleneck (IB) framework offers a well-suited theoretical foundation for multiple imputations, allowing us to account for the uncertainty associated with the imputed values. However, directly applying the IB framework to time series data without considering their temporal context can lead to a substantial loss of temporal dependencies, which, in turn, can degrade the overall imputation performance. To address such a challenge, we propose a novel conditional information bottleneck (CIB) approach for time series imputation, which aims to mitigate the potentially negative consequences of the regularization constraint by focusing on reducing the redundant information conditioned on the temporal context. We provide a theoretical analysis of its effect by adapting variational decomposition. We use the resulting insight and propose a novel deep learning method that can approximately achieve the proposed CIB objective for time series imputation as a combination of evidence lower bound and novel temporal kernel-enhanced contrastive optimization. Our experiments, conducted on multiple real-world datasets, consistently demonstrate that our method significantly improves imputation performance (including both interpolation and extrapolation), and also enhances classification performance based on the imputed values.
****
Denizalp Goktas,Amy Greenwald,Sadie Zhao,Alec Koppel,Sumitra Ganesh
Generative Adversarial Inverse Multiagent Learning
In this paper, we study inverse game theory (resp. inverse multiagent learning) in
which the goal is to find parameters of a game’s payoff functions for which the expected (resp. sampled) behavior is an equilibrium. We formulate these problems as a generative-adversarial (i.e., min-max) optimization problem, based on which

we develop polynomial-time algorithms the solve them, the former of which
relies on an exact first-order oracle, and the latter, a stochastic one. We exte
nd
our approach to solve inverse multiagent apprenticeship learning in polynomial
time and number of samples, where we seek a simulacrum, i.e., parameters and
an associated equilibrium, which replicate observations in expectation. We find
that our approach outperforms other widely-used methods in predicting prices in
Spanish electricity markets based on time-series data.
****
Jung Hwan Heo,Jeonghoon Kim,Beomseok Kwon,Byeongwook Kim,Se Jung Kwon,Dongsoo Le
e
Rethinking Channel Dimensions to Isolate Outliers for Low-bit Weight Quantizatio
n of Large Language Models
Large Language Models (LLMs) have recently demonstrated a remarkable success acr
oss various tasks. However, efficiently serving LLMs has been a challenge due to
 its large memory bottleneck, specifically in small batch inference settings (e.
g. mobile devices). Weight-only quantization can be a promising approach, but su
b-4 bit quantization remains a challenge due to large-magnitude activation outli
ers. To mitigate the undesirable outlier effect, we first propose per-IC quantiz
ation, a simple yet effective method that creates quantization groups within eac
h input channel (IC) rather than the conventional per-output channel (OC). Our m
ethod is motivated by the observation that activation outliers affect the input
dimension of the weight matrix, so similarly grouping the weights in the IC dire
ction can $\textit{isolate outliers to be within a group}$. We also find that ac
tivation outliers do not dictate quantization difficulty, and inherent weight se
nsitivities also exist. With per-IC quantization as a new outlier-friendly schem
e, we then propose Adaptive Dimensions ($\textbf{AdaDim}$), a versatile quantiza
tion framework that can adapt to various weight sensitivity patterns. We demonst
rate the effectiveness of AdaDim by augmenting prior methods such as Round-To-Ne
arest and GPTQ, showing significant improvements across various language modelin
g benchmarks for both base (up to $+4.7\%$ on MMLU) and instruction-tuned (up to
 $+10\%$ on HumanEval) LLMs.
****
Shangyu Wu,Ying Xiong,Yufei CUI,Xue Liu,Buzhou Tang,Tei-Wei Kuo,Chun Jason Xue
ReFusion: Improving Natural Language Understanding with Computation-Efficient Re
trieval Representation Fusion
Retrieval-based augmentations (RA) incorporating knowledge from an external data
base into language models have greatly succeeded in various knowledge-intensive
(KI) tasks. However, integrating retrievals in non-knowledge-intensive (NKI) tas
ks is still challenging.
Existing works focus on concatenating retrievals with inputs to improve model pe
rformance. Unfortunately, the use of retrieval concatenation-based augmentations
 causes an increase in the input length, substantially raising the computational
 demands of attention mechanisms.
This paper proposes a new paradigm of RA named \textbf{ReFusion}, a computation-
efficient \textbf{Re}trieval representation \textbf{Fusion} with bi-level optimi
zation. Unlike previous works, ReFusion directly fuses the retrieval representat
ions into the hidden states of models.
Specifically, ReFusion leverages an adaptive retrieval integrator to seek the op
timal combination of the proposed ranking schemes across different model layers.
 Experimental results demonstrate that the proposed ReFusion can achieve superio
r and robust performance in various NKI tasks.
****
Naoya Hasegawa,Issei Sato
Exploring Weight Balancing on Long-Tailed Recognition Problem
Recognition problems in long-tailed data, in which the sample size per class is
heavily skewed, have gained importance because the distribution of the sample si
ze per class in a dataset is generally exponential unless the sample size is int
entionally adjusted. Various methods have been devised to address these problems
.

Recently, weight balancing, which combines well-known classical regularization t echniques with two-stage training, has been proposed. Despite its simplicity, it is known for its high performance compared with existing methods devised in var ious ways.
However, there is a lack of understanding as to why this method is effective for long-tailed data. In this study, we analyze weight balancing by focusing on neu ral collapse and the cone effect at each training stage and found that it can be decomposed into an increase in Fisher's discriminant ratio of the feature extra ctor caused by weight decay and cross entropy loss and implicit logit adjustment caused by weight decay and class-balanced loss. Our analysis enables the traini ng method to be further simplified by reducing the number of training stages to one while increasing accuracy. Code is available at https://github.com/HN410/Exp loring-Weight-Balancing-on-Long-Tailed-Recognition-Problem.
****
Zhengbo Wang,Jian Liang,Lijun Sheng,Ran He,Zilei Wang,Tieniu Tan
A Hard-to-Beat Baseline for Training-free CLIP-based Adaptation
Contrastive Language-Image Pretraining (CLIP) has gained popularity for its rema rkable zero-shot capacity.
Recent research has focused on developing efficient fine-tuning methods, such as prompt learning and adapter, to enhance CLIP's performance in downstream tasks.
However, these methods still require additional training time and computational resources, which is undesirable for devices with limited resources.
In this paper, we revisit a classical algorithm, Gaussian Discriminant Analysis (GDA), and apply it to the downstream classification of CLIP.
Typically, GDA assumes that features of each class follow Gaussian distributions with identical covariance.
By leveraging Bayes' formula, the classifier can be expressed in terms of the cl ass means and covariance, which can be estimated from the data without the need for training.
To integrate knowledge from both visual and textual modalities, we ensemble it w ith the original zero-shot classifier within CLIP.
Extensive results on 17 datasets validate that our method surpasses or achieves comparable results with state-of-the-art methods on few-shot classification, imb alanced learning, and out-of-distribution generalization.
In addition, we extend our method to base-to-new generalization and unsupervised learning, once again demonstrating its superiority over competing approaches.
Our code is publicly available at https://github.com/mrflogs/ICLR24.
****
Christos Louizos,Matthias Reisser,Denis Korzhenkov
A Mutual Information Perspective on Federated Contrastive Learning
We investigate contrastive learning in the federated setting through the lens of Sim- CLR and multi-view mutual information maximization. In doing so, we uncove r a connection between contrastive representation learning and user verification ; by adding a user verification loss to each client’s local SimCLR loss we recov er a lower bound to the global multi-view mutual information. To accommodate for the case of when some labelled data are available at the clients, we extend our SimCLR variant to the federated semi-supervised setting. We see that a supervis ed SimCLR objective can be obtained with two changes: a) the contrastive loss is computed between datapoints that share the same label and b) we require an addi tional auxiliary head that predicts the correct labels from either of the two vi ews. Along with the proposed SimCLR extensions, we also study how different sour ces of non-i.i.d.-ness can impact the performance of federated unsupervised lear ning through global mutual information maximization; we find that a global objec tive is beneficial for some sources of non-i.i.d.-ness but can be detrimental fo r others. We empirically evaluate our proposed extensions in various tasks to va lidate our claims and furthermore demonstrate that our proposed modifications ge neralize to other pretraining methods.
****
Badih Ghazi,Yangsibo Huang,Pritish Kamath,Ravi Kumar,Pasin Manurangsi,Chiyuan Zh ang

LabelDP-Pro: Learning with Label Differential Privacy via Projections
Label differentially private (label DP) algorithms seek to preserve the privacy
of the labels in a training dataset in settings where the features are known to
the adversary. In this work, we study a new family of label DP training algorith
ms. Unlike most prior label DP algorithms that have been based on label randomiz
ation, our algorithm naturally leverages the power of the central model of DP. I
t interleaves gradient projection operations with private stochastic gradient de
scent steps in order to improve the utility of the trained model while guarantee
ing the privacy of the labels. We show that such projection-based algorithms can
 be made practical and that they improve on the state-of-the art for label DP tr
aining in the high-privacy regime. We complement our empirical evaluation with t
heoretical results shedding light on the efficacy of our method through the lens
 of bias-variance trade-offs.
****
Patrik Okanovic,Roger Waleffe,Vasilis Mageirakos,Konstantinos Nikolakakis,Amin K
arbasi,Dionysios Kalogerias,Nezihe Merve Gürel,Theodoros Rekatsinas
Repeated Random Sampling for Minimizing the Time-to-Accuracy of Learning
Methods for carefully selecting or generating a small set of training data to le
arn from, i.e., data pruning, coreset selection, and dataset distillation, have
been shown to be effective in reducing the ever-increasing cost of training neur
al networks. Behind this success are rigorously designed, yet expensive, strateg
ies for identifying the most informative training examples out of large datasets
. In this work, we revisit these methods to understand if the additional computa
tional costs associated with such strategies are justified from the perspective
of time-to-accuracy, which has become a critical efficiency measure of deep neur
al network training over large datasets. Surprisingly, we find that many of the
recently proposed methods underperform what we call Repeated Sampling of Random
Subsets (RSRS or RS2), a powerful yet overlooked extension of the standard rando
m baseline that learns from repeatedly sampled data throughout training instead
of a fixed random subset. We test RS2 against thirty-two state-of-the-art data p
runing and distillation methods across four datasets including ImageNet. Our res
ults demonstrate that RS2 significantly reduces time-to-accuracy, particularly i
n practical regimes where accuracy, but not runtime, is similar to that of train
ing on full dataset. For example, when training ResNet-18 on ImageNet, with 10\%
 of the dataset each epoch RS2 reaches an accuracy of 66\% versus 69\% when trai
ning with the full dataset. The best competing method achieves only 55\% while t
raining 1.6$\times$ slower than RS2. Beyond the above meta-study, we discuss the
 theoretical properties of RS2 such as its convergence rate and generalization e
rror. Our primary goal is to highlight that future works that aim to minimize to
tal training cost by using subset selection, need to consider 1) the total compu
tation cost (including preparing the subset) and 2) should aim to outperform a s
imple extension of random sampling (i.e., RS2).
****
Lifan Zhao,Yanyan Shen
Rethinking Channel Dependence for Multivariate Time Series Forecasting: Learning
 from Leading Indicators
Recently, channel-independent methods have achieved state-of-the-art performance
 in multivariate time series (MTS) forecasting. Despite reducing overfitting ris
ks, these methods miss potential opportunities in utilizing channel dependence f
or accurate predictions. We argue that there exist locally stationary lead-lag r
elationships between variates, i.e., some lagged variates may follow the leading
 indicators within a short time period. Exploiting such channel dependence is be
neficial since leading indicators offer advance information that can be used to
reduce the forecasting difficulty of the lagged variates. In this paper, we prop
ose a new method named LIFT that first efficiently estimates leading indicators
and their leading steps at each time step and then judiciously allows the lagged
 variates to utilize the advance information from leading indicators. LIFT plays
 as a plugin that can be seamlessly collaborated with arbitrary time series fore
casting methods. Extensive experiments on six real-world datasets demonstrate th
at LIFT improves the state-of-the-art methods by 5.5% in average forecasting per

formance.
****
Yuren Cong,Mengmeng Xu,christian simon,Shoufa Chen,Jiawei Ren,Yanping Xie,Juan-Manuel Perez-Rua,Bodo Rosenhahn,Tao Xiang,Sen He
FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing
Text-to-video editing aims to edit the visual appearance of a source video conditional on textual prompts.
A major challenge in this task is to ensure that all frames in the edited video are visually consistent.
Most recent works apply advanced text-to-image diffusion models to this task by inflating 2D spatial attention in the U-Net into spatio-temporal attention.
Although temporal context can be added through spatio-temporal attention, it may introduce some irrelevant information for each patch and therefore cause inconsistency in the edited video.
In this paper, for the first time, we introduce optical flow into the attention module in diffusion model's U-Net to address the inconsistency issue for text-to-video editing.
Our method, FLATTEN, enforces the patches on the same flow path across different frames to attend to each other in the attention module, thus improving the visual consistency in the edited videos.
Additionally, our method is training-free and can be seamlessly integrated into any diffusion based text-to-video editing methods and improve their visual consistency.
Experiment results on existing text-to-video editing benchmarks show that our proposed method achieves the new state-of-the-art performance. In particular, our method excels in maintaining the visual consistency in the edited videos.
****
Jianfa Lai,zhifan Li,Dongming Huang,Qian Lin
The optimality of kernel classifiers in Sobolev space
Kernel methods are widely used in machine learning, especially for classification problems. However, the theoretical analysis of kernel classification is still limited. This paper investigates the statistical performances of kernel classifiers. With some mild assumptions on the conditional probability $\eta(x)=\mathbb{P}(Y=1\mid X=x)$, we derive an upper bound on the classification excess risk of a kernel classifier using recent advances in the theory of kernel regression. We also obtain a minimax lower bound for Sobolev spaces, which shows the optimality of the proposed classifier. Our theoretical results can be extended to the generalization error of overparameterized neural network classifiers. To make our theoretical results more applicable in realistic settings, we also propose a simple method to estimate the interpolation smoothness of $2\eta(x)-1$ and apply the method to real datasets.
****
Dingli Yu,Simran Kaur,Arushi Gupta,Jonah Brown-Cohen,Anirudh Goyal,Sanjeev Arora
SKILL-MIX: a Flexible and Expandable Family of Evaluations for AI Models
With LLMs shifting their role from statistical modeling of language to serving as general-purpose AI agents, how should LLM evaluations change? Arguably, a key ability of an AI agent is to flexibly combine, as needed, the basic skills it has learned. The capability to combine skills plays an important role in (human) pedagogy and also in a paper on emergence phenomena (Arora & Goyal, 2023).

This work introduces SKILL-MIX, a new evaluation to measure ability to combine skills. Using a list of N skills the evaluator repeatedly picks random subsets of k skills and asks the LLM to produce text combining that subset of skills. Since the number of subsets grows like N^k, for even modest k this evaluation will, with high probability, require the LLM to produce text significantly different from any text in the training set.
The paper develops a methodology for (a) designing and administering such an evaluation, and (b) automatic grading (plus spot-checking by humans) of the results using GPT-4 as well as the open LLaMA-2 70B model.

Administering a version of SKILL-MIX to popular chatbots gave results that, whi
le generally in line with prior expectations, contained surprises. Sizeable diff
erences exist among model capabilities that are not captured by their ranking on
 popular LLM leaderboards ("cramming for the leaderboard"). Furthermore, simple
probability calculations indicate that GPT-4's reasonable performance on $k=5$ i
s suggestive of going beyond "stochastic parrot" behavior (Bender et al., 2021),
 i.e., it combines skills in ways that it had not seen during training.

We sketch how the methodology can lead to a SKILL-MIX based eco-system of open e
valuations for AI capabilities of future models. We maintain a leaderboard of SK
ILL-MIX at https://skill-mix.github.io.
****
Aleksandar Petrov,Philip Torr,Adel Bibi
When Do Prompting and Prefix-Tuning Work? A Theory of Capabilities and Limitatio
ns
Context-based fine-tuning methods, including prompting, in-context learning, sof
t prompting (also known as prompt tuning), and prefix-tuning, have gained popula
rity due to their ability to often match the performance of full fine-tuning wit
h a fraction of the parameters. Despite their empirical successes, there is litt
le theoretical understanding of how these techniques influence the internal comp
utation of the model and their expressiveness limitations. We show that despite
the continuous embedding space being more expressive than the discrete token spa
ce, soft-prompting and prefix-tuning are potentially less expressive than full f
ine-tuning, even with the same number of learnable parameters. Concretely, conte
xt-based fine-tuning cannot change the relative attention pattern over the conte
nt and can only bias the outputs of an attention layer in a fixed direction. Thi
s suggests that while techniques like prompting, in-context learning, soft promp
ting, and prefix-tuning can effectively elicit skills present in the pretrained
model, they may not be able to learn novel tasks that require new attention patt
erns.
****
Yong Liu,Tengge Hu,Haoran Zhang,Haixu Wu,Shiyu Wang,Lintao Ma,Mingsheng Long
iTransformer: Inverted Transformers Are Effective for Time Series Forecasting
The recent boom of linear forecasting models questions the ongoing passion for a
rchitectural modifications of Transformer-based forecasters. These forecasters l
everage Transformers to model the global dependencies over temporal tokens of ti
me series, with each token formed by multiple variates of the same timestamp. Ho
wever, Transformers are challenged in forecasting series with larger lookback wi
ndows due to performance degradation and computation explosion. Besides, the emb
edding for each temporal token fuses multiple variates that represent potential
delayed events and distinct physical measurements, which may fail in learning va
riate-centric representations and result in meaningless attention maps. In this
work, we reflect on the competent duties of Transformer components and repurpose
 the Transformer architecture without any modification to the basic components.
We propose iTransformer that simply applies the attention and feed-forward netwo
rk on the inverted dimensions. Specifically, the time points of individual serie
s are embedded into variate tokens which are utilized by the attention mechanism
 to capture multivariate correlations; meanwhile, the feed-forward network is ap
plied for each variate token to learn nonlinear representations. The iTransforme
r model achieves state-of-the-art on challenging real-world datasets, which furt
her empowers the Transformer family with promoted performance, generalization ab
ility across different variates, and better utilization of arbitrary lookback wi
ndows, making it a nice alternative as the fundamental backbone of time series f
orecasting. Code is available at this repository: https://github.com/thuml/iTran
sformer.
****
Zhenbang Wu,Anant Dadu,Nicholas Tustison,Brian Avants,Mike Nalls,Jimeng Sun,Fara
z Faghri
Multimodal Patient Representation Learning with Missing Modalities and Labels
Multimodal patient representation learning aims to integrate information from mu

ltiple modalities and generate comprehensive patient representations for subsequent clinical predictive tasks. However, many existing approaches either presuppose the availability of all modalities and labels for each patient or only deal with missing modalities. In reality, patient data often comes with both missing modalities and labels for various reasons (i.e., the missing modality and label issue). Moreover, multimodal models might over-rely on certain modalities, causing sub-optimal performance when these modalities are absent (i.e., the modality collapse issue). To address these issues, we introduce MUSE: a mutual-consistent graph contrastive learning method. MUSE uses a flexible bipartite graph to represent the patient-modality relationship, which can adapt to various missing modality patterns. To tackle the modality collapse issue, MUSE learns to focus on modality-general and label-decisive features via a mutual-consistent contrastive learning loss. Notably, the unsupervised component of the contrastive objective only requires self-supervision signals, thereby broadening the training scope to incorporate patients with missing labels. We evaluate MUSE on three publicly available datasets: MIMIC-IV, eICU, and ADNI. Results show that MUSE outperforms all baselines, and MUSE+ further elevates the absolute improvement to ~4% by extending the training scope to patients with absent labels.
****
Keita Suzuki,Taiji Suzuki
Optimal criterion for feature learning of two-layer linear neural network in high dimensional interpolation regime
Deep neural networks with feature learning have shown surprising generalization performance in high dimensional settings, but it has not been fully understood how and when they enjoy the benefit of feature learning. In this paper, we theoretically analyze the statistical properties of the benefits from feature learning in a two-layer linear neural network with multiple outputs in a high-dimensional setting. For that purpose, we propose a new criterion that allows feature learning of a two-layer linear neural network in a high-dimensional setting. Interestingly, we can show that models with smaller values of the criterion generalize even in situations where normal ridge regression fails to generalize. This is because the proposed criterion contains a proper regularization for the feature mapping and acts as an upper bound on the predictive risk. As an important characterization of the criterion, the two-layer linear neural network that minimizes this criterion can achieve the optimal Bayes risk that is determined by the distribution of the true signals across the multiple outputs. To the best of our knowledge, this is the first study to specifically identify the conditions under which a model obtained by proper feature learning can outperform normal ridge regression in a high-dimensional multiple-output linear regression problem.
****
Heejun Lee,Jina Kim,Jeffrey Willette,Sung Ju Hwang
SEA: Sparse Linear Attention with Estimated Attention Mask
The transformer architecture has driven breakthroughs in recent years on tasks which require modeling pairwise relationships between sequential elements, as is the case in natural language understanding. However, long seqeuences pose a problem due to the quadratic complexity of the attention operation. Previous research has aimed to lower the complexity by sparsifying or linearly approximating the attention matrix. Yet, these approaches cannot straightforwardly distill knowledge from a teacher’s attention matrix, and often require complete retraining from scratch. Furthermore, previous sparse and linear approaches lose interpretability if they cannot produce full attention matrices. To address these challenges, we propose SEA: Sparse linear attention with an Estimated Attention mask. SEA estimates the attention matrix with linear complexity via kernel-based linear attention, then subsequently creates a sparse attention matrix with a top-k■ selection

to perform a sparse attention operation. For language modeling tasks (Wikitext2)
,
previous linear and sparse attention methods show roughly two-fold worse per-
plexity scores over the quadratic OPT-1.3B baseline, while SEA achieves better
perplexity than OPT-1.3B, using roughly half the memory of OPT-1.3B. More-
over, SEA maintains an interpretable attention matrix and can utilize knowledge
distillation to lower the complexity of existing pretrained transformers. We be-
lieve that our work will have a large practical impact, as it opens the possibil
ity of
running large transformers on resource-limited devices with less memory.
****
Lean Wang,Wenkai Yang,Deli Chen,Hao Zhou,Yankai Lin,Fandong Meng,Jie Zhou,Xu Sun
Towards Codable Watermarking for Injecting Multi-Bits Information to LLMs
As large language models (LLMs) generate texts with increasing fluency and reali
sm, there is a growing need to identify the source of texts to prevent the abuse
 of LLMs. Text watermarking techniques have proven reliable in distinguishing wh
ether a text is generated by LLMs by injecting hidden patterns. However, we argu
e that existing LLM watermarking methods are encoding-inefficient and cannot fle
xibly meet the diverse information encoding needs (such as encoding model versio
n, generation time, user id, etc.). In this work, we conduct the first systemati
c study on the topic of **Codable Text Watermarking for LLMs** (CTWL) that allow
s text watermarks to carry multi-bit customizable information. First of all, we
study the taxonomy of LLM watermarking technologies and give a mathematical form
ulation for CTWL. Additionally, we provide a comprehensive evaluation system for
 CTWL: (1) watermarking success rate, (2) robustness against various corruptions
, (3) coding rate of payload information, (4) encoding and decoding efficiency,
(5) impacts on the quality of the generated text. To meet the requirements of th
ese non-Pareto-improving metrics, we follow the most prominent vocabulary partit
ion-based watermarking direction, and devise an advanced CTWL method named **Bal
ance-Marking**. The core idea of our method is to use a proxy language model to
split the vocabulary into probability-balanced parts, thereby effectively mainta
ining the quality of the watermarked text. Our code is available at https://gith
ub.com/lancopku/codable-watermarking-for-llm.
****
Samuel Sokota,Gabriele Farina,David J Wu,Hengyuan Hu,Kevin A. Wang,J Zico Kolter
,Noam Brown
The Update-Equivalence Framework for Decision-Time Planning
The process of revising (or constructing) a policy at execution time---known as
decision-time planning---has been key to achieving superhuman performance in per
fect-information games like chess and Go. A recent line of work has extended dec
ision-time planning to imperfect-information games, leading to superhuman perfor
mance in poker. However, these methods involve solving subgames whose sizes grow
 quickly in the amount of non-public information, making them unhelpful when the
 amount of non-public information is large.
Motivated by this issue, we introduce an alternative framework for decision-time
 planning that is not based on solving subgames, but rather on update equivalenc
e. In this update-equivalence framework, decision-time planning algorithms repli
cate the updates of last-iterate algorithms, which need not rely on public infor
mation. This facilitates scalability to games with large amounts of non-public i
nformation. Using this framework, we derive a provably sound search algorithm fo
r fully cooperative games based on mirror descent and a search algorithm for adv
ersarial games based on magnetic mirror descent. We validate the performance of
these algorithms in cooperative and adversarial domains, notably in Hanabi, the
standard benchmark for search in fully cooperative imperfect-information games.
Here, our mirror descent approach exceeds or matches the performance of public i
nformation-based search while using two orders of magnitude less search time. Th
is is the first instance of a non-public-information-based algorithm outperformi
ng public-information-based approaches in a domain they have historically domina
ted.
****

Jiecheng Lu,Xu Han,Shihao Yang
ARM: Refining Multivariate Forecasting with Adaptive Temporal-Contextual Learning
Long-term time series forecasting (LTSF) is important for various domains but is confronted by challenges in handling the complex temporal-contextual relationships. As multivariate input models underperforming some recent univariate counterparts, we posit that the issue lies in the inefficiency of existing multivariate LTSF Transformers to model series-wise relationships: the characteristic differences between series are often captured incorrectly. To address this, we introduce ARM: a multivariate temporal-contextual adaptive learning method, which is an enhanced architecture specifically designed for multivariate LTSF modelling. ARM employs Adaptive Univariate Effect Learning (**A**UEL), Random Dropping (**R**D) training strategy, and Multi-kernel Local Smoothing (**M**KLS), to better handle individual series temporal patterns and correctly learn inter-series dependencies. ARM demonstrates superior performance on multiple benchmarks without significantly increasing computational costs compared to vanilla Transformer, thereby advancing the state-of-the-art in LTSF. ARM is also generally applicable to other LTSF architecture beyond vanilla Transformer.
****
Kethmi Hirushini Hettige,Jiahao Ji,Shili Xiang,Cheng Long,Gao Cong,Jingyuan Wang
AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction
Air quality prediction and modelling plays a pivotal role in public health and environment management, for individuals and authorities to make informed decisions. Although traditional data-driven models have shown promise in this domain, their long-term prediction accuracy can be limited, especially in scenarios with sparse or incomplete data and they often rely on black-box deep learning structures that lack solid physical foundation leading to reduced transparency and interpretability in predictions. To address these limitations, this paper presents a novel approach named Physics guided Neural Network for Air Quality Prediction (AirPhyNet). Specifically, we leverage two well-established physics principles of air particle movement (diffusion and advection) by representing them as differential equation networks. Then, we utilize a graph structure to integrate physics knowledge into a neural network architecture and exploit latent representations to capture spatio-temporal relationships within the air quality data. Experiments on two real-world benchmark datasets demonstrate that AirPhyNet outperforms state-of-the-art models for different testing scenarios including different lead time (24h, 48h, 72h), sparse data and sudden change prediction, achieving reduction in prediction errors up to 10\%. Moreover, a case study further validates that our model captures underlying physical processes of particle movement and generates accurate predictions with real physical meaning. The code is available at: https://github.com/kethmih/AirPhyNet
****
Chenjie Mao,Qiaosheng Zhang,Zhen Wang,Xuelong Li
On the Role of General Function Approximation in Offline Reinforcement Learning
We study offline reinforcement learning (RL) with general function approximation. General function approximation is a powerful tool for algorithm design and analysis, but its adaptation to offline RL encounters several challenges due to varying approximation targets and assumptions that blur the real meanings of function assumptions. In this paper, we try to formulate and clarify the treatment of general function approximation in offline RL in two aspects: (1) analyzing different types of assumptions and their practical usage, and (2) understanding its role as a restriction on underlying MDPs from information-theoretic perspectives. Additionally, we introduce a new insight for lower bound establishing: one can exploit model-realizability to establish general-purposed lower bounds that can be generalized into other functions. Building upon this insight, we propose two generic lower bounds that contribute to a better understanding of offline RL with general function approximation.
****
Quentin Bertrand,Joey Bose,Alexandre Duplessis,Marco Jiralerspong,Gauthier Gidel
On the Stability of Iterative Retraining of Generative Models on their own Data

Deep generative models have made tremendous progress in modeling complex data, o
ften exhibiting generation quality that surpasses a typical human's ability to d
iscern the authenticity of samples. Undeniably, a key driver of this success is
enabled by the massive amounts of web-scale data consumed by these models. Due t
o these models' striking performance and ease of availability, the web will inev
itably be increasingly populated with synthetic content. Such a fact directly im
plies that future iterations of generative models will be trained on both clean
and artificially generated data from past models. In this paper, we develop a fr
amework to rigorously study the impact of training generative models on mixed da
tasets---from classical training on real data to self-consuming generative model
s trained on purely synthetic data. We first prove the stability of iterative tr
aining under the condition that the initial generative models approximate the da
ta distribution well enough and the proportion of clean training data (w.r.t. sy
nthetic data) is large enough. We empirically validate our theory on both synthe
tic and natural images by iteratively training normalizing flows and state-of-th
e-art diffusion models on CIFAR10 and FFHQ.
****
Xiaoran Liu,Hang Yan,Chenxin An,Xipeng Qiu,Dahua Lin
Scaling Laws of RoPE-based Extrapolation
The extrapolation capability of Large Language Models (LLMs) based on Rotary Pos
ition Embedding \citep{su2021roformer} is currently a topic of considerable inte
rest. The mainstream approach to addressing extrapolation with LLMs involves mod
ifying RoPE by replacing 10000, the rotary base of $\theta_n={10000}^{-2n/d}$ in
 the original RoPE, with a larger value and providing longer fine-tuning text. I
n this work, we first observe that fine-tuning a RoPE-based LLM with either a sm
aller or larger base in pre-training context length could significantly enhance
its extrapolation performance. After that, we propose \textbf{\textit{Scaling La
ws of RoPE-based Extrapolation}}, a unified framework from the periodic perspect
ive, to describe the relationship between the extrapolation performance and base
 value as well as tuning context length. In this process, we also explain the or
igin of the RoPE-based extrapolation issue by \textbf{\textit{critical dimension
 for extrapolation}}. Besides these observations and analyses, we achieve extrap
olation up to 1 million context length within only 16K training length on LLaMA2
 7B and 13B \citep{touvron2023llama2}.
****
Yuka Hashimoto,Sho Sonoda,Isao Ishikawa,Atsushi Nitanda,Taiji Suzuki
Koopman-based generalization bound: New aspect for full-rank weights
We propose a new bound for generalization of neural networks using Koopman opera
tors. Whereas most of existing works focus on low-rank weight matrices, we focus
 on full-rank weight matrices. Our bound is tighter than existing norm-based bou
nds when the condition numbers of weight matrices are small. Especially, it is c
ompletely independent of the width of the network if the weight matrices are ort
hogonal. Our bound does not contradict to the existing bounds but is a complemen
t to the existing bounds. As supported by several existing empirical results, lo
w-rankness is not the only reason for generalization. Furthermore, our bound can
 be combined with the existing bounds to obtain a tighter bound. Our result shed
s new light on understanding generalization of neural networks with full-rank we
ight matrices, and it provides a connection between operator-theoretic analysis
and generalization of neural networks.
****
Ke Xue,Ren-Jian Wang,Pengyi Li,Dong Li,Jianye HAO,Chao Qian
Sample-Efficient Quality-Diversity by Cooperative Coevolution
Quality-Diversity (QD) algorithms, as a subset of evolutionary algorithms, have
emerged as a powerful optimization paradigm with the aim of generating a set of
high-quality and diverse solutions. Although QD has demonstrated competitive per
formance in reinforcement learning, its low sample efficiency remains a signific
ant impediment for real-world applications. Recent research has primarily focuse
d on augmenting sample efficiency by refining selection and variation operators
of QD. However, one of the less considered yet crucial factors is the inherently
 large-scale issue of the QD optimization problem. In this paper, we propose a n

ovel Cooperative Coevolution QD (CCQD) framework, which decomposes a policy network naturally into two types of layers, corresponding to representation and decision respectively, and thus simplifies the problem significantly. The resulting two (representation and decision) subpopulations are coevolved cooperatively. CCQD can be implemented with different selection and variation operators. Experiments on several popular tasks within the QDAX suite demonstrate that an instantiation of CCQD achieves approximately a 200% improvement in sample efficiency.
****
Vijay Lingam,Mohammad Sadegh Akhondzadeh,Aleksandar Bojchevski
Rethinking Label Poisoning for GNNs: Pitfalls and Attacks
Node labels for graphs are usually generated using an automated process or crowd-sourced from human users. This opens up avenues for malicious users to compromise the training labels, making it unwise to blindly rely on them. While robustness against noisy labels is an active area of research, there are only a handful of papers in the literature that address this for graph-based data. Even more so, the effects of adversarial label perturbations is sparsely studied. More critically, we reveal that the entire literature on label poisoning for GNNs is plagued by serious evaluation pitfalls. Thus making it hard to conclude how robust GNNs are against label perturbations. After course correcting the state of label poisoning attacks with our faithful evaluation, we identify a discrepancy in attack efficiency of $\sim9\%$ on average. Additionally, we introduce two new simple yet effective attacks that are significantly stronger (up to $\sim8\%$) than the previous strongest attack. Our strongest proposed attack can be efficiently computed and is theoretically backed.
****
Fuxiao Liu,Kevin Lin,Linjie Li,Jianfeng Wang,Yaser Yacoob,Lijuan Wang
Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning
Despite the promising progress in multi-modal tasks, current large multi-modal models (LMMs) are prone to hallucinating inconsistent descriptions with respect to the associated image and human instructions. This paper addresses this issue by introducing the first large and diverse visual instruction tuning dataset, named Large-scale Robust Visual (LRV)-Instruction. Our dataset comprises 400k visual
instructions generated by GPT4, covering 16 vision-and-language tasks with open-ended instructions and answers. Unlike existing studies that primarily focus on positive instruction samples, we design LRV-Instruction to include both positive and negative instructions for more robust visual instruction tuning. Our negative instructions are designed at three semantic levels: (i) Nonexistent Object Manipulation, (ii) Existent Object Manipulation and (iii) Knowledge Manipulation. To efficiently measure the hallucination generated by LMMs, we propose GPT4-Assisted Visual Instruction Evaluation (GAVIE), a stable approach to evaluate visual instruction tuning like human experts. GAVIE does not require human-annotated groundtruth answers and can adapt to diverse instruction formats. We conduct comprehensive experiments to investigate the hallucination of LMMs. Our results demonstrate existing LMMs exhibit significant hallucinations when presented with our negative instructions, particularly Existent Object and Knowledge Manipulation instructions. Moreover, we successfully mitigate hallucination by finetuning MiniGPT4 and mPLUG-Owl on LRV-Instruction while improving performance on several public
datasets compared to state-of-the-art methods. Additionally, we observed that a balanced ratio of positive and negative instances in the training data leads to a more robust model. Code and data will be released upon publication.
****
Stephen Marcus McAleer,JB Lanier,Kevin A. Wang,Pierre Baldi,Tuomas Sandholm,Roy Fox
Toward Optimal Policy Population Growth in Two-Player Zero-Sum Games
In competitive two-agent environments, deep reinforcement learning (RL) methods like Policy Space Response Oracles (PSRO) often increase exploitability between iterations, which is problematic when training in large games. To address this i

ssue, we introduce anytime double oracle (ADO), an algorithm that ensures exploi
tability does not increase between iterations, and its approximate extensive-for
m version, anytime PSRO (APSRO). ADO converges to a Nash equilibrium while itera
tively reducing exploitability. However, convergence in these algorithms may req
uire adding all of a game's deterministic policies. To improve this, we propose
Self-Play PSRO (SP-PSRO), which incorporates an approximately optimal stochastic
 policy into the population in each iteration. APSRO and SP-PSRO demonstrate low
er exploitability and near-monotonic exploitability reduction in games like Ledu
c poker and Liar's Dice. Empirically, SP-PSRO often converges much faster than A
PSRO and PSRO, requiring only a few iterations in many games.
****
Yuhao Huang,Qingsong Wang,Akwum Onwunta,Bao Wang
Efficient Score Matching with Deep Equilibrium Layers
Score matching methods -- estimate probability densities without computing the n
ormalization constant -- are particularly useful in deep learning. However, comp
utational and memory costs of score matching methods can be prohibitive for high
-dimensional data or complex models, particularly due to the derivatives or Hess
ians of the log density function appearing in the objective function. Some exist
ing approaches modify the objective function to reduce the quadratic computation
al complexity for Hessian computation. However, the memory bottleneck of score m
atching methods remains for deep learning. This study improves the memory effici
ency of score matching by leveraging deep equilibrium models. We provide a theor
etical analysis of deep equilibrium models for scoring matching and applying imp
licit differentiation to higher-order derivatives. Empirical evaluations demonst
rate that our approach enables the development of deep and expressive models wit
h improved performance and comparable computational and memory costs over shallo
w architectures.
****
Qinzi Zhang,Hoang Tran,Ashok Cutkosky
Private Zeroth-Order Nonsmooth Nonconvex Optimization
We introduce a new zeroth-order algorithm for private stochastic optimization on
 nonconvex and nonsmooth objectives.
Given a dataset of size M, our algorithm ensures $(\alpha,\alpha\rho^2/2)$-Ren
yi differential privacy and finds a (δ,ϵ)-stationary point so long
 as $M=\tilde\Omega(\frac{d}{\delta\epsilon^3} + \frac{d^{3/2}}{\rho\delta\epsil
on^2})$.
This matches the optimal complexity found in its non-private zeroth-order analog
.
Notably, although the objective is not smooth, we have privacy ``for free'' when
 $\rho \ge \sqrt{d}\epsilon$.
****
Bao Nguyen,Binh Nguyen,Viet Anh Nguyen
Bellman Optimal Stepsize Straightening of Flow-Matching Models
Flow matching is a powerful framework for generating high-quality samples in var
ious applications, especially image synthesis. However, the intensive computatio
nal demands of these models, especially during the finetuning process and sampli
ng processes, pose significant challenges for low-resource scenarios. This paper
 introduces Bellman Optimal Stepsize Straightening (BOSS) technique for distilli
ng flow-matching generative models: it aims specifically for a few-step efficien
t image sampling while adhering to a computational budget constraint. First, thi
s technique involves a dynamic programming algorithm that optimizes the stepsize
s of the pretrained network. Then, it refines the velocity network to match the
optimal step sizes, aiming to straighten the generation paths. Extensive experim
ental evaluations across image generation tasks demonstrate the efficacy of BOSS
 in terms of both resource utilization and image quality. Our results reveal tha
t BOSS achieves substantial gains in efficiency while maintaining competitive sa
mple quality, effectively bridging the gap between low-resource constraints and
the demanding requirements of flow-matching generative models. Our paper also fo
rtifies the responsible development of artificial intelligence, offering a more
sustainable generative model that reduces computational costs and environmental

footprints. Our code can be found at https://github.com/nguyenngocbaocmt02/BOSS.
****
Kashif Rasul,Andrew Bennett,Pablo Vicente,Umang Gupta,Hena Ghonia,Anderson Schneider,Yuriy Nevmyvaka
VQ-TR: Vector Quantized Attention for Time Series Forecasting
Probabilistic time series forecasting is a challenging problem due to the long sequences involved, the large number of samples needed for accurate probabilistic inference, and the need for real-time inference in many applications. These challenges necessitate methods that are not only accurate but computationally efficient. Unfortunately, most current state-of-the-art methods for time series forecasting are based on Transformers, which scale poorly due to quadratic complexity in sequence length, and are therefore needlessly computationally inefficient. Moreover, with a few exceptions, these methods have only been evaluated for non-probabilistic point estimation. In this work, we address these two shortcomings. For the first, we introduce VQ-TR, which maps large sequences to a discrete set of latent representations as part of the Attention module. This not only allows us to attend over larger context windows with linear complexity in sequence length but also allows for effective regularization to avoid overfitting.
For the second, we provide what is to the best of our knowledge the first systematic comparison of modern Transformer-based time series forecasting methods for probabilistic forecasting. In this comparison, we find that VQ-TR performs better or comparably to all other methods while being computationally efficient.
****
Ayan Sengupta,Shantanu Dixit,Md Shad Akhtar,Tanmoy Chakraborty
A Good Learner can Teach Better: Teacher-Student Collaborative Knowledge Distillation
Knowledge distillation (KD) is a technique used to transfer knowledge from a larger ''teacher'' model into a smaller ''student'' model. Recent advancements in meta-learning-based knowledge distillation (MetaKD) emphasize that the fine-tuning of teacher models should be aware of the student's need to achieve better knowledge distillation. However, existing MetaKD methods often lack incentives for the teacher model to improve itself. In this study, we introduce MPDistil, a meta-policy distillation technique, that utilizes novel optimization strategies to foster both *collaboration* and *competition* during the fine-tuning of the teacher model in the meta-learning step. Additionally, we propose a curriculum learning framework for the student model in a competitive setup, in which the student model aims to outperform the teacher model by self-training on various tasks. Exhaustive experiments on SuperGLUE and GLUE benchmarks demonstrate the efficacy of MPDistil compared to 20 conventional KD and advanced MetaKD baselines, showing significant performance enhancements in the student model -- e.g., a distilled 6-layer BERT model outperforms a 12-layer BERT model on five out of six SuperGLUE tasks. Furthermore, MPDistil, while applied to a large language teacher model (DeBERTa-v2-xxlarge), significantly narrows the performance gap of its smaller student counterpart (DeBERTa-12) by just 4.6% on SuperGLUE. We further demonstrate how higher rewards and customized training curricula strengthen the student model and enhance generalizability.
****
Bahare Fatemi,Jonathan Halcrow,Bryan Perozzi
Talk like a Graph: Encoding Graphs for Large Language Models
Graphs are a powerful tool for representing and analyzing complex relationships in real-world applications such as social networks, recommender systems, and computational finance. Reasoning on graphs is essential for drawing inferences about the relationships between entities in a complex system, and to identify hidden patterns and trends. Despite the remarkable progress in automated reasoning with natural text, reasoning on graphs with large language models (LLMs) remains an understudied problem. In this work, we perform the first comprehensive study of encoding graph-structured data as text for consumption by LLMs. We show that LLM performance on graph reasoning tasks varies on three fundamental levels: (1) the graph encoding method, (2) the nature of the graph task itself, and (3) interestingly, the very structure of the graph considered. These novel results provid

e valuable insight on strategies for encoding graphs as text. Using these insigh
ts we illustrate how the correct choice of encoders can boost performance on gra
ph reasoning tasks inside LLMs by 4.8% to 61.8%, depending on the task.
****
Jiaxin Cheng,Tianjun Xiao,Tong He
Consistent Video-to-Video Transfer Using Synthetic Dataset
We introduce a novel and efficient approach for text-based video-to-video editin
g that eliminates the need for resource-intensive per-video-per-model finetuning
. At the core of our approach is a synthetic paired video dataset tailored for v
ideo-to-video transfer tasks. Inspired by Instruct Pix2Pix's image transfer via
editing instruction, we adapt this paradigm to the video domain. Extending the P
rompt-to-Prompt to videos, we efficiently generate paired samples, each with an
input video and its edited counterpart. Alongside this, we introduce the Long Vi
deo Sampling Correction during sampling, ensuring consistent long videos across
batches. Our method surpasses current methods like Tune-A-Video, heralding subst
antial progress in text-based video-to-video editing and suggesting exciting ave
nues for further exploration and deployment.
****
Jie Huang,Xinyun Chen,Swaroop Mishra,Huaixiu Steven Zheng,Adams Wei Yu,Xinying S
ong,Denny Zhou
Large Language Models Cannot Self-Correct Reasoning Yet
Large Language Models (LLMs) have emerged as a groundbreaking technology with th
eir unparalleled text generation capabilities across various applications. Never
theless, concerns persist regarding the accuracy and appropriateness of their ge
nerated content. A contemporary methodology, self-correction, has been proposed
as a remedy to these issues. Building upon this premise, this paper critically e
xamines the role and efficacy of self-correction within LLMs, shedding light on
its true potential and limitations. Central to our investigation is the notion o
f intrinsic self-correction, whereby an LLM attempts to correct its initial resp
onses based solely on its inherent capabilities, without the crutch of external
feedback. In the context of reasoning, our research indicates that LLMs struggle
 to self-correct their responses without external feedback, and at times, their
performance even degrades after self-correction. Drawing from these insights, we
 offer suggestions for future research and practical applications in this field.
****
Peter Müller,Lukas Faber,Karolis Martinkus,Roger Wattenhofer
GraphChef: Decision-Tree Recipes to Explain Graph Neural Networks
We propose a new self-explainable Graph Neural Network (GNN) model: GraphChef. G
raphChef integrates decision trees into the GNN message passing framework. Given
 a dataset, GraphChef returns a set of rules (a recipe) that explains each class
 in the dataset unlike existing GNNs and explanation methods that reason on indi
vidual graphs. Thanks to the decision trees, GraphChef recipes are human unders
tandable. We also present a new pruning method to produce small and easy to dig
est trees. Experiments demonstrate that GraphChef reaches comparable accuracy to
 not self-explainable GNNs and produced decision trees are indeed small. We furt
her validate the correctness of the discovered recipes on datasets where explana
tion ground truth is available: Reddit-Binary, MUTAG, BA-2Motifs, BA-Shapes, Tre
e-Cycle, and Tree-Grid.
****
Junyuan Hong,Jiachen T. Wang,Chenhui Zhang,Zhangheng LI,Bo Li,Zhangyang Wang
DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer
Large Language Models (LLMs) have emerged as dominant tools for various tasks, p
articularly when tailored for a specific target by prompt tuning. Nevertheless,
concerns surrounding data privacy present obstacles due to the tuned prompts' de
pendency on sensitive private information. A practical solution is to host a loc
al LLM and optimize a soft prompt privately using data. Yet, hosting a local mod
el becomes problematic when model ownership is protected. Alternative methods, l
ike sending data to the model's provider for training, intensify these privacy i
ssues facing an untrusted provider. In this paper, we present a novel solution c
alled Differentially-Private Offsite Prompt Tuning (DP-OPT) to address this chal

lenge. Our approach involves tuning a discrete prompt on the client side and the
n applying it to the desired cloud models. We demonstrate that prompts suggested
 by LLMs themselves can be transferred without compromising performance signific
antly. To ensure that the prompts do not leak private information, we introduce
the first private prompt generation mechanism, by a differentially-private (DP)
ensemble of in-context learning with private demonstrations. With DP-OPT, gener
ating privacy-preserving prompts by Vicuna-7b can yield competitive performance
compared to non-private in-context learning on GPT3.5 or local private prompt tu
ning.
Codes are available at https://github.com/VITA-Group/DP-OPT.
****
Noam Razin,Hattie Zhou,Omid Saremi,Vimal Thilak,Arwen Bradley,Preetum Nakkiran,J
oshua M. Susskind,Etai Littwin
Vanishing Gradients in Reinforcement Finetuning of Language Models
Pretrained language models are commonly aligned with human preferences and downs
tream tasks via reinforcement finetuning (RFT), which refers to maximizing a (po
ssibly learned) reward function using policy gradient algorithms. This work iden
tifies a fundamental optimization obstacle in RFT: we prove that the expected gr
adient for an input vanishes when its reward standard deviation under the model
is small, even if the expected reward is far from optimal. Through experiments o
n an RFT benchmark and controlled environments, as well as a theoretical analysi
s, we then demonstrate that vanishing gradients due to small reward standard dev
iation are prevalent and detrimental, leading to extremely slow reward maximizat
ion. Lastly, we explore ways to overcome vanishing gradients in RFT. We find the
 common practice of an initial supervised finetuning (SFT) phase to be the most
promising candidate, which sheds light on its importance in an RFT pipeline. Mor
eover, we show that a relatively small number of SFT optimization steps on as fe
w as 1% of the input samples can suffice, indicating that the initial SFT phase
need not be expensive in terms of compute and data labeling efforts. Overall, ou
r results emphasize that being mindful for inputs whose expected gradient vanish
es, as measured by the reward standard deviation, is crucial for successful exec
ution of RFT.
****
Fan Shi,Bin Li,Xiangyang Xue
Towards Generative Abstract Reasoning: Completing Raven’s Progressive Matrix via
 Rule Abstraction and Selection
Endowing machines with abstract reasoning ability has been a long-term research
topic in artificial intelligence. Raven's Progressive Matrix (RPM) is widely use
d to probe abstract visual reasoning in machine intelligence, where models will
analyze the underlying rules and select one image from candidates to complete th
e image matrix. Participators of RPM tests can show powerful reasoning ability b
y inferring and combining attribute-changing rules and imagining the missing ima
ges at arbitrary positions of a matrix. However, existing solvers can hardly man
ifest such an ability in realistic RPM tests. In this paper, we propose a deep l
atent variable model for answer generation problems through Rule AbstractIon and
 SElection (RAISE). RAISE can encode image attributes into latent concepts and a
bstract atomic rules that act on the latent concepts. When generating answers, R
AISE selects one atomic rule out of the global knowledge set for each latent con
cept to constitute the underlying rule of an RPM. In the experiments of bottom-r
ight and arbitrary-position answer generation, RAISE outperforms the compared so
lvers in most configurations of realistic RPM datasets. In the odd-one-out task
and two held-out configurations, RAISE can leverage acquired latent concepts and
 atomic rules to find the rule-breaking image in a matrix and handle problems wi
th unseen combinations of rules and attributes.
****
Junzhe Zhu,Peiye Zhuang,Sanmi Koyejo
HIFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance
The advancements in automatic text-to-3D generation have been remarkable. Most e
xisting methods use pre-trained text-to-image diffusion models to optimize 3D re
presentations like Neural Radiance Fields (NeRFs) via latent-space denoising sco

re matching. Yet, these methods often result in artifacts and inconsistencies ac
ross different views due to their suboptimal optimization approaches and limited
 understanding of 3D geometry. Moreover, the inherent constraints of NeRFs in re
ndering crisp geometry and stable textures usually lead to a two-stage optimizat
ion to attain high-resolution details. This work proposes holistic sampling and
smoothing approaches to achieve high-quality text-to-3D generation, all in a sin
gle-stage optimization. We compute denoising scores in the text-to-image diffusi
on model's latent and image spaces. Instead of randomly sampling timesteps (also
 referred to as noise levels in denoising score matching), we introduce a novel
timestep annealing approach that progressively reduces the sampled timestep thro
ughout optimization. To generate high-quality renderings in a single-stage optim
ization, we propose regularization for the variance of z-coordinates along NeRF
rays. To address texture flickering issues in NeRFs, we introduce a kernel smoot
hing technique that refines importance sampling weights coarse-to-fine, ensuring
 accurate and thorough sampling in high-density regions. Extensive experiments d
emonstrate the superiority of our method over previous approaches, enabling the
generation of highly detailed and view-consistent 3D assets through a single-sta
ge training process.
****
Atsushi Shimizu,Xiaoou Cheng,Christopher Musco,Jonathan Weare
Improved Active Learning via Dependent Leverage Score Sampling
We show how to obtain improved active learning methods in the agnostic (adversar
ial noise) setting by combining marginal leverage score sampling with non-indepe
ndent sampling strategies that promote spatial coverage. In particular, we propo
se an easily implemented method based on the \emph{pivotal sampling algorithm},
which we test on problems motivated by learning-based methods for parametric PDE
s and uncertainty quantification. In comparison to independent sampling, our met
hod reduces the number of samples needed to reach a given target accuracy by up
to 50%.

We support our findings with two theoretical results. First, we show that any no
n-independent leverage score sampling method that obeys a weak \emph{one-sided $
\ell_{\infty}$ independence condition} (which includes pivotal sampling) can act
ively learn d dimensional linear functions with $O(d\log d)$ samples, matching
 independent sampling. This result extends recent work on matrix Chernoff bounds
 under ℓ_{∞} independence, and may be of interest for analyzing other
sampling strategies beyond pivotal sampling. Second, we show that, for the impor
tant case of polynomial regression, our pivotal method obtains an improved bound
 of $O(d)$ samples.
****
Jaemin Cho,Yushi Hu,Jason Michael Baldridge,Roopal Garg,Peter Anderson,Ranjay Kr
ishna,Mohit Bansal,Jordi Pont-Tuset,Su Wang
Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Te
xt-to-Image Generation
Evaluating text-to-image models is notoriously difficult. A strong recent approa
ch for assessing text-image faithfulness is based on QG/A (question generation a
nd answering), which uses pre-trained foundational models to automatically gener
ate a set of questions and answers from the prompt, and output images are scored
 based on whether these answers extracted with a visual question answering model
 are consistent with the prompt-based answers. This kind of evaluation is natura
lly dependent on the quality of the underlying QG and VQA models. We identify an
d address several reliability challenges in existing QG/A work: (a) QG questions
 should respect the prompt (avoiding hallucinations, duplications, and omissions
) and (b) VQA answers should be consistent (not asserting that there is no motor
cycle in an image while also claiming the motorcycle is blue). We address these
issues with Davidsonian Scene Graph (DSG), an empirically grounded evaluation fr
amework inspired by formal semantics, which is adaptable to any QG/A frameworks.
 DSG produces atomic and unique questions organized in dependency graphs, which
(i) ensure appropriate semantic coverage and (ii) sidestep inconsistent answers.
 With extensive experimentation and human evaluation on a range of model configu

rations (LLM, VQA, and T2I), we empirically demonstrate that DSG addresses the c
hallenges noted above. Finally, we present DSG-1k, an open-sourced evaluation be
nchmark that includes 1,060 prompts, covering a wide range of fine-grained seman
tic categories with a balanced distribution. We release the DSG-1k prompts and t
he corresponding DSG questions.
****
Tsung-Wei Ke,Sangwoo Mo,Stella X. Yu
Learning Hierarchical Image Segmentation For Recognition and By Recognition
Image segmentation and recognition occur simultaneously, with recognition relyin
g on the underlying segmentation to form a continuous visual grouping hierarchy.
 For example, the same object can be parsed into different part-to-whole structu
res, resulting in varying recognitions. Despite this, most prior works treated s
egmentation and recognition as separate tasks. In this paper, we aim to devise a
 learning framework that involves segmentation in the recognition process, utili
zing hierarchical segmentation for recognition, which is learned by recognition.
 Specifically, we propose CAST, which realizes this concept through designs insp
ired by vision transformers, enabling concurrent segmentation and recognition wi
th a single model. The core idea of CAST is to employ adaptive segment tokens th
at group the finest pixels into coarser segments, using the latest embedding to
represent the entire image for recognition. Trained solely on image recognition
objectives, CAST automatically discovers the hierarchy of segments. Our experime
nts demonstrate that CAST provides consistent hierarchical segmentation and reco
gnition, which is impossible with state-of-the-art segmentation methods such as
SAM. Additionally, CAST offers several advantages over the standard ViT, includi
ng improved semantic segmentation, computational efficiency, and object-centric
attention.
****
Daniel Bolya,Chaitanya Ryali,Judy Hoffman,Christoph Feichtenhofer
Window Attention is Bugged: How not to Interpolate Position Embeddings
Window attention, position embeddings, and high resolution finetuning are core c
oncepts in the modern transformer era of computer vision. However, we find that
naively combining these near ubiquitous components can have a detrimental effect
 on performance. The issue is simple: interpolating position embeddings while us
ing window attention is wrong. We study two state-of-the-art methods that have t
hese three components, namely Hiera and ViTDet, and find that both do indeed suf
fer from this bug. To fix it, we introduce a simple absolute window position emb
edding strategy, which solves the bug outright in Hiera and allows us to increas
e both speed and performance of the model in ViTDet. We finally combine the two
to obtain HieraDet, which achieves 61.7 box mAP on COCO, making it state-of-the-
art for models that only use ImageNet-1k pretraining. This all stems from what i
s essentially a 3 line bug fix, which we name "absolute win".
****
Dong HUANG,Qingwen Bu
Adversarial Feature Map Pruning for Backdoor
Deep neural networks have been widely used in many critical applications, such a
s autonomous vehicles and medical diagnosis. However, their security is threaten
ed by backdoor attacks, which are achieved by adding artificial patterns to spec
ific training data. Existing defense strategies primarily focus on using reverse
 engineering to reproduce the backdoor trigger generated by attackers and subseq
uently repair the DNN model by adding the trigger into inputs and fine-tuning th
e model with ground truth labels. However, once the trigger generated by the att
ackers is complex and invisible, the defender cannot reproduce the trigger succe
ssfully then the DNN model will not be repaired, as the trigger is not effective
ly removed.

In this work, we propose Adversarial Feature Map Pruning for Backdoor (FMP) to m
itigate backdoor from the DNN. Unlike existing defense strategies, which focus o
n reproducing backdoor triggers, FMP attempts to prune backdoor feature maps, wh
ich are trained to extract backdoor information from inputs. After pruning these
 backdoor feature maps, FMP will fine-tune the model with a secure subset of tra

ining data. Our experiments demonstrate that, compared to existing defense strategies, FMP can effectively reduce the Attack Success Rate (ASR) even against the most complex and invisible attack triggers (e.g., FMP decreases the ASR to 2.86% in CIFAR10, which is 19.2% to 65.41% lower than baselines). Second, unlike conventional defense methods that tend to exhibit low robust accuracy (that is, the accuracy of the model on poisoned data), FMP achieves a higher RA, indicating its superiority in maintaining model performance while mitigating the effects of backdoor attacks (e.g., FMP obtains 87.40% RA in CIFAR10). Our code is publicly available at: https://github.com/hku-systems/FMP.
****
Hugo Lebeau,Mohamed El Amine Seddik,José Henrique De Morais Goulart
Performance Gaps in Multi-view Clustering under the Nested Matrix-Tensor Model
We study the estimation of a planted signal hidden in a recently introduced nested matrix-tensor model, which is an extension of the classical spiked rank-one tensor model, motivated by multi-view clustering. Prior work has theoretically examined the performance of a tensor-based approach, which relies on finding a best rank-one approximation, a problem known to be computationally hard. A tractable alternative approach consists in computing instead the best rank-one (matrix) approximation of an unfolding of the observed tensor data, but its performance was hitherto unknown. We quantify here the performance gap between these two approaches, in particular by deriving the precise algorithmic threshold of the unfolding approach and demonstrating that it exhibits a BBP-type transition behavior. This work is therefore in line with recent contributions which deepen our understanding of why tensor-based methods surpass matrix-based methods in handling structured tensor data.
****
Sheng Xu,Guiliang Liu
Uncertainty-aware Constraint Inference in Inverse Constrained Reinforcement Learning
Aiming for safe control, Inverse Constrained Reinforcement Learning (ICRL) considers inferring the constraints respected by expert agents from their demonstrations and learning imitation policies that adhere to these constraints. While previous ICRL works often neglected underlying uncertainties during training, we contend that modeling these uncertainties is crucial for facilitating robust constraint inference. This insight leads to the development of an Uncertainty-aware Inverse Constrained Reinforcement Learning (UAICRL) algorithm. Specifically, 1) aleatoric uncertainty arises from the inherent stochasticity of environment dynamics, leading to constraint-violating behaviors in imitation policies. To address this, UAICRL constructs risk-sensitive constraints by incorporating distributional Bellman updates into the cumulative costs model. 2) Epistemic uncertainty, resulting from the model's limited knowledge of Out-of-Distribution (OoD) samples, affects the accuracy of step-wise cost predictions. To tackle this issue, UAICRL develops an information-theoretic quantification of the epistemic uncertainty and mitigates its impact through flow-based generative data augmentation. Empirical results demonstrate that UAICRL consistently outperforms other baselines in continuous and discrete environments with stochastic dynamics. The code is available at https://github.com/Jasonxu1225/UAICRL.
****
Lirui Wang,Kaiqing Zhang,Allan Zhou,Max Simchowitz,Russ Tedrake
Robot Fleet Learning via Policy Merging
Fleets of robots ingest massive amounts of heterogeneous streaming data silos generated by interacting with their environments, far more than what can be stored or transmitted with ease. At the same time, teams of robots should co-acquire diverse skills through their heterogeneous experiences in varied settings. How can we enable such fleet-level learning without having to transmit or centralize fleet-scale data? In this paper, we investigate policy merging (PoMe) from such distributed heterogeneous datasets as a potential solution. To efficiently merge policies in the fleet setting, we propose FLEET-MERGE, an instantiation of distributed learning that accounts for the permutation invariance that arises when parameterizing the control policies with recurrent neural networks. We show that F

LEET-MERGE consolidates the behavior of policies trained on 50 tasks in the Meta-World environment, with good performance on nearly all training tasks at test time. Moreover, we introduce a novel robotic tool-use benchmark, FLEET-TOOLS, for fleet policy learning in compositional and contact-rich robot manipulation tasks, to validate the efficacy of FLEET-MERGE on the benchmark.
****
zaishuo xia,Han Yang,Binghui Wang,Jinyuan Jia
GraphGuard: Provably Robust Graph Classification against Adversarial Attacks
Graph classification, which aims to predict a label for a graph, has many real-world applications such as malware detection, fraud detection, and healthcare. However, many studies show an attacker could carefully perturb the structure and/or node features in a graph such that a graph classifier misclassifies the perturbed graph. Such vulnerability impedes the deployment of graph classification in security/safety-critical applications. Existing empirical defenses lack formal robustness guarantees and could be broken by adaptive or unknown attacks. Existing provable defenses have the following limitations: 1) they achieve sub-optimal robustness guarantees for graph structure perturbation, 2) they cannot provide robustness guarantees for arbitrarily node feature perturbations, 3) their robustness guarantees are probabilistic, meaning they could be incorrect with a non-zero probability, and 4) they incur large computation costs. We aim to address those limitations in this work. We propose GraphGuard, a certified defense against both graph structure and node feature perturbations for graph classification. Our GraphGuard provably predicts the same label for a graph when the number of perturbed edges and the number of nodes with perturbed features are bounded. Our results on 8 benchmark datasets show GraphGuard outperforms three state-of-the-art methods.
****
Yecheng Jason Ma,William Liang,Guanzhi Wang,De-An Huang,Osbert Bastani,Dinesh Jayaraman,Yuke Zhu,Linxi Fan,Anima Anandkumar
Eureka: Human-Level Reward Design via Coding Large Language Models
Large Language Models (LLMs) have excelled as high-level semantic planners for sequential decision-making tasks. However, harnessing them to learn complex low-level manipulation tasks, such as dexterous pen spinning, remains an open problem. We bridge this fundamental gap and present Eureka, a human-level reward design algorithm powered by LLMs. Eureka exploits the remarkable zero-shot generation, code-writing, and in-context improvement capabilities of state-of-the-art LLMs, such as GPT-4, to perform evolutionary optimization over reward code. The resulting rewards can then be used to acquire complex skills via reinforcement learning. Without any task-specific prompting or pre-defined reward templates, Eureka generates reward functions that outperform expert human-engineered rewards. In a diverse suite of 29 open-source RL environments that include 10 distinct robot morphologies, Eureka outperforms human experts on 83% of the tasks, leading to an average normalized improvement of 52%. The generality of Eureka also enables a new gradient-free in-context learning approach to reinforcement learning from human feedback (RLHF), readily incorporating human inputs to improve the quality and the safety of the generated rewards without model updating. Finally, using Eureka rewards in a curriculum learning setting, we demonstrate for the first time, a simulated Shadow Hand capable of performing pen spinning tricks, adeptly manipulating a pen in circles at rapid speed.
****
Samar Khanna,Patrick Liu,Linqi Zhou,Chenlin Meng,Robin Rombach,Marshall Burke,David B. Lobell,Stefano Ermon
DiffusionSat: A Generative Foundation Model for Satellite Imagery
Diffusion models have achieved state-of-the-art results on many modalities including images, speech, and video. However, existing models are not tailored to support remote sensing data, which is widely used in important applications including environmental monitoring and crop-yield prediction. Satellite images are significantly different from natural images -- they can be multi-spectral, irregularly sampled across time -- and existing diffusion models trained on images from the Web do not support them. Furthermore, remote sensing data is inherently spati

o-temporal, requiring conditional generation tasks not supported by traditional methods based on captions or images. In this paper, we present DiffusionSat, to date the largest generative foundation model trained on a collection of publicly available large, high-resolution remote sensing datasets .
As text-based captions are sparsely available for satellite images, we incorporate the associated metadata such as geolocation as conditioning information.
Our method produces realistic samples and can be used to solve multiple generative tasks including temporal generation, multi-spectral superrresolution and in-painting. Our method outperforms previous state-of-the-art methods for satellite image generation and is the first large-scale _generative_ foundation model for satellite imagery.
The project website can be found here: https://samar-khanna.github.io/DiffusionSat/
****
Ruipeng Zhang,Ziqing Fan,Jiangchao Yao,Ya Zhang,Yanfeng Wang
Domain-Inspired Sharpness-Aware Minimization Under Domain Shifts
This paper presents a Domain-Inspired Sharpness-Aware Minimization (DISAM) algorithm for optimization under domain shifts. It is motivated by the inconsistent convergence degree of SAM across different domains, which induces optimization bias towards certain domains and thus impairs the overall convergence. To address this issue, we consider the domain-level convergence consistency in the sharpness estimation to prevent the overwhelming (deficient) perturbations for less (well) optimized domains. Specifically, DISAM introduces the constraint of minimizing variance in the domain loss, which allows the elastic gradient calibration in perturbation generation: when one domain is optimized above the averaging level w.r.t. loss, the gradient perturbation towards that domain will be weakened automatically, and vice versa. Under this mechanism, we theoretically show that DISAM can achieve faster overall convergence and improved generalization in principle when inconsistent convergence emerges. Extensive experiments on various domain generalization benchmarks show the superiority of DISAM over a range of state-of-the-art methods. Furthermore, we show the superior efficiency of DISAM in parameter-efficient fine-tuning combined with the pretraining models. The source code is released at https://github.com/MediaBrain-SJTU/DISAM.
****
Rui Qiao,Bryan Kian Hsiang Low
Understanding Domain Generalization: A Noise Robustness Perspective
Despite the rapid development of machine learning algorithms for domain generalization (DG), there is no clear empirical evidence that the existing DG algorithms outperform the classic empirical risk minimization (ERM) across standard benchmarks. To better understand this phenomenon, we investigate whether there are benefits of DG algorithms over ERM through the lens of label noise.
Specifically, our finite-sample analysis reveals that label noise exacerbates the effect of spurious correlations for ERM, undermining generalization.
Conversely, we illustrate that DG algorithms exhibit implicit label-noise robustness during finite-sample training even when spurious correlation is present.
Such desirable property helps mitigate spurious correlations and improve generalization in synthetic experiments.
However, additional comprehensive experiments on real-world benchmark datasets indicate that label-noise robustness does not necessarily translate to better performance compared to ERM.
We conjecture that the failure mode of ERM arising from spurious correlations may be less pronounced in practice. Our code is available at https://github.com/qiaoruiyt/NoiseRobustDG
****
Xinyue Xu,Yi Qin,Lu Mi,Hao Wang,Xiaomeng Li
Energy-Based Concept Bottleneck Models: Unifying Prediction, Concept Intervention, and Probabilistic Interpretations
Existing methods, such as concept bottleneck models (CBMs), have been successful in providing concept-based interpretations for black-box deep learning models.
They typically work by predicting concepts given the input and then predicting t

he final class label given the predicted concepts. However, (1) they often fail to capture the high-order, nonlinear interaction between concepts, e.g., correcting a predicted concept (e.g., “yellow breast”) does not help correct highly correlated concepts (e.g., “yellow belly”), leading to suboptimal final accuracy; (2) they cannot naturally quantify the complex conditional dependencies between different concepts and class labels (e.g., for an image with the class label “Kentucky Warbler” and a concept “black bill”, what is the probability that the model correctly predicts another concept “black crown”), therefore failing to provide deeper insight into how a black-box model works. In response to these limitations, we propose Energy-based Concept Bottleneck Models (ECBMs). Our ECBMs use a set of neural networks to define the joint energy of candidate (input, concept, class) tuples. With such a unified interface, prediction, concept correction, and conditional dependency quantification are then represented as conditional probabilities, which are generated by composing different energy functions. Our ECBMs address both limitations of existing CBMs, providing higher accuracy and richer concept interpretations. Empirical results show that our approach outperforms the state-of-the-art on real-world datasets.
****
Krishna Acharya,Eshwar Ram Arunachaleswaran,Sampath Kannan,Aaron Roth,Juba Ziani
Oracle Efficient Algorithms for Groupwise Regret
We study the problem of online prediction, in which at each time step $t \in \{1,2, \cdots T\}$, an individual x_t arrives, whose label we must predict. Each individual is associated with various groups, defined based on their features such as age, sex, race etc., which may intersect. Our goal is to make predictions that have regret guarantees not just overall but also simultaneously on each sub-sequence comprised of the members of any single group. Previous work such as [Blum & Lykouris][1] and [Lee et al][2] provide attractive regret guarantees for these problems; however, these are computationally intractable on large model classes (e.g., the set of all linear models, as used in linear regression). We show that a simple modification of the sleeping experts technique of [Blum & Lykouris][1] yields an efficient *reduction* to the well-understood problem of obtaining diminishing external regret *absent group considerations*.
Our approach gives similar regret guarantees compared to [Blum & Lykouris][1]; however, we run in time linear in the number of groups, and are oracle-efficient in the hypothesis class. This in particular implies that our algorithm is efficient whenever the number of groups is polynomially bounded and the external-regret problem can be solved efficiently, an improvement on [Blum & Lykouris][1]'s stronger condition that the model class must be small. Our approach can handle online linear regression and online combinatorial optimization problems like online shortest paths. Beyond providing theoretical regret bounds, we evaluate this algorithm with an extensive set of experiments on synthetic data and on two real data sets --- Medical costs and the Adult income dataset, both instantiated with intersecting groups defined in terms of race, sex, and other demographic characteristics.
We find that uniformly across groups, our algorithm gives substantial error improvements compared to running a standard online linear regression algorithm with no groupwise regret guarantees.
****
Ziyang Xiao,Dongxiang Zhang,Yangjun Wu,Lilin Xu,Yuan Jessica Wang,Xiongwei Han,Xiaojin Fu,Tao Zhong,Jia Zeng,Mingli Song,Gang Chen
Chain-of-Experts: When LLMs Meet Complex Operations Research Problems
Large language models (LLMs) have emerged as powerful techniques for various NLP tasks, such as mathematical reasoning and plan generation. In this paper, we study automatic modeling and programming for complex operation research (OR) problems, so as to alleviate the heavy dependence on domain experts and benefit a spectrum of industry sectors. We present the first LLM-based solution, namely Chain-of-Experts (CoE), a novel multi-agent cooperative framework to enhance reasoning capabilities. Specifically, each agent is assigned a specific role and endowed with domain knowledge related to OR. We also introduce a conductor to orchestrate these agents via forward thought construction and backward reflection mechani

sm. Furthermore, we release a benchmark dataset (ComplexOR) of complex OR problems to facilitate OR research and community development. Experimental results show that CoE significantly outperforms the state-of-the-art LLM-based approaches both on LPWP and ComplexOR.
****
Johnathan Wenjia Xie,Yoonho Lee,Annie S Chen,Chelsea Finn
Self-Guided Masked Autoencoders for Domain-Agnostic Self-Supervised Learning
Self-supervised learning excels in learning representations from large amounts of unlabeled data, demonstrating success across multiple data modalities. Yet, extending self-supervised learning to new modalities is non-trivial because the specifics of existing methods are tailored to each domain, such as domain-specific augmentations which reflect the invariances in the target task. While masked modeling is promising as a domain-agnostic framework for self-supervised learning because it does not rely on input augmentations, its mask sampling procedure remains domain-specific. We present Self-guided Masked Autoencoders (SMA), a fully domain-agnostic masked modeling method. SMA trains an attention based model using a masked modeling objective, by learning masks to sample without any domain-specific assumptions. We evaluate SMA on three self-supervised learning benchmarks in protein biology, chemical property prediction, and particle physics. We find SMA is capable of learning representations without domain-specific knowledge and achieves state-of-the-art performance on these three benchmarks.
****
Tian Qiu,Xu Wenxiang,lin chen,Zhou Linyun,Zunlei Feng,Mingli Song
Dynamic Neural Response Tuning
Artificial Neural Networks (ANNs) have gained widespread applications across various areas in recent years. The ANN design was initially inspired by principles of biology. The biological neural network's fundamental response process comprises information transmission and aggregation. The information transmission in biological neurons is often achieved by triggering action potentials that propagate through axons. ANNs utilize activation mechanisms to simulate such biological behavior. However, previous studies have only considered static response conditions, while the biological neuron's response conditions are typically dynamic, depending on multiple factors such as neuronal properties and the real-time environment. Therefore, the dynamic response conditions of biological neurons could help improve the static ones of existing activations in ANNs. Additionally, the biological neuron's aggregated response exhibits high specificity for different categories, allowing the nervous system to differentiate and identify objects. Inspired by these biological patterns, we propose a novel Dynamic Neural Response Tuning (DNRT) mechanism, which aligns the response patterns of ANNs with those of biological neurons. DNRT comprises Response-Adaptive Activation (RAA) and Aggregated Response Regularization (ARR), mimicking the biological neuron's information transmission and aggregation behaviors. RAA dynamically adjusts the response condition based on the characteristics and strength of the input signal. ARR is devised to enhance the network's ability to learn category specificity by imposing constraints on the network's response distribution. Extensive experimental studies indicate that the proposed DNRT is highly interpretable, applicable to various mainstream network architectures, and can achieve remarkable performance compared with existing neural response mechanisms in multiple tasks and domains. Code is available at https://github.com/horrible-dong/DNRT.
****
Germain Kolossov,Andrea Montanari,Pulkit Tandon
Towards a statistical theory of data selection under weak supervision
Given a sample of size N, it is often useful to select a subsample of smaller size $n<N$ to be used for statistical estimation or learning. Such a data selection step is useful to reduce the requirements of data labeling and the computational complexity of learning. We assume to be given N unlabeled samples x_{i}, and to be given access to a `surrogate model' that can predict labels y_i better than random guessing. Our goal is to select a subset of the samples, to be denoted by {x_{i}}$_{i\in G}$, of size $|G|=n<N$. We then acquire labels for this set and we use them to train a model via regularized empirical risk minimi

zation. By using a mixture of numerical experiments on real and synthetic data, and mathematical derivations under low- and high- dimensional asymptotics, we show that: (i) Data selection can be very effective, in particular beating training on the full sample in some cases; (ii) Certain popular choices in data selection methods (e.g. unbiased reweighted subsampling, or influence function-based subsampling) can be substantially suboptimal.
****
Chuan Wen,Dinesh Jayaraman,Yang Gao
Can Transformers Capture Spatial Relations between Objects?
Spatial relationships between objects represent key scene information for humans to understand and interact with the world. To study the capability of current computer vision systems to recognize physically grounded spatial relations, we start by proposing precise relation definitions that permit consistently annotating a benchmark dataset. Despite the apparent simplicity of this task relative to others in the recognition literature, we observe that existing approaches perform poorly on this benchmark. We propose new approaches exploiting the long-range attention capabilities of transformers for this task, and evaluating key design principles. We identify a simple ``RelatiViT'' architecture and demonstrate that it outperforms all current approaches. To our knowledge, this is the first method to convincingly outperform naive baselines on spatial relation prediction in in-the-wild settings. The code and datasets are available in \url{https://sites.google.com/view/spatial-relation}.
****
Margalit Glasgow
SGD Finds then Tunes Features in Two-Layer Neural Networks with near-Optimal Sample Complexity: A Case Study in the XOR problem
In this work, we consider the optimization process of minibatch stochastic gradient descent (SGD) on a 2-layer neural network with data separated by a quadratic ground truth function. We prove that with data drawn from the Boolean hypercube labeled by the quadratic ``XOR'' function $y = -x_ix_j$, it is possible to train to a population error $o(1)$
with $\Theta(d\text{polylog}(d))$ samples. Our result considers simultaneously training both layers of the two-layer-neural network with ReLU activations via standard minibatch SGD on the logistic loss. To our knowledge, this work is the first to give a sample complexity of
for efficiently learning the XOR function on isotropic data on a standard neural network with standard training. Our main technique is showing that the network evolves in two phases: a \em signal-finding \em phase where the network is small and many of the neurons evolve independently to find features, and a \em signal-heavy \em phase, where SGD maintains and balances the features. We leverage the simultaneous training of the layers to show that it is sufficient for only a small fraction of the neurons to learn features, since those neurons will be amplified by the simultaneous growth of their second layer weights.
****
Patricia Pauli,Aaron J Havens,Alexandre Araujo,Siddharth Garg,Farshad Khorrami,Frank Allgöwer,Bin Hu
Novel Quadratic Constraints for Extending LipSDP beyond Slope-Restricted Activations
Recently, semidefinite programming (SDP) techniques have shown great promise in providing accurate Lipschitz bounds for neural networks. Specifically, the LipSDP approach (Fazlyab et al., 2019) has received much attention and provides the least conservative Lipschitz upper bounds that can be computed with polynomial time guarantees. However, one main restriction of LipSDP is that its formulation requires the activation functions to be slope-restricted on $[0,1]$, preventing its further use for more general activation functions such as GroupSort, MaxMin, and Householder. One can rewrite MaxMin activations for example as residual ReLU networks. However, a direct application of LipSDP to the resultant residual ReLU networks is conservative and even fails in recovering the well-known fact that the MaxMin activation is 1-Lipschitz. Our paper bridges this gap and extends LipSDP beyond slope-restricted activation functions. To this end, we provide nove

l quadratic constraints for GroupSort, MaxMin, and Householder activations via leveraging their underlying properties such as sum preservation. Our proposed analysis is general and provides a unified approach for estimating ℓ_2 and ℓ_∞ Lipschitz bounds for a rich class of neural network architectures, including non-residual and residual neural networks and implicit models, with GroupSort, MaxMin, and HouseHolder activations. Finally, we illustrate the utility of our approach with a variety of experiments and show that our proposed SDPs generate less conservative Lipschitz bounds in comparison to existing approaches.
****
Fred Zhang,Neel Nanda
Towards Best Practices of Activation Patching in Language Models: Metrics and Methods
Mechanistic interpretability seeks to understand the internal mechanisms of machine learning models, where localization—identifying the important model components—is a key step. Activation patching, also known as causal tracing or interchange intervention, is a standard technique for this task (Vig et al., 2020), but
the literature contains many variants with little consensus on the choice of hyperparameters or methodology. In this work, we systematically examine the impact of methodological details in activation patching, including evaluation metrics and
corruption methods. In several settings of localization and circuit discovery in language models, we find that varying these hyperparameters could lead to disparate
interpretability results. Backed by empirical observations, we give conceptual arguments for why certain metrics or methods may be preferred. Finally, we provide
recommendations for the best practices of activation patching going forwards.
****
Thong Thanh Nguyen,Xiaobao Wu,Xinshuai Dong,Cong-Duy T Nguyen,See-Kiong Ng,Anh Tuan Luu
Topic Modeling as Multi-Objective Contrastive Optimization
Recent representation learning approaches enhance neural topic models by optimizing the weighted linear combination of the evidence lower bound (ELBO) of the log-likelihood and the contrastive learning objective that contrasts pairs of input documents. However, document-level contrastive learning might capture low-level mutual information, such as word ratio, which disturbs topic modeling. Moreover, there is a potential conflict between the ELBO loss that memorizes input details for better reconstruction quality, and the contrastive loss which attempts to learn topic representations that generalize among input documents. To address these issues, we first introduce a novel contrastive learning method oriented towards sets of topic vectors to capture useful semantics that are shared among a set of input documents. Secondly, we explicitly cast contrastive topic modeling as a gradient-based multi-objective optimization problem, with the goal of achieving a Pareto stationary solution that balances the trade-off between the ELBO and the contrastive objective. Extensive experiments demonstrate that our framework consistently produces higher-performing neural topic models in terms of topic coherence, topic diversity, and downstream performance.
****
Wuyang Chen,Junru Wu,Zhangyang Wang,Boris Hanin
Principled Architecture-aware Scaling of Hyperparameters
Training a high-quality deep neural network requires choosing suitable hyperparameters, which is a non-trivial and expensive process. Current works try to automatically optimize or design principles of hyperparameters, such that they can generalize to diverse unseen scenarios. However, most designs of principles or optimization methods are agnostic to the choice of network structures, and thus largely ignore the impact of neural architectures on hyperparameters. In this work, we precisely characterize the dependence of initializations and maximal learning rates on the network architecture, which includes the network depth, width, convolutional kernel size, and connectivity patterns. By pursuing every parameter

to be maximally updated with the same mean squared change in pre-activations, we can generalize our initialization and learning rates across MLPs (multi-layer perception) and CNNs (convolutional neural network) with sophisticated graph topologies. We verify our principles with comprehensive experiments. More importantly, our strategy further sheds light on advancing current benchmarks for architecture design. A fair comparison of AutoML algorithms requires accurate network rankings. However, we demonstrate that network rankings can be easily changed by better training networks in benchmarks with our architecture-aware learning rates and initialization.
****
Lukas Muttenthaler,Robert A. Vandermeulen,Qiuyi Zhang,Thomas Unterthiner,Klaus Robert Muller
Set Learning for Accurate and Calibrated Models
Model overconfidence and poor calibration are common in machine learning and difficult to account for when applying standard empirical risk minimization. In this work, we propose a novel method to alleviate these problems that we call odd-k-out learning (OKO), which minimizes the cross-entropy error for sets rather than for single examples. This naturally allows the model to capture correlations across data examples and achieves both better accuracy and calibration, especially in limited training data and class-imbalanced regimes. Perhaps surprisingly, OKO often yields better calibration even when training with hard labels and dropping any additional calibration parameter tuning, such as temperature scaling. We demonstrate this in extensive experimental analyses and provide a mathematical theory to interpret our findings. We emphasize that OKO is a general framework that can be easily adapted to many settings and a trained model can be applied to single examples at inference time, without significant run-time overhead or architecture changes.
****
Yi-Fu Wu,Minseung Lee,Sungjin Ahn
Structured World Modeling via Semantic Vector Quantization
Neural discrete representations are crucial components of modern neural networks. However, their main limitation is that the primary strategies such as VQ-VAE can only provide representations at the patch level. Therefore, one of the main goals of representation learning, acquiring structured, semantic, and compositional abstractions such as the color and shape of an object, remains elusive. In this paper, we present the first approach to semantic neural discrete representation learning. The proposed model, called Semantic Vector-Quantized Variational Autoencoder (SVQ), leverages recent advances in unsupervised object-centric learning to address this limitation. Specifically, we observe that a simple approach quantizing at the object level poses a significant challenge and propose constructing scene representations hierarchically, from low-level discrete concept schemas to object representations. Additionally, we suggest a novel method for structured semantic world modeling by training a prior over these representations, enabling the ability to generate images by sampling the semantic properties of the objects in the scene. In experiments on various 2D and 3D object-centric datasets, we find that our model achieves superior generation performance compared to nonsemantic vector quantization methods such as VQ-VAE and previous object-centric generative models. Furthermore, we find that the semantic discrete representations can solve downstream scene understanding tasks that require reasoning about the properties of different objects in the scene.
****
Sepehr Dehdashtian,Lan Wang,Vishnu Boddeti
FairerCLIP: Debiasing CLIP's Zero-Shot Predictions using Functions in RKHSs
Large pre-trained vision-language models such as CLIP provide compact and general-purpose representations of text and images that are demonstrably effective across multiple downstream zero-shot prediction tasks. However, owing to the nature of their training process, these models have the potential to 1) propagate or amplify societal biases in the training data and 2) learn to rely on spurious features. This paper proposes FairerCLIP, a general approach for making zero-shot predictions of CLIP more fair and robust to spurious correlations. We formulate t

he problem of jointly debiasing CLIP's image and text representations in reproducing kernel Hilbert spaces (RKHSs), which affords multiple benefits: 1) Flexibility: Unlike existing approaches, which are specialized to either learn with or without ground-truth labels, FairerCLIP is adaptable to learning in both scenarios. 2) Ease of Optimization: FairerCLIP lends itself to an iterative optimization involving closed-form solvers, which leads to 4×-10× faster training than the existing methods. 3) Sample Efficiency: Under sample-limited conditions, FairerCLIP significantly outperforms baselines when they fail entirely. And, 4) Performance: Empirically, FairerCLIP achieves appreciable accuracy gains on benchmark fairness and spurious correlation datasets over their respective baselines.
****
Suttisak Wizadwongsa,Worameth Chinchuthakun,Pramook Khungurn,Amit Raj,Supasorn Suwajanakorn
Diffusion Sampling with Momentum for Mitigating Divergence Artifacts
Despite the remarkable success of diffusion models in image generation, slow sampling remains a persistent issue. To accelerate the sampling process, prior studies have reformulated diffusion sampling as an ODE/SDE and introduced higher-order numerical methods. However, these methods often produce divergence artifacts, especially with a low number of sampling steps, which limits the achievable acceleration. In this paper, we investigate the potential causes of these artifacts and suggest that the small stability regions of these methods could be the principal cause. To address this issue, we propose two novel techniques. The first technique involves the incorporation of Heavy Ball (HB) momentum, a well-known technique for improving optimization, into existing diffusion numerical methods to expand their stability regions. We also prove that the resulting methods have first-order convergence. The second technique, called Generalized Heavy Ball (GHVB), constructs a new high-order method that offers a variable trade-off between accuracy and artifact suppression. Experimental results show that our techniques are highly effective in reducing artifacts and improving image quality, surpassing state-of-the-art diffusion solvers on both pixel-based and latent-based diffusion models for low-step sampling. Our research provides novel insights into the design of numerical methods for future diffusion work.
****
Byeonghu Na,Yeongmin Kim,HeeSun Bae,Jung Hyun Lee,Se Jung Kwon,Wanmo Kang,Il-chul Moon
Label-Noise Robust Diffusion Models
Conditional diffusion models have shown remarkable performance in various generative tasks, but training them requires large-scale datasets that often contain noise in conditional inputs, a.k.a. noisy labels. This noise leads to condition mismatch and quality degradation of generated data. This paper proposes Transition-aware weighted Denoising Score Matching (TDSM) for training conditional diffusion models with noisy labels, which is the first study in the line of diffusion models. The TDSM objective contains a weighted sum of score networks, incorporating instance-wise and time-dependent label transition probabilities. We introduce a transition-aware weight estimator, which leverages a time-dependent noisy-label classifier distinctively customized to the diffusion process. Through experiments across various datasets and noisy label settings, TDSM improves the quality of generated samples aligned with given conditions. Furthermore, our method improves generation performance even on prevalent benchmark datasets, which implies the potential noisy labels and their risk of generative model learning. Finally, we show the improved performance of TDSM on top of conventional noisy label corrections, which empirically proving its contribution as a part of label-noise robust generative models. Our code is available at: https://github.com/byeonghu-na/tdsm.
****
Madhur Panwar,Kabir Ahuja,Navin Goyal
In-Context Learning through the Bayesian Prism
In-context learning (ICL) is one of the surprising and useful features of large language models and subject of intense research. Recently, stylized meta-learning-like ICL setups have been devised that train transformers on sequences of inpu

t-output pairs $(x, f(x))$. The function f comes from a function class and gen
eralization is checked by evaluation on sequences for unseen functions from the
same class. One of the main discoveries in this line of research has been that f
or several function classes, such as linear regression, transformers successfull
y generalize to new functions in the class. However, the inductive biases of the
se models resulting in this behavior are not clearly understood. A model with un
limited training data and compute is a Bayesian predictor: it learns the pretrai
ning distribution.
In this paper we empirically examine how far this Bayesian perspective can help
us understand ICL. To this end, we generalize the previous meta-ICL setup to hie
rarchical meta-ICL setup which involve unions of multiple task families. We inst
antiate this setup on a diverse range of linear and nonlinear function families
and find that transformers can do ICL in this setting as well. Where Bayesian in
ference is tractable, we find evidence that high-capacity transformers mimic the
 Bayesian predictor. The Bayesian perspective provides insights into the inducti
ve bias of ICL and how transformers perform a particular task when they are trai
ned on multiple tasks. We also find that transformers can learn to generalize to
 new function classes that were not seen during pretraining. This involves devia
tion from the Bayesian predictor. We examine these deviations in more depth offe
ring new insights and hypotheses.
****
Whie Jung,Jaehoon Yoo,Sungjin Ahn,Seunghoon Hong
Learning to Compose: Improving Object Centric Learning by Injecting Compositiona
lity
Learning compositional representation is a key aspect of object-centric learning
 as it enables flexible systematic generalization and supports complex visual re
asoning. However, most of the existing approaches rely on auto-encoding objectiv
e, while the compositionality is implicitly imposed by the architectural or algo
rithmic bias in the encoder. This misalignment between auto-encoding objective a
nd learning compositionality often results in failure of capturing meaningful ob
ject representations. In this study, we propose a novel objective that explicitl
y encourages compositionality of the representations. Built upon the existing ob
ject-centric learning framework (e.g., slot attention), our method incorporates
additional constraints that an arbitrary mixture of object representations from
two images should be valid by maximizing the likelihood of the composite data. W
e demonstrate that incorporating our objective to the existing framework consist
ently improves the objective-centric learning and enhances the robustness to the
 architectural choices.
****
Bohang Zhang,Jingchu Gai,Yiheng Du,Qiwei Ye,Di He,Liwei Wang
Beyond Weisfeiler-Lehman: A Quantitative Framework for GNN Expressiveness
Designing expressive Graph Neural Networks (GNNs) is a fundamental topic in the
graph learning community. So far, GNN expressiveness has been primarily assessed
 via the Weisfeiler-Lehman (WL) hierarchy. However, such an expressivity measure
 has notable limitations: it is inherently coarse, qualitative, and may not well
 reflect practical requirements (e.g., the ability to encode substructures). In
this paper, we introduce a novel framework for quantitatively studying the expre
ssiveness of GNN architectures, addressing all the above limitations. Specifical
ly, we identify a fundamental expressivity measure termed homomorphism expressiv
ity, which quantifies the ability of GNN models to count graphs under homomorphi
sm. Homomorphism expressivity offers a complete and practical assessment tool: t
he completeness enables direct expressivity comparisons between GNN models, whil
e the practicality allows for understanding concrete GNN abilities such as subgr
aph counting. By examining four classes of prominent GNNs as case studies, we de
rive simple, unified, and elegant descriptions of their homomorphism expressivit
y for both invariant and equivariant settings. Our results provide novel insight
s into a series of previous work, unify the landscape of different subareas in t
he community, and settle several open questions. Empirically, extensive experime
nts on both synthetic and real-world tasks verify our theory, showing that the p
ractical performance of GNN models aligns well with the proposed metric.

****
Yeda Song,Dongwook Lee,Gunhee Kim
Compositional Conservatism: A Transductive Approach in Offline Reinforcement Learning
Offline reinforcement learning (RL) is a compelling framework for learning optimal policies from past experiences without additional interaction with the environment. Nevertheless, offline RL inevitably faces the problem of distributional shifts, where the states and actions encountered during policy execution may not be in the training dataset distribution. A common solution involves incorporating conservatism into the policy or the value function to safeguard against uncertainties and unknowns. In this work, we focus on achieving the same objectives of conservatism but from a different perspective. We propose COmpositional COnservatism with Anchor-seeking (COCOA) for offline RL, an approach that pursues conservatism in a _compositional_ manner on top of the transductive reparameterization (Netanyahu et al., 2023), which decomposes the input variable (the state in our case) into an anchor and its difference from the original input. Our COCOA seeks both in-distribution anchors and differences by utilizing the learned reverse dynamics model, encouraging conservatism in the compositional input space for the policy or value function. Such compositional conservatism is independent of and agnostic to the prevalent _behavioral_ conservatism in offline RL. We apply COCOA to four state-of-the-art offline RL algorithms and evaluate them on the D4RL benchmark, where COCOA generally improves the performance of each algorithm. The code is available at https://github.com/runamu/compositional-conservatism.
****
Arturs Backurs,Zinan Lin,Sepideh Mahabadi,Sandeep Silwal,Jakub Tarnawski
Efficiently Computing Similarities to Private Datasets
Many methods in differentially private model training rely on computing the similarity between a query point (such as public or synthetic data) and private data. We abstract out this common subroutine and study the following fundamental algorithmic problem: Given a similarity function f and a large high-dimensional private dataset $X \subset \mathbb{R}^d$, output a differentially private (DP) data-structure which approximates $\sum_{x \in X} f(x,y)$ for any query y. We consider the cases where f is a kernel function, such as $f(x,y) = e^{-\|x-y\|_2^2/\sigma^2}$ (also known as DP kernel density estimation), or a distance function such as $f(x,y) = \|x-y\|_2$, among others.

Our theoretical results improve upon prior work and give better privacy-utility trade-offs as well as faster query times for a wide range of kernels and distance functions. The unifying approach behind our results is leveraging `low-dimensional structures' present in the specific functions f that we study, using tools such as provable dimensionality reduction, approximation theory, and one-dimensional decomposition of the functions. Our algorithms empirically exhibit improved query times and accuracy over prior state of the art. We also present an application to DP classification. Our experiments demonstrate that the simple methodology of classifying based on average similarity is orders of magnitude faster than prior DP-SGD based approaches for comparable accuracy.
****
Rene Winchenbach,Nils Thuerey
Symmetric Basis Convolutions for Learning Lagrangian Fluid Mechanics
Learning physical simulations has been an essential and central aspect of many recent research efforts in machine learning, particularly for Navier-Stokes-based fluid mechanics. Classic numerical solvers have traditionally been computationally expensive and challenging to use in inverse problems, whereas Neural solvers aim to address both concerns through machine learning. We propose a general formulation for continuous convolutions using separable basis functions as a superset of existing methods and evaluate a large set of basis functions in the context of (a) a compressible 1D SPH simulation, (b) a weakly compressible 2D SPH simulation, and (c) an incompressible 2D SPH Simulation. We demonstrate that even and odd symmetries included in the basis functions are key aspects of stability and accuracy.

Our broad evaluation shows that Fourier-based continuous convolutions outperform all other architectures regarding accuracy and generalization. Finally, using these Fourier-based networks, we show that prior inductive biases, such as window functions, are no longer necessary. An implementation of our approach, as well as complete datasets and solver implementations, is available at https://github.com/orgs/tum-pbs/SFBC.
****
Han Li,Shaohui Li,Wenrui Dai,Chenglin Li,Junni Zou,Hongkai Xiong
Frequency-Aware Transformer for Learned Image Compression
Learned image compression (LIC) has gained traction as an effective solution for image storage and transmission in recent years. However, existing LIC methods are redundant in latent representation due to limitations in capturing anisotropic frequency components and preserving directional details. To overcome these challenges, we propose a novel frequency-aware transformer (FAT) block that for the first time achieves multiscale directional ananlysis for LIC. The FAT block comprises frequency-decomposition window attention (FDWA) modules to capture multiscale and directional frequency components of natural images. Additionally, we introduce frequency-modulation feed-forward network (FMFFN) to adaptively modulate different frequency components, improving rate-distortion performance. Furthermore, we present a transformer-based channel-wise autoregressive (T-CA) model that effectively exploits channel dependencies. Experiments show that our method achieves state-of-the-art rate-distortion performance compared to existing LIC methods, and evidently outperforms latest standardized codec VTM-12.1 by 14.5\%, 15.1\%, 13.0\% in BD-rate on the Kodak, Tecnick, and CLIC datasets.
****
Liu Yang,Kangwook Lee,Robert D Nowak,Dimitris Papailiopoulos
Looped Transformers are Better at Learning Learning Algorithms
Transformers have demonstrated effectiveness in in-context solving data-fitting problems from various (latent) models, as reported by Garg et al. (2022). However, the absence of an inherent iterative structure in the transformer architecture presents a challenge in emulating the iterative algorithms, which are commonly employed in traditional machine learning methods. To address this, we propose the utilization of looped transformer architecture and its associated training methodology, with the aim of incorporating iterative characteristics into the transformer architectures. Experimental results suggest that the looped transformer achieves performance comparable to the standard transformer in solving various data-fitting problems, while utilizing less than 10% of the parameter count.
****
Qianxu Wang,Haotong Zhang,Congyue Deng,Yang You,Hao Dong,Yixin Zhu,Leonidas Guibas
SparseDFF: Sparse-View Feature Distillation for One-Shot Dexterous Manipulation
Humans demonstrate remarkable skill in transferring manipulation abilities across objects of varying shapes, poses, and appearances, a capability rooted in their understanding of semantic correspondences between different instances. To equip robots with a similar high-level comprehension, we present SparseDFF, a novel DFF for 3D scenes utilizing large 2D vision models to extract semantic features from sparse RGBD images, a domain where research is limited despite its relevance to many tasks with fixed-camera setups. SparseDFF generates view-consistent 3D DFFs, enabling efficient one-shot learning of dexterous manipulations by mapping image features to a 3D point cloud. Central to SparseDFF is a feature refinement network, optimized with a contrastive loss between views and a point-pruning mechanism for feature continuity. This facilitates the minimization of feature discrepancies w.r.t. end-effector parameters, bridging demonstrations and target manipulations. Validated in real-world scenarios with a dexterous hand, SparseDFF proves effective in manipulating both rigid and deformable objects, demonstrating significant generalization capabilities across object and scene variations.
****
Bin Lu,Tingyan Ma,Xiaoying Gan,Xinbing Wang,Yunqiang Zhu,Chenghu Zhou,Shiyu Liang
Temporal Generalization Estimation in Evolving Graphs

Graph Neural Networks (GNNs) are widely deployed in vast fields, but they often struggle to maintain accurate representations as graphs evolve. We theoretically establish a lower bound, proving that under mild conditions, representation distortion inevitably occurs over time. To estimate the temporal distortion without human annotation after deployment, one naive approach is to pre-train a recurrent model (e.g., RNN) before deployment and use this model afterwards, but the estimation is far from satisfactory. In this paper, we analyze the representation distortion from an information theory perspective, and attribute it primarily to inaccurate feature extraction during evolution. Consequently, we introduce Smart, a straightforward and effective baseline enhanced by an adaptive feature extractor through self-supervised graph reconstruction. In synthetic random graphs, we further refine the former lower bound to show the inevitable distortion over time and empirically observe that Smart achieves good estimation performance. Moreover, we observe that Smart consistently shows outstanding generalization estimation on four real-world evolving graphs. The ablation studies underscore the necessity of graph reconstruction. For example, on OGB-arXiv dataset, the estimation metric MAPE deteriorates from 2.19% to 8.00% without reconstruction.
****
Anshuman Chhabra,Peizhao Li,Prasant Mohapatra,Hongfu Liu
"What Data Benefits My Classifier?" Enhancing Model Performance and Interpretability through Influence-Based Data Selection
Classification models are ubiquitously deployed in society and necessitate high utility, fairness, and robustness performance. Current research efforts mainly focus on improving model architectures and learning algorithms on fixed datasets to achieve this goal. In contrast, in this paper, we address an orthogonal yet crucial problem: given a fixed convex learning model (or a convex surrogate for a non-convex model) and a function of interest, we assess what data benefits the model by interpreting the feature space, and then aim to improve performance as measured by this function. To this end, we propose the use of influence estimation models for interpreting the classifier's performance from the perspective of the data feature space. Additionally, we propose data selection approaches based on influence that enhance model utility, fairness, and robustness. Through extensive experiments on synthetic and real-world datasets, we validate and demonstrate the effectiveness of our approaches not only for conventional classification scenarios, but also under more challenging scenarios such as distribution shifts, fairness poisoning attacks, utility evasion attacks, online learning, and active learning.
****
Chengrui Li,Yule Wang,Weihan Li,Anqi Wu
Forward χ^2 Divergence Based Variational Importance Sampling
Maximizing the marginal log-likelihood is a crucial aspect of learning latent variable models, and variational inference (VI) stands as the commonly adopted method. However, VI can encounter challenges in achieving a high marginal log-likelihood when dealing with complicated posterior distributions. In response to this limitation, we introduce a novel variational importance sampling (VIS) approach that directly estimates and maximizes the marginal log-likelihood. VIS leverages the optimal proposal distribution, achieved by minimizing the forward χ^2 divergence, to enhance marginal log-likelihood estimation. We apply VIS to various popular latent variable models, including mixture models, variational auto-encoders, and partially observable generalized linear models. Results demonstrate that our approach consistently outperforms state-of-the-art baselines, in terms of both log-likelihood and model parameter estimation. Code: \url{https://github.com/JerrySoybean/vis}.
****
Ryo Ueda,Tadahiro Taniguchi
Lewis's Signaling Game as beta-VAE For Natural Word Lengths and Segments
As a sub-discipline of evolutionary and computational linguistics, emergent communication (EC) studies communication protocols, called emergent languages, arising in simulations where agents communicate. A key goal of EC is to give rise to languages that share statistical properties with natural languages. In this pape

r, we reinterpret Lewis's signaling game, a frequently used setting in EC, as beta-VAE and reformulate its objective function as ELBO. Consequently, we clarify the existence of prior distributions of emergent languages and show that the choice of the priors can influence their statistical properties. Specifically, we address the properties of word lengths and segmentation, known as Zipf's law of abbreviation (ZLA) and Harris's articulation scheme (HAS), respectively. It has been reported that the emergent languages do not follow them when using the conventional objective. We experimentally demonstrate that by selecting an appropriate prior distribution, more natural segments emerge, while suggesting that the conventional one prevents the languages from following ZLA and HAS.
****
Simon Schug,Seijin Kobayashi,Yassir Akram,Maciej Wolczyk,Alexandra Maria Proca,Johannes Von Oswald,Razvan Pascanu,Joao Sacramento,Angelika Steger
Discovering modular solutions that generalize compositionally
Many complex tasks can be decomposed into simpler, independent parts. Discovering such underlying compositional structure has the potential to enable compositional generalization. Despite progress, our most powerful systems struggle to compose flexibly. It therefore seems natural to make models more modular to help capture the compositional nature of many tasks. However, it is unclear under which circumstances modular systems can discover hidden compositional structure. To shed light on this question, we study a teacher-student setting with a modular teacher where we have full control over the composition of ground truth modules. This allows us to relate the problem of compositional generalization to that of identification of the underlying modules. In particular we study modularity in hypernetworks representing a general class of multiplicative interactions. We show theoretically that identification up to linear transformation purely from demonstrations is possible without having to learn an exponential number of module combinations. We further demonstrate empirically that under the theoretically identified conditions, meta-learning from finite data can discover modular policies that generalize compositionally in a number of complex environments.
****
Yinghao Xu,Hao Tan,Fujun Luan,Sai Bi,Peng Wang,Jiahao Li,Zifan Shi,Kalyan Sunkavalli,Gordon Wetzstein,Zexiang Xu,Kai Zhang
DMV3D: Denoising Multi-view Diffusion Using 3D Large Reconstruction Model
We propose DMV3D, a novel 3D generation approach that uses a transformer-based 3D large reconstruction model to denoise multi-view diffusion. Our reconstruction model incorporates a triplane NeRF representation and, functioning as a denoiser, can denoise noisy multi-view images via 3D NeRF reconstruction and rendering, achieving single-stage 3D generation in the 2D diffusion denoising process. We train DMV3D on large-scale multi-view image datasets of extremely diverse objects using only image reconstruction losses, without accessing 3D assets. We demonstrate state-of-the-art results for the single-image reconstruction problem where probabilistic modeling of unseen object parts is required for generating diverse reconstructions with sharp textures. We also show high-quality text-to-3D generation results outperforming previous 3D diffusion models. Our project website is at: https://dmv3d.github.io/.
****
Mohammad Pedramfar,Yididiya Y. Nadew,Christopher John Quinn,Vaneet Aggarwal
Unified Projection-Free Algorithms for Adversarial DR-Submodular Optimization
This paper introduces unified projection-free Frank-Wolfe type algorithms for adversarial continuous DR-submodular optimization, spanning scenarios such as full information and (semi-)bandit feedback, monotone and non-monotone functions, different constraints, and types of stochastic queries. For every problem considered in the non-monotone setting, the proposed algorithms are either the first with proven sub-linear α-regret bounds or have better α-regret bounds than the state of the art, where α is a corresponding approximation bound in the offline setting. In the monotone setting, the proposed approach gives state-of-the-art sub-linear α-regret bounds among projection-free algorithms in 7 of the 8 considered cases while matching the result of the remaining case. Additionally, this paper addresses semi-bandit and bandit feedback for advers

arial DR-submodular optimization, advancing the understanding of this optimization area.
****
Jen-tse Huang,Wenxuan Wang,Eric John Li,Man Ho LAM,Shujie Ren,Youliang Yuan,Wenxiang Jiao,Zhaopeng Tu,Michael Lyu
On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs
Large Language Models (LLMs) have recently showcased their remarkable capacities, not only in natural language processing tasks but also across diverse domains such as clinical medicine, legal consultation, and education. LLMs become more than mere applications, evolving into assistants capable of addressing diverse user requests. This narrows the distinction between human beings and artificial intelligence agents, raising intriguing questions regarding the potential manifestation of personalities, temperaments, and emotions within LLMs. In this paper, we propose a framework, PsychoBench, for evaluating diverse psychological aspects of LLMs. Comprising thirteen scales commonly used in clinical psychology, PsychoBench further classifies these scales into four distinct categories: personality traits, interpersonal relationships, motivational tests, and emotional abilities. Our study examines five popular models, namely text-davinci-003, ChatGPT, GPT-4, LLaMA-2-7b, and LLaMA-2-13b. Additionally, we employ a jailbreak approach to bypass the safety alignment protocols and test the intrinsic natures of LLMs. We have made PsychoBench openly accessible via https://github.com/CUHK-ARISE/PsychoBench.
****
Haobo SONG,Hao Zhao,Soumajit Majumder,Tao Lin
Increasing Model Capacity for Free: A Simple Strategy for Parameter Efficient Fine-tuning
Fine-tuning large pre-trained foundation models, such as the 175B GPT-3, has become the prevailing approach for downstream tasks. While parameter-efficient fine-tuning methods have been proposed and proven effective without retraining all model parameters, their performance is limited by the capacity of incremental modules, especially under constrained parameter budgets.
To overcome this challenge, we propose CAPABOOST, a simple yet effective strategy that enhances model capacity by leveraging low-rank updates through parallel weight modules in target layers. By applying static random masks to the shared weight matrix, CAPABOOST constructs a diverse set of weight matrices, effectively increasing the rank of incremental weights without adding parameters. Notably, our approach can be seamlessly integrated into various existing parameter-efficient fine-tuning methods. We extensively validate the efficacy of CAPABOOST through experiments on diverse downstream tasks, including natural language understanding, question answering, and image classification. Our results demonstrate significant improvements over baselines, without incurring additional computation or storage costs. We will make our code and benchmark publicly available.
****
Jinxuan Wang,Shiting Xu,Tong Zhang
A unique M-pattern for micro-expression spotting in long videos
Micro-expression spotting (MES) is challenging since the small magnitude of micro-expression (ME) makes them susceptible to global movements like head rotation. However, the unique movement pattern and inherent characteristics of ME allow them to be distinguished from other movements. Existing MES methods based on fixed reference frame degrade optical flow accuracy and are overly dependent on facial alignment. In this paper, we propose a skip-k-frame block-wise main directional mean optical flow (MDMO) feature for MES based on unfixed reference frame. By employing skip-k-frame strategy, we substantiate the existence of a distinct and exclusive movement pattern in ME, called M-pattern due to its feature curve resembling the letter `M'. Based on M-pattern and characteristics of ME, we then provide a novel spotting rules to precisely locate ME intervals. Block-wise MDMO feature is capable of removing global movements without compromising complete ME movements in the early feature extraction stage. Besides, A novel pixelmatch-based facial alignment algorithm with dynamic update of reference frame is pro

posed to better align facial images and reduce jitter between frames. Experimental results on CAS(ME)^2, SAMM-LV and CASME II validate the proposed methods are superior to the state-of-the-art methods.
****
Linwei Tao,Younan Zhu,Haolan Guo,Minjing Dong,Chang Xu
A Benchmark Study on Calibration
Deep neural networks are increasingly utilized in various machine learning tasks. However, as these models grow in complexity, they often face calibration issues, despite enhanced prediction accuracy. Many studies have endeavored to improve calibration performance through the use of specific loss functions, data preprocessing and training frameworks. Yet, investigations into calibration properties have been somewhat overlooked. Our study leverages the Neural Architecture Search (NAS) search space, offering an exhaustive model architecture space for thorough calibration properties exploration. We specifically create a model calibration dataset. This dataset evaluates 90 bin-based and 12 additional calibration measurements across 117,702 unique neural networks within the widely employed NATS-Bench search space. Our analysis aims to answer several longstanding questions in the field, using our proposed dataset: (i) Can model calibration be generalized across different datasets? (ii) Can robustness be used as a calibration measurement? (iii) How reliable are calibration metrics? (iv) Does a post-hoc calibration method affect all models uniformly? (v) How does calibration interact with accuracy? (vi) What is the impact of bin size on calibration measurement? (vii) Which architectural designs are beneficial for calibration? Additionally, our study bridges an existing gap by exploring calibration within NAS. By providing this dataset, we enable further research into NAS calibration. As far as we are aware, our research represents the first large-scale investigation into calibration properties and the premier study of calibration issues within NAS.
****
Gregory Kang Ruey Lau,Apivich Hemachandra,See-Kiong Ng,Bryan Kian Hsiang Low
PINNACLE: PINN Adaptive ColLocation and Experimental points selection
Physics-Informed Neural Networks (PINNs), which incorporate PDEs as soft constraints, train with a composite loss function that contains multiple training point types: different types of collocation points chosen during training to enforce each PDE and initial/boundary conditions, and experimental points which are usually costly to obtain via experiments or simulations. Training PINNs using this loss function is challenging as it typically requires selecting large numbers of points of different types, each with different training dynamics. Unlike past works that focused on the selection of either collocation or experimental points, this work introduces PINN Adaptive ColLocation and Experimental points selection (PINNACLE), the first algorithm that jointly optimizes the selection of all training point types, while automatically adjusting the proportion of collocation point types as training progresses. PINNACLE uses information on the interactions among training point types, which had not been considered before, based on an analysis of PINN training dynamics via the Neural Tangent Kernel (NTK). We theoretically show that the criterion used by PINNACLE is related to the PINN generalization error, and empirically demonstrate that PINNACLE is able to outperform existing point selection methods for forward, inverse, and transfer learning problems.
****
Javier Rando,Florian Tramèr
Universal Jailbreak Backdoors from Poisoned Human Feedback
Reinforcement Learning from Human Feedback (RLHF) is used to align large language models to produce helpful and harmless responses. Yet, these models can be jailbroken by finding adversarial prompts that revert the model to its unaligned behavior. In this paper, we consider a new threat where an attacker poisons the RLHF data to embed a jailbreak trigger into the model as a backdoor. The trigger then acts like a universal sudo command, enabling arbitrary harmful responses without the need to search for an adversarial prompt. Universal jailbreak backdoors are much more powerful than previously studied backdoors on language models, and we find they are significantly harder to plant using common backdoor attack t

echniques. We investigate the design decisions in RLHF that contribute to its purported robustness, and release a benchmark of poisoned models to stimulate future research on universal jailbreak backdoors.
****
Vint Lee,Pieter Abbeel,Youngwoon Lee
DreamSmooth: Improving Model-based Reinforcement Learning via Reward Smoothing
Model-based reinforcement learning (MBRL) has gained much attention for its ability to learn complex behaviors in a sample-efficient way: planning actions by generating imaginary trajectories with predicted rewards. Despite its success, we found that surprisingly, reward prediction is often a bottleneck of MBRL, especially for sparse rewards that are challenging (or even ambiguous) to predict. Motivated by the intuition that humans can learn from rough reward estimates, we propose a simple yet effective reward smoothing approach, DreamSmooth, which learns to predict a temporally-smoothed reward, instead of the exact reward at the given timestep. We empirically show that DreamSmooth achieves state-of-the-art performance on long-horizon sparse-reward tasks both in sample efficiency and final performance without losing performance on common benchmarks, such as Deepmind Control Suite and Atari benchmarks.
****
Joey Hong,Anca Dragan,Sergey Levine
Offline RL with Observation Histories: Analyzing and Improving Sample Complexity
Offline reinforcement learning (RL) can in principle synthesize more optimal behavior from a dataset consisting only of suboptimal trials. One way that this can happen is by "stitching" together the best parts of otherwise suboptimal trajectories that overlap on similar states, to create new behaviors where each individual state is in-distribution, but the overall returns are higher. However, in many interesting and complex applications, such as autonomous navigation and dialogue systems, the state is partially observed. Even worse, the state representation is unknown or not easy to define. In such cases, policies and value functions are often conditioned on observation histories instead of states. In these cases, it is not clear if the same kind of "stitching" is feasible at the level of observation histories, since two different trajectories would always have different histories, and thus "similar states" that might lead to effective stitching cannot be leveraged. Theoretically, we show that standard offline RL algorithms conditioned on observation histories suffer from poor sample complexity, in accordance with the above intuition. We then identify sufficient conditions under which offline RL can still be efficient -- intuitively, it needs to learn a compact representation of history comprising only features relevant for action selection. We introduce a bisimulation loss that captures the extent to which this happens, and propose that offline RL can explicitly optimize this loss to aid worst-case sample complexity. Empirically, we show that across a variety of tasks either our proposed loss improves performance, or the value of this loss is already minimized as a consequence of standard offline RL, indicating that it correlates well with good performance.
****
Jiashun Jin,Tracy Ke,Gabriel Moryoussef,Jiajun Tang,Jingming Wang
Improved algorithm and bounds for successive projection
Consider a K-vertex simplex in a d-dimensional space. We measure n points on the simplex, but due to the measurement noise,
some of the observed points fall outside the simplex. The interest is vertex hunting (i.e., estimating the vertices of the simplex). The successive projection algorithm (SPA) is one of the most popular approaches to vertex hunting, but it is vulnerable to noise and outliers, and may perform unsatisfactorily. We propose pseudo-point SPA (pp-SPA) as a new approach to vertex hunting. The approach contains
two novel ideas (a projection step and a denoise step) and generates roughly n pseudo-points, which can be fed in to SPA for vertex hunting. For theory, we first derive an improved non-asymptotic bound for the orthodox SPA, and then use the result to derive the bounds for pp-SPA. Compared with the orthodox SPA, pp-SPA has a faster rate and more satisfactory numerical performance in a broad set

ting. The analysis is quite delicate: the non-asymptotic bound is hard to derive, and we need precise results on the extreme values of (possibly) high-dimensional random vectors.
****
Mengkang Hu,Yao Mu,Xinmiao Chelsey Yu,Mingyu Ding,Shiguang Wu,Wenqi Shao,Qiguang Chen,Bin Wang,Yu Qiao,Ping Luo
Tree-Planner: Efficient Close-loop Task Planning with Large Language Models
This paper studies close-loop task planning, which refers to the process of generating a sequence of skills (a plan) to accomplish a specific goal while adapting the plan based on real-time observations.
Recently, prompting Large Language Models (LLMs) to generate actions iteratively has become a prevalent paradigm due to its superior performance and user-friendliness.
However, this paradigm is plagued by two inefficiencies: high token consumption and redundant error correction, both of which hinder its scalability for large-scale testing and applications.
To address these issues, we propose Tree-Planner, which reframes task planning with LLMs into three distinct phases:
plan sampling, action tree construction, and grounded deciding.
Tree-Planner starts by using an LLM to sample a set of potential plans before execution, followed by the aggregation of them to form an action tree.
Finally, the LLM performs a top-down decision-making process on the tree, taking into account real-time environmental information.
Experiments show that Tree-Planner achieves state-of-the-art performance while maintaining high efficiency.
By decomposing LLM queries into a single plan-sampling call and multiple grounded-deciding calls,
a considerable part
of the prompt are less likely to be repeatedly consumed.
As a result, token consumption is reduced by 92.2\% compared to the previously best-performing model.
Additionally, by enabling backtracking on the action tree as needed, the correction process becomes more flexible, leading to a 40.5\% decrease in error corrections.
****
Stephanie Fu,Mark Hamilton,Laura E. Brandt,Axel Feldmann,Zhoutong Zhang,William T. Freeman
FeatUp: A Model-Agnostic Framework for Features at Any Resolution
Deep features are a cornerstone of computer vision research, capturing image semantics and enabling the community to solve downstream tasks even in the zero- or few-shot regime. However, these features often lack the spatial resolution to directly perform dense prediction tasks like segmentation and depth prediction because models aggressively pool information over large areas. In this work, we introduce FeatUp, a task- and model-agnostic framework to restore lost spatial information in deep features. We introduce two variants of FeatUp: one that guides features with high-resolution signal in a single forward pass, and one that fits an implicit model to a single image to reconstruct features at any resolution. Both approaches use a multi-view consistency loss with deep analogies to NeRFs. Our features retain their original semantics and can be swapped into existing applications to yield resolution and performance gains even without re-training. We show that FeatUp significantly outperforms other feature upsampling and image super-resolution approaches in class activation map generation, transfer learning for segmentation and depth prediction, and end-to-end training for semantic segmentation.
****
Matthew Morris,Bernardo Cuenca Grau,Ian Horrocks
Orbit-Equivariant Graph Neural Networks
Equivariance is an important structural property that is captured by architectures such as graph neural networks (GNNs). However, equivariant graph functions cannot produce different outputs for similar nodes, which may be undesirable when

the function is trying to optimize some global graph property. In this paper, we define orbit-equivariance, a relaxation of equivariance which allows for such functions whilst retaining important structural inductive biases. We situate the property in the hierarchy of graph functions, define a taxonomy of orbit-equivariant functions, and provide four different ways to achieve non-equivariant GNNs. For each, we analyze their expressivity with respect to orbit-equivariance and evaluate them on two novel datasets, one of which stems from a real-world use-case of designing optimal bioisosteres.
****
Yihao Xue,Eric Gan,Jiayi Ni,Siddharth Joshi,Baharan Mirzasoleiman
Investigating the Benefits of Projection Head for Representation Learning
An effective technique for obtaining high-quality representations is adding a projection head on top of the encoder during training, then discarding it and using the pre-projection representations. Despite its proven practical effectiveness, the reason behind the success of this technique is poorly understood. The pre-projection representations are not directly optimized by the loss function, raising the question: what makes them better? In this work, we provide a rigorous theoretical answer to this question. We start by examining linear models trained with self-supervised contrastive loss. We reveal that the implicit bias of training algorithms leads to layer-wise progressive feature weighting, where features become increasingly unequal as we go deeper into the layers. Consequently, lower layers tend to have more normalized and less specialized representations. We theoretically characterize scenarios where such representations are more beneficial, highlighting the intricate interplay between data augmentation and input features. Additionally, we demonstrate that introducing non-linearity into the network allows lower layers to learn features that are completely absent in higher layers. Finally, we show how this mechanism improves the robustness in supervised contrastive learning and supervised learning. We empirically validate our results through various experiments on CIFAR-10/100, UrbanCars and shifted versions of ImageNet. We also introduce a potential alternative to projection head, which offers a more interpretable and controllable design.
****
Suhwan Choi,Myeongho Jeon,Yeonjung Hwang,Jeonglyul Oh,Sungjun Lim,Joonseok Lee,Myungjoo Kang
Dictionary Contrastive Learning for Efficient Local Supervision without Auxiliary Networks
While backpropagation (BP) has achieved widespread success in deep learning, it faces two prominent challenges: computational inefficiency and biological implausibility.
In response to these challenges, local supervision, encompassing Local
Learning (LL) and Forward Learning (FL), has emerged as a promising research
direction. LL employs module-wise BP to achieve competitive results yet relies on
module-wise auxiliary networks, which increase memory and parameter demands.
Conversely, FL updates layer weights without BP and auxiliary networks but falls
short of BP’s performance. This paper proposes a simple yet effective objective
within a contrastive learning framework for local supervision without auxiliary
networks. Given the insight that the existing contrastive learning framework for
local supervision is susceptible to task-irrelevant information without auxiliary
networks, we present DICTIONARY CONTRASTIVE LEARNING (DCL) that optimizes
the similarity between local features and label embeddings. Our method
using static label embeddings yields substantial performance improvements in the
FL scenario, outperforming state-of-the-art FL approaches. Moreover, our method
using adaptive label embeddings closely approaches the performance achieved by
LL while achieving superior memory and parameter efficiency.
****
David Ireland,Giovanni Montana
REValueD: Regularised Ensemble Value-Decomposition for Factorisable Markov Decision Processes

Discrete-action reinforcement learning algorithms often falter in tasks with hig
h-dimensional discrete action spaces due to the vast number of possible actions.
 A recent advancement leverages value-decomposition, a concept from multi-agent
reinforcement learning, to tackle this challenge. This study delves deep into th
e effects of this value-decomposition, revealing that whilst it curtails the ove
r-estimation bias inherent to Q-learning algorithms, it amplifies target varianc
e. To counteract this, we present an ensemble of critics to mitigate target vari
ance. Moreover, we introduce a regularisation loss that helps to mitigate the ef
fects that exploratory actions in one dimension can have on the value of optimal
 actions in other dimensions. Our novel algorithm, REValueD, tested on discretis
ed versions of the DeepMind Control Suite tasks, showcases superior performance,
 especially in the challenging humanoid and dog tasks. We further dissect the fa
ctors influencing REValueD's performance, evaluating the significance of the reg
ularisation loss and the scalability of REValueD with increasing sub-actions per
 dimension.
****
Minjun Sung,Sambhu Harimanas Karumanchi,Aditya Gahlawat,Naira Hovakimyan
Robust Model Based Reinforcement Learning Using $\mathcal{L}_1$ Adaptive Control
We introduce $\mathcal{L}_1$-MBRL, a control-theoretic augmentation scheme for M
odel-Based Reinforcement Learning (MBRL) algorithms. Unlike model-free approache
s, MBRL algorithms learn a model of the transition function using data and use i
t to design a control input. Our approach generates a series of approximate cont
rol-affine models of the learned transition function according to the proposed s
witching law. Using the approximate model, control input produced by the underly
ing MBRL is perturbed by the $\mathcal{L}_1$ adaptive control, which is designed
 to enhance the robustness of the system against uncertainties. Importantly, thi
s approach is agnostic to the choice of MBRL algorithm, enabling the use of the
 scheme with various MBRL algorithms. MBRL algorithms with $\mathcal{L}_1$ augme
ntation exhibit enhanced performance and sample efficiency across multiple MuJoC
o environments, outperforming the original MBRL algorithms, both with and withou
t system noise.
****
Yan Sun,Jicong Fan
MMD Graph Kernel: Effective Metric Learning for Graphs via Maximum Mean Discrepa
ncy
This paper focuses on graph metric learning. First, we present a class of maximu
m mean discrepancy (MMD) based graph kernels, called MMD-GK. These kernels are c
omputed by applying MMD to the node representations of two graphs with message-p
assing propagation.
Secondly, we provide a class of deep MMD-GKs that are able to learn graph kernel
s and implicit graph features adaptively in an unsupervised manner. Thirdly, we
propose a class of supervised deep MMD-GKs that are able to utilize label inform
ation of graphs and hence yield more discriminative metrics. Besides the algorit
hms, we provide theoretical analysis for the proposed methods. The proposed meth
ods are evaluated in comparison to many baselines such as graph kernels and grap
h neural networks in the tasks of graph clustering and graph classification. The
 numerical results demonstrate the effectiveness and superiority of our methods.
****
Gabriele Tiboni,Pascal Klink,Jan Peters,Tatiana Tommasi,Carlo D'Eramo,Georgia Ch
alvatzaki
Domain Randomization via Entropy Maximization
Varying dynamics parameters in simulation is a popular Domain Randomization (DR)
 approach for overcoming the reality gap in Reinforcement Learning (RL). Neverth
eless, DR heavily hinges on the choice of the sampling distribution of the dynam
ics parameters, since high variability is crucial to regularize the agent's beha
vior but notoriously leads to overly conservative policies when randomizing exce
ssively. In this paper, we propose a novel approach to address sim-to-real trans
fer, which automatically shapes dynamics distributions during training in simula
tion without requiring real-world data. We introduce DOmain RAndomization via En
tropy MaximizatiON (DORAEMON), a constrained optimization problem that directly

maximizes the entropy of the training distribution while retaining generalizatio n capabilities. In achieving this, DORAEMON gradually increases the diversity of sampled dynamics parameters as long as the probability of success of the curren t policy is sufficiently high. We empirically validate the consistent benefits o f DORAEMON in obtaining highly adaptive and generalizable policies, i.e. solving the task at hand across the widest range of dynamics parameters, as opposed to representative baselines from the DR literature. Notably, we also demonstrate th e Sim2Real applicability of DORAEMON through its successful zero-shot transfer i n a robotic manipulation setup under unknown real-world parameters.
****
Yuxin Dong,Tieliang Gong,Hong Chen,Shujian Yu,Chen Li
Rethinking Information-theoretic Generalization: Loss Entropy Induced PAC Bounds
Information-theoretic generalization analysis has achieved astonishing success i n characterizing the generalization capabilities of noisy and iterative learning algorithms. However, current advancements are mostly restricted to average-case scenarios and necessitate the stringent bounded loss assumption, leaving a gap with regard to computationally tractable PAC generalization analysis, especially for long-tailed loss distributions. In this paper, we bridge this gap by introd ucing a novel class of PAC bounds through leveraging loss entropies. These bound s simplify the computation of key information metrics in previous PAC informatio n-theoretic bounds to one-dimensional variables, thereby enhancing computational tractability. Moreover, our data-independent bounds provide novel insights into the generalization behavior of the minimum error entropy criterion, while our d ata-dependent bounds improve over previous results by alleviating the bounded lo ss assumption under both leave-one-out and supersample settings. Extensive numer ical studies indicate strong correlations between the generalization error and t he induced loss entropy, showing that the presented bounds adeptly capture the p atterns of the true generalization gap under various learning scenarios.
****
Li Siyao,Tianpei Gu,Zhitao Yang,Zhengyu Lin,Ziwei Liu,Henghui Ding,Lei Yang,Chen Change Loy
Duolando: Follower GPT with Off-Policy Reinforcement Learning for Dance Accompan iment
We introduce a novel task within the field of human motion generation, termed da nce accompaniment, which necessitates the generation of responsive movements fro m a dance partner, the "follower", synchronized with the lead dancer’s movements and the underlying musical rhythm. Unlike existing solo or group dance generati on tasks, a duet dance scenario entails a heightened degree of interaction betwe en the two participants, requiring delicate coordination in both pose and positi on. To support this task, we first build a large-scale and diverse duet interact ive dance dataset, DD100, by recording about 117 minutes of professional dancers ’ performances. To address the challenges inherent in this task, we propose a GP T based model, Duolando, which autoregressively predicts the subsequent tokenize d motion conditioned on the coordinated information of the music, the leader’s a nd the follower’s movements. To further enhance the GPT’s capabilities of genera ting stable results on unseen conditions (music and leader motions), we devise a n off-policy reinforcement learning strategy that allows the model to explore vi able trajectories from out-of-distribution samplings, guided by human-defined re wards. Based on the collected dataset and proposed method, we establish a benchm ark with several carefully designed metrics.
****
Kyungmin Lee,Kihyuk Sohn,Jinwoo Shin
DreamFlow: High-quality text-to-3D generation by Approximating Probability Flow
Recent progress in text-to-3D generation has been achieved through the utilizati on of score distillation methods: they make use of the pre-trained text-to-image (T2I) diffusion models by distilling via the diffusion model training objective . However, such an approach inevitably results in the use of random timesteps at each update, which increases the variance of the gradient and ultimately prolon gs the optimization process. In this paper, we propose to enhance the text-to-3D optimization by leveraging the T2I diffusion prior in the generative sampling p

rocess with a predetermined timestep schedule. To this end, we interpret text-to-3D optimization as a multi-view image-to-image translation problem, and propose a solution by approximating the probability flow. By leveraging the proposed novel optimization algorithm, we design DreamFlow, a practical three-stage coarse-to-fine text-to-3D optimization framework that enables fast generation of high-quality and high-resolution (i.e., 1024×1024) 3D contents. For example, we demonstrate that DreamFlow is 5 times faster than the existing state-of-the-art text-to-3D method, while producing more photorealistic 3D contents.
****
Roi Benita,Michael Elad,Joseph Keshet
DiffAR: Denoising Diffusion Autoregressive Model for Raw Speech Waveform Generation
Diffusion models have recently been shown to be relevant for high-quality speech generation. Most work has been focused on generating spectrograms, and as such, they further require a subsequent model to convert the spectrogram to a waveform (i.e., a vocoder). This work proposes a diffusion probabilistic end-to-end model for generating a raw speech waveform. The proposed model is autoregressive, generating overlapping frames sequentially, where each frame is conditioned on a portion of the previously generated one. Hence, our model can effectively synthesize an unlimited speech duration while preserving high-fidelity synthesis and temporal coherence. We implemented the proposed model for unconditional and conditional speech generation, where the latter can be driven by an input sequence of phonemes, amplitudes, and pitch values. Working on the waveform directly has some empirical advantages. Specifically, it allows the creation of local acoustic behaviors, like vocal fry, which makes the overall waveform sounds more natural. Furthermore, the proposed diffusion model is stochastic and not deterministic; therefore, each inference generates a slightly different waveform variation, enabling abundance of valid realizations. Experiments show that the proposed model generates speech with superior quality compared with other state-of-the-art neural speech generation systems.
****
Nirmit Joshi,Gal Vardi,Nathan Srebro
Noisy Interpolation Learning with Shallow Univariate ReLU Networks
Understanding how overparameterized neural networks generalize despite perfect interpolation of noisy training data is a fundamental question. Mallinar et. al. (2022) noted that neural networks seem to often exhibit ``tempered overfitting'', wherein the population risk does not converge to the Bayes optimal error, but neither does it approach infinity, yielding non-trivial generalization. However, this has not been studied rigorously. We provide the first rigorous analysis of the overfiting behaviour of regression with minimum norm (ℓ_2 of weights), focusing on univariate two-layer ReLU networks. We show overfitting is tempered (with high probability) when measured with respect to the L_1 loss, but also show that the situation is more complex than suggested by Mallinar et. al., and overfitting is catastrophic with respect to the L_2 loss, or when taking an expectation over the training set.
****
Fan-Ming Luo,Tian Xu,Xingchen Cao,Yang Yu
Reward-Consistent Dynamics Models are Strongly Generalizable for Offline Reinforcement Learning
Learning a precise dynamics model can be crucial for offline reinforcement learning, which, unfortunately, has been found to be quite challenging. Dynamics models that are learned by fitting historical transitions often struggle to generalize to unseen transitions. In this study, we identify a hidden but pivotal factor termed dynamics reward that remains consistent across transitions, offering a pathway to better generalization. Therefore, we propose the idea of reward-consistent dynamics models: any trajectory generated by the dynamics model should maximize the dynamics reward derived from the data. We implement this idea as the MOREC (Model-based Offline reinforcement learning with Reward Consistency) method, which can be seamlessly integrated into previous offline model-based reinforcement learning (MBRL) methods. MOREC learns a generalizable dynamics reward functi

on from offline data, which is subsequently employed as a transition filter in a
ny offline MBRL method: when generating transitions, the dynamics model generate
s a batch of transitions and selects the one with the highest dynamics reward va
lue. On a synthetic task, we visualize that MOREC has a strong generalization ab
ility and can surprisingly recover some distant unseen transitions. On 21 offlin
e tasks in D4RL and NeoRL benchmarks, MOREC improves the previous state-of-the-a
rt performance by a significant margin, i.e., 4.6\% on D4RL tasks and 25.9\% on
NeoRL tasks. Notably, MOREC is the first method that can achieve above 95\% onli
ne RL performance in 6 out of 12 D4RL tasks and 3 out of 9 NeoRL tasks. Code is
available at https://github.com/polixir/morec.
****
Xuanlei Zhao,Shenggan Cheng,Guangyang LU,Haotian Zhou,Bin Jia,Yang You
AutoChunk: Automated Activation Chunk for Memory-Efficient Deep Learning Inferen
ce
Large deep learning models have achieved impressive performance across a range o
f applications. However, their large memory requirements, including parameter me
mory and activation memory, have become a significant challenge for their practi
cal serving. While existing methods mainly address parameter memory, the importa
nce of activation memory has been overlooked. Especially for long input sequence
s, activation memory is expected to experience a significant exponential growth
as the length of sequences increases. In this approach, we propose AutoChunk, an
 automatic and adaptive compiler system that efficiently reduces activation memo
ry for long sequence inference by chunk strategies. The proposed system generate
s chunk plans by optimizing through multiple stages. In each stage, the chunk se
arch pass explores all possible chunk candidates and the chunk selection pass id
entifies the optimal one. At runtime, AutoChunk employs code generation to autom
atically apply chunk strategies. The experiments demonstrate that AutoChunk can
reduce over 80% of activation memory while maintaining speed loss within 10%, ex
tend max sequence length by 3.2x to 11.7x, and outperform state-of-the-art metho
ds by a large margin.
****
Lukas Berglund,Meg Tong,Maximilian Kaufmann,Mikita Balesni,Asa Cooper Stickland,
Tomasz Korbak,Owain Evans
The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”
We expose a surprising failure of generalization in auto-regressive large langua
ge models (LLMs). If a model is trained on a sentence of the form ''_A_ is _B_''
, it will not automatically generalize to the reverse direction ''_B_ is _A_''.
This is the **Reversal Curse**. For instance, if a model is trained on ''Valenti
na Tereshkova was the first woman to travel to space'', it will not automaticall
y be able to answer the question, ''Who was the first woman to travel to space?'
'. Moreover, the likelihood of the correct answer (''Valentina Tershkova'') will
 not be higher than for a random name. Thus, models do not generalize a prevalen
t pattern in their training set: if ''_A_ is _B_'' occurs, ''_B_ is _A_'' is mor
e likely to occur. It is worth noting, however, that if ''_A_ is _B_'' appears _
in-context_, models can deduce the reverse relationship.

We provide evidence for the Reversal Curse by finetuning GPT-3 and Llama-1 on fi
ctitious statements such as ''Uriah Hawthorne is the composer of _Abyssal Melodi
es_'' and showing that they fail to correctly answer ''Who composed _Abyssal Mel
odies?_''. The Reversal Curse is robust across model sizes and model families an
d is not alleviated by data augmentation.

We also evaluate ChatGPT (GPT-3.5 and GPT-4) on questions about real-world celeb
rities, such as ''Who is Tom Cruise's mother? [A: Mary Lee Pfeiffer]'' and the r
everse ''Who is Mary Lee Pfeiffer's son?''. GPT-4 correctly answers questions li
ke the former 79\% of the time, compared to 33\% for the latter.
****
Parth Sarthi,Salman Abdullah,Aditi Tuli,Shubh Khanna,Anna Goldie,Christopher D M
anning
RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval

Retrieval-augmented language models can better adapt to changes in world state and incorporate long-tail knowledge. However, most existing methods retrieve only short contiguous chunks from a retrieval corpus, limiting holistic understanding of the overall document context. We introduce the novel approach of recursively embedding, clustering, and summarizing chunks of text, constructing a tree with differing levels of summarization from the bottom up. At inference time, our RAPTOR model retrieves from this tree, integrating information across lengthy documents at different levels of abstraction. Controlled experiments show that retrieval with recursive summaries offers significant improvements over traditional retrieval-augmented LMs on several tasks. On question-answering tasks that involve complex, multi-step reasoning, we show state-of-the-art results; for example, by coupling RAPTOR retrieval with the use of GPT-4, we can improve the best performance on the QuALITY benchmark by 20\% in absolute accuracy.
****
Lukas Fesser,Melanie Weber
Effective Structural Encodings via Local Curvature Profiles
Structural and Positional Encodings can significantly improve the performance of Graph Neural Networks in downstream tasks. Recent literature has begun to systematically investigate differences in the structural properties that these approaches encode, as well as performance trade-offs between them. However, the question of which structural properties yield the most effective encoding remains open. In this paper, we investigate this question from a geometric perspective. We propose a novel structural encoding based on discrete Ricci curvature (Local Curvature Profiles, short LCP) and show that it significantly outperforms existing encoding approaches. We further show that combining local structural encodings, such as LCP, with global positional encodings improves downstream performance, suggesting that they capture complementary geometric information. Finally, we compare different encoding types with (curvature-based) rewiring techniques. Rewiring has recently received a surge of interest due to its ability to improve the performance of Graph Neural Networks by mitigating over-smoothing and over-squashing effects. Our results suggest that utilizing curvature information for structural encodings delivers significantly larger performance increases than rewiring.
****
Noam Itzhak Levi,Alon Beck,Yohai Bar-Sinai
Grokking in Linear Estimators -- A Solvable Model that Groks without Understanding
Grokking is the intriguing phenomenon where a model learns to generalize long after it has fit the training data.

We show both analytically and numerically that grokking can surprisingly occur in linear networks performing linear tasks in a simple teacher-student setup. In this setting, the full training dynamics is derived in terms of the expected training and generalization data covariance matrix. We present exact predictions on how the grokking time depends on input and output dimensionality, train sample size, regularization, and network parameters initialization.

The key findings are that late generalization increase may not imply a transition from "memorization" to "understanding", but can simply be an artifact of the accuracy measure.

We provide empirical verification for these propositions, along with preliminary results indicating that some predictions also hold for deeper networks, with non-linear activations.
****
Yangjun Ruan,Honghua Dong,Andrew Wang,Silviu Pitis,Yongchao Zhou,Jimmy Ba,Yann Dubois,Chris J. Maddison,Tatsunori Hashimoto
Identifying the Risks of LM Agents with an LM-Emulated Sandbox
Recent advances in Language Model (LM) agents and tool use, exemplified by applications like ChatGPT Plugins, enable a rich set of capabilities but also amplify potential risks—such as leaking private data or causing financial losses. Identifying these risks is labor-intensive, necessitating implementing the tools, setting up the environment for each test scenario manually, and finding risky cases. As tools and agents become more complex, the high cost of testing these agents

will make it increasingly difficult to find high-stakes, long-tail risks. To ad dress these challenges, we introduce ToolEmu: a framework that uses an LM to emu late tool execution and enables scalable testing of LM agents against a diverse range of tools and scenarios. Alongside the emulator, we develop an LM-based aut omatic safety evaluator that examines agent failures and quantifies associated r isks. We test both the tool emulator and evaluator through human evaluation and find that 68.8% of failures identified with ToolEmu would be valid real-world ag ent failures. Using our curated initial benchmark consisting of 36 high-stakes t oolkits and 144 test cases, we provide a quantitative risk analysis of current L M agents and identify numerous failures with potentially severe outcomes. Notabl y, even the safest LM agent exhibits such failures 23.9% of the time according t o our evaluator, underscoring the need to develop safer LM agents for real-world deployment.
****

Pangpang Liu,Yichuan Zhao
Empirical Likelihood for Fair Classification
Machine learning algorithms are commonly being deployed in decision-making syste ms that have a direct impact on human lives. However, if these algorithms are tr ained solely to minimize training/test errors, they may inadvertently discrimina te against individuals based on their sensitive attributes, such as gender, race or age. Recently, algorithms that ensure the fairness are developed in the mach ine learning community. Fairness criteria are applied by these algorithms to mea sure the fairness, but they often use the point estimate to assess the fairness and fail to consider the uncertainty of the sample fairness criterion once the a lgorithms are deployed. We suggest that assessing the fairness should take the u ncertainty into account. In this paper, we use the covariance as a proxy for the fairness and develop the confidence region of the covariance vector using empir ical likelihood \citep{Owen1988}. Our confidence region based fairness constrain ts for classification take uncertainty into consideration during fairness assess ment. The proposed confidence region can be used to test the fairness and impose fairness constraint using the significant level as a tool to balance the accura cy and fairness.   Simulation studies show that our method exactly covers the ta rget Type I error rate and effectively balances the trade-off between accuracy a nd fairness. Finally, we conduct data analysis to demonstrate the effectiveness of our method.
****

Ruocheng Wang,Eric Zelikman,Gabriel Poesia,Yewen Pu,Nick Haber,Noah Goodman
Hypothesis Search: Inductive Reasoning with Language Models
Inductive reasoning is a core problem-solving capacity: humans can identify unde rlying principles from a few examples, which can then be robustly generalized to novel scenarios. Recent work has evaluated large language models (LLMs) on indu ctive reasoning tasks by directly prompting them yielding "in context learning." This can work well for straightforward inductive tasks, but performs very poorl y on more complex tasks such as the Abstraction and Reasoning Corpus (ARC). In t his work, we propose to improve the inductive reasoning ability of LLMs by gener ating explicit hypotheses at multiple levels of abstraction: we prompt the LLM t o propose multiple abstract hypotheses about the problem, in natural language, t hen implement the natural language hypotheses as concrete Python programs. These programs can be directly verified by running on the observed examples and gener alized to novel inputs. To reduce the hypothesis search space, we explore steps to filter the set of hypotheses to be implemented as programs: we either ask the LLM to summarize them into a smaller set of hypotheses, or ask human annotators to select a subset. We verify our pipeline's effectiveness on the ARC visual in ductive reasoning benchmark, its variant 1D-ARC, and string transformation datas et SyGuS. On a random 40-problem subset of ARC, our automated pipeline using LLM summaries achieves 27.5% accuracy, significantly outperforming the direct promp ting baseline (accuracy of 12.5%). With the minimal human input of selecting fro m LLM-generated candidates, the performance is boosted to 37.5%. Our ablation st udies show that abstract hypothesis generation and concrete program representati ons are both beneficial for LLMs to perform inductive reasoning tasks.

****
Hangting Ye,Wei Fan,Xiaozhuang Song,Shun Zheng,He Zhao,Dan dan Guo,Yi Chang
PTaRL: Prototype-based Tabular Representation Learning via Space Calibration
Tabular data have been playing a mostly important role in diverse real-world fie lds, such as healthcare, engineering, finance, etc.
With the recent success of deep learning, many tabular machine learning (ML) met hods based on deep networks (e.g., Transformer, ResNet) have achieved competitiv e performance on tabular benchmarks. However, existing deep tabular ML methods s uffer from the representation entanglement and localization, which largely hinde rs their prediction performance and leads to performance inconsistency on tabul ar tasks.
To overcome these problems, we explore a novel direction of applying prototype l earning for tabular ML and propose a prototype-based tabular representation lear ning framework, PTaRL, for tabular prediction tasks. The core idea of PTaRL is t o construct prototype-based projection space (P-Space) and learn the disentangle d representation around global data prototypes. Specifically, PTaRL mainly invol ves two stages: (i) Prototype Generating, that constructs global prototypes as t he basis vectors of P-Space for representation, and (ii) Prototype Projecting, t hat projects the data samples into P-Space and keeps the core global data inform ation via Optimal Transport. Then, to further acquire the disentangled represent ations, we constrain PTaRL with two strategies: (i) to diversify the coordinates towards global prototypes of different representations within P-Space, we bring up a diversifying constraint for representation calibration; (ii) to avoid prot otype entanglement in P-Space, we introduce a matrix orthogonalization constrain t to ensure the independence of global prototypes.
Finally, we conduct extensive experiments in PTaRL coupled with state-of-the-art deep tabular ML models on various tabular benchmarks and the results have shown our consistent superiority.
****
Sehyun Kwon,Jaeseung Park,Minkyu Kim,Jaewoong Cho,Ernest K. Ryu,Kangwook Lee
Image Clustering Conditioned on Text Criteria
Classical clustering methods do not provide users with direct control of the clu stering results, and the clustering results may not be consistent with the relev ant criterion that a user has in mind. In this work, we present a new methodolog y for performing image clustering based on user-specified criteria in the form o f text by leveraging modern Vision-Language Models and Large Language Models. We call our method Image Clustering Conditioned on Text Criteria (IC$|$TC), and it represents a different paradigm of image clustering. IC$|$TC requires a minimal and practical degree of human intervention and grants the user significant cont rol over the clustering results in return. Our experiments show that IC$|$TC can effectively cluster images with various criteria, such as human action, physica l location, or the person's mood, significantly outperforming baselines.
****
Haowen Wang,Tao Sun,Congyun Jin,Yingbo Wang,Yibo Fan,Yunqi Xu,Yuliang Du,Cong Fa n
Customizable Combination of Parameter-Efficient Modules for Multi-Task Learning
Modular skill learning is an emerging direction in the field of Parameter Effici ent Fine-Tuning (PEFT), as it enables neural networks to better organize and cla rify various aspects of knowledge, leading to improved knowledge transfer for ne w tasks. In this paper, we introduce a novel approach that categorizes skills in to shared domain skills and specialized skills, with the skill parameters being highly parameterized using low-rank or sparse techniques. Each task is associate d with an exclusive specialized skill while also benefiting from shared domain s kills. Moreover, tasks can selectively utilize specialized skills from other tas ks as needed. To facilitate this approach, we propose a skill assignment matrix that can be jointly learned, and the task network is instantiated based on the s kill parameters. To evaluate the effectiveness of our approach, we conducted ext ensive experiments on the Super Natural Instructions and SuperGLUE datasets. Our results demonstrate that compared to fully-shared, task-specific, or skill-indi stinguishable baselines. Modular learning with skill-type discrimination signifi

cantly enhances the sample efficiency of multi-task learning. Furthermore, the f reezing of a substantial number of base model parameters greatly improves parame ter efficiency, leading to boosted training efficiency.
****
Lifan Yuan,Yangyi Chen,Xingyao Wang,Yi Fung,Hao Peng,Heng Ji
CRAFT: Customizing LLMs by Creating and Retrieving from Specialized Toolsets
Large language models (LLMs) are often augmented with tools to solve complex tas ks. By generating code snippets and executing them through task-specific Applica tion Programming Interfaces (APIs), they can offload certain functions to dedica ted external modules, such as image encoding and performing calculations. Howeve r, most existing approaches to augment LLMs with tools are constrained
by general-purpose APIs and lack the flexibility for tailoring them to specific tasks. In this work, we present CRAFT, a general tool creation and retrieval fra mework for LLMs. It creates toolsets specifically curated for the tasks and equi ps LLMs with a component that retrieves tools from these sets to enhance their c apability to solve complex tasks. For each task, we collect specific code soluti ons by prompting
GPT-4 to solve the training examples. Following a validation step ensuring the c orrectness, these solutions are abstracted into code snippets to enhance reusabi lity, and deduplicated for higher quality. At inference time, the language model retrieves snippets from the toolsets and then executes them or generates the ou tput conditioning on the retrieved snippets. Our method is designed to be flexib le and
offers a plug-and-play approach to adapt off-the-shelf LLMs to unseen domains an d modalities, without any finetuning. Experiments on vision-language, tabular pr ocessing, and mathematical reasoning tasks show that our approach achieves subst antial improvements compared to strong baselines. In addition, our in-depth anal ysis reveals that: (1) consistent performance improvement can be achieved by
scaling up the number of tools and the capability of the backbone models; (2) ea ch component of our approach contributes to the performance gains; (3) the creat ed tools are well-structured and reliable with low complexity and atomicity.
****
Yuwei Guo,Ceyuan Yang,Anyi Rao,Zhengyang Liang,Yaohui Wang,Yu Qiao,Maneesh Agraw ala,Dahua Lin,Bo Dai
AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Sp ecific Tuning
With the advance of text-to-image (T2I) diffusion models (e.g., Stable Diffusion) and corresponding personalization techniques such as DreamBooth and LoRA, ever yone can manifest their imagination into high-quality images at an affordable co st. However, adding motion dynamics to existing high-quality personalized T2Is a nd enabling them to generate animations remains an open challenge. In this paper , we present AnimateDiff, a practical framework for animating personalized T2I m odels without requiring model-specific tuning. At the core of our framework is a plug-and-play motion module that can be trained once and seamlessly integrated into any personalized T2Is originating from the same base T2I. Through our propo sed training strategy, the motion module effectively learns transferable motion priors from real-world videos. Once trained, the motion module can be inserted i nto a personalized T2I model to form a personalized animation generator. We furt her propose MotionLoRA, a lightweight fine-tuning technique for AnimateDiff that enables a pre-trained motion module to adapt to new motion patterns, such as di fferent shot types, at a low training and data collection cost. We evaluate Anim ateDiff and MotionLoRA on several public representative personalized T2I models collected from the community. The results demonstrate that our approaches help t hese models generate temporally smooth animation clips while preserving the visu al quality and motion diversity. Codes and pre-trained weights are available at
https://github.com/guoyww/AnimateDiff.
****
Xueyi Liu,Li Yi
GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via De noising Diffusion

In this work, we tackle the challenging problem of denoising hand-object interac
tions (HOI). Given an erroneous interaction sequence, the objective is to refine
 the incorrect hand trajectory to remove interaction artifacts for a perceptuall
y realistic sequence. This challenge involves intricate interaction noise, incl
uding unnatural hand poses and incorrect hand-object relations, alongside the ne
cessity for robust generalization to new interactions and diverse noise patterns
. We tackle those challenges through a novel approach, GeneOH Diffusion, incorpo
rating two key designs: an innovative contact-centric HOI representation named G
eneOH and a new domain-generalizable denoising scheme. The contact-centric repre
sentation GeneOH informatively parameterizes the HOI process, facilitating enhan
ced generalization across various HOI scenarios. The new denoising scheme consis
ts of a canonical denoising model trained to project noisy data samples from a w
hitened noise space to a clean data manifold and a ``denoising via diffusion'' s
trategy which can handle input trajectories with various noise patterns by first
 diffusing them to align with the whitened noise space and cleaning via the cano
nical denoiser. Extensive experiments on four benchmarks with significant domain
 variations demonstrate the superior effectiveness of our method. GeneOH Diffusi
on also shows promise for various downstream applications. We include [a website
](https://meowuu7.github.io/GeneOH-Diffusion/) for introducing the work.
****
Zhiyu Zhu,Huaming Chen,Jiayu Zhang,Xinyi Wang,Zhibo Jin,Jason Xue,Flora D. Salim
AttEXplore: Attribution for Explanation with model parameters eXploration
Due to the real-world noise and human-added perturbations, attaining the trustwo
rthiness of deep neural networks (DNNs) is a challenging task. Therefore, it bec
omes essential to offer explanations for the decisions made by these non-linear
and complex parameterized models. Attribution methods are promising for this goa
l, yet its performance can be further improved. In this paper, for the first tim
e, we present that the decision boundary exploration approaches of attribution a
re consistent with the process for transferable adversarial attacks. Specificall
y, the transferable adversarial attacks craft general adversarial samples from t
he source model, which is consistent with the generation of adversarial samples
that can cross multiple decision boundaries in attribution. Utilizing this consi
stency, we introduce a novel attribution method via model parameter exploration.
 Furthermore, inspired by the capability of frequency exploration to investigate
 the model parameters, we provide enhanced explainability for DNNs by manipulati
ng the input features based on frequency information to explore the decision bou
ndaries of different models. Large-scale experiments demonstrate that our \textb
f{A}ttribution method for \textbf{E}xplanation with model parameter e\textbf{X}p
loration (AttEXplore) outperforms other state-of-the-art interpretability method
s. Moreover, by employing other transferable attack techniques, AttEXplore can e
xplore potential variations in attribution outcomes. Our code is available at: h
ttps://github.com/LMBTough/ATTEXPLORE.
****
Xuan Ju,Ailing Zeng,Yuxuan Bian,Shaoteng Liu,Qiang Xu
PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code
Text-guided diffusion models have revolutionized image generation and editing, o
ffering exceptional realism and diversity. Specifically, in the context of diffu
sion-based editing, where a source image is edited according to a target prompt,
 the process commences by acquiring a noisy latent vector corresponding to the s
ource image via the diffusion model. This vector is subsequently fed into separa
te source and target diffusion branches for editing. The accuracy of this invers
ion process significantly impacts the final editing outcome, influencing both es
sential content preservation of the source image and edit fidelity according to
the target prompt.
Prior inversion techniques aimed at finding a unified solution in both the sourc
e and target diffusion branches. However, our theoretical and empirical analyses
 reveal that disentangling these branches leads to a distinct separation of resp
onsibilities for preserving essential content and ensuring edit fidelity. Buildi
ng on this insight, we introduce “PnP Inversion,” a novel technique achieving op
timal performance of both branches with just three lines of code. To assess imag

e editing performance, we present PIE-Bench, an editing benchmark with 700 images showcasing diverse scenes and editing types, accompanied by versatile annotations and comprehensive evaluation metrics. Compared to state-of-the-art optimization-based inversion techniques, our solution not only yields superior performance across 8 editing methods but also achieves nearly an order of speed-up.
****
S. Fatemeh Seyyedsalehi,Mahdieh Soleymani Baghshah,Hamid R. Rabiee
SOInter: A Novel Deep Energy-Based Interpretation Method for Explaining Structured Output Models
This paper proposes a novel interpretation technique to explain the behavior of structured output models, which simultaneously learn mappings between an input vector and a set of output variables. As a result of the complex relationships between the computational path of output variables in structured models, a feature may impact the output value via another feature. We focus on one of the outputs as the target and try to find the most important features adopted by the structured model to decide on the target in each locality of the input space. We consider an arbitrary structured output model available as a black-box and argue that considering correlations among output variables can improve explanation quality. The goal is to train a function as an interpreter for the target output variable over the input space. We introduce an energy-based training process for the interpreter function, which effectively considers the structural information incorporated into the model to be explained. The proposed method's effectiveness is confirmed using various simulated and real data sets.
****
Yang Jin,Kun Xu,Kun Xu,Liwei Chen,Chao Liao,Jianchao Tan,Quzhe Huang,Bin CHEN,Chengru Song,dai meng,Di ZHANG,Wenwu Ou,Kun Gai,Yadong MU
Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization
Recently, the remarkable advance of the Large Language Model (LLM) has inspired researchers to transfer its extraordinary reasoning capability to both vision and language data. However, the prevailing approaches primarily regard the visual input as a prompt and focus exclusively on optimizing the text generation process conditioned upon vision content by a frozen LLM. Such an inequitable treatment of vision and language heavily constrains the model's potential. In this paper, we break through this limitation by representing both vision and language in a unified form. Specifically, we introduce a well-designed visual tokenizer to translate the non-linguistic image into a sequence of discrete tokens like a foreign language that LLM can read. The resulting visual tokens encompass high-level semantics worthy of a word and also support dynamic sequence length varying from the image. Coped with this tokenizer, the presented foundation model called LaVIT can handle both image and text indiscriminately under the same generative learning paradigm. This unification empowers LaVIT to serve as an impressive generalist interface to understand and generate multi-modal content simultaneously. Extensive experiments further showcase that it outperforms the existing models by a large margin on massive vision-language tasks. Our code and models are available at https://github.com/jy0205/LaVIT.
****
Hong Chen,Yipeng Zhang,Simin Wu,Xin Wang,Xuguang Duan,Yuwei Zhou,Wenwu Zhu
DisenBooth: Identity-Preserving Disentangled Tuning for Subject-Driven Text-to-Image Generation
Subject-driven text-to-image generation aims to generate customized images of the given subject based on the text descriptions, which has drawn increasing attention. Existing methods mainly resort to finetuning a pretrained generative model, where the identity-relevant information (e.g., the boy) and the identity-irrelevant information (e.g., the background or the pose of the boy) are entangled in the latent embedding space. However, the highly entangled latent embedding may lead to the failure of subject-driven text-to-image generation as follows: (i) the identity-irrelevant information hidden in the entangled embedding may dominate the generation process, resulting in the generated images heavily dependent on the irrelevant information while ignoring the given text descriptions; (ii) the

identity-relevant information carried in the entangled embedding can not be appropriately preserved, resulting in identity change of the subject in the generated images. To tackle the problems, we propose DisenBooth, an identity-preserving disentangled tuning framework for subject-driven text-to-image generation. Specifically, DisenBooth finetunes the pretrained diffusion model in the denoising process. Different from previous works that utilize an entangled embedding to denoise each image, DisenBooth instead utilizes disentangled embeddings to respectively preserve the subject identity and capture the identity-irrelevant information. We further design the novel weak denoising and contrastive embedding auxiliary tuning objectives to achieve the disentanglement. Extensive experiments show that our proposed DisenBooth framework outperforms baseline models for subject-driven text-to-image generation with the identity-preserved embedding. Additionally, by combining the identity-preserved embedding and identity-irrelevant embedding, DisenBooth demonstrates more generation flexibility and controllability.
****
Zhaoxuan Wu,Mohammad Mohammadi Amiri,Ramesh Raskar,Bryan Kian Hsiang Low
Incentive-Aware Federated Learning with Training-Time Model Rewards
In federated learning (FL), incentivizing contributions of training resources (e.g., data, compute) from potentially competitive clients is crucial. Existing incentive mechanisms often distribute post-training monetary rewards, which suffer from practical challenges of timeliness and feasibility of the rewards. Rewarding the clients after the completion of training may incentivize them to abort the collaboration, and monetizing the contribution is challenging in practice. To address these problems, we propose an incentive-aware algorithm that offers differentiated training-time model rewards for each client at each FL iteration. We theoretically prove that such a $\textit{local}$ design ensures the $\textit{global}$ objective of client incentivization. Through theoretical analyses, we further identify the issue of error propagation in model rewards and thus propose a stochastic reference-model recovery strategy to ensure theoretically that all the clients eventually obtain the optimal model in the limit. We perform extensive experiments to demonstrate the superior incentivizing performance of our method compared to existing baselines.
****
Shuo He,Chaojie Wang,Guowu Yang,Lei Feng
Candidate Label Set Pruning: A Data-centric Perspective for Deep Partial-label Learning
Partial-label learning (PLL) allows each training example to be equipped with a set of candidate labels. Existing deep PLL research focuses on a \emph{learning-centric} perspective to design various training strategies for label disambiguation i.e., identifying the concealed true label from the candidate label set, for model training. However, when the size of the candidate label set becomes excessively large, these learning-centric strategies would be unable to find the true label for model training, thereby causing performance degradation. This motivates us to think from a \emph{data-centric} perspective and pioneer a new PLL-related task called candidate label set pruning (CLSP) that aims to filter out certain potential false candidate labels in a training-free manner. To this end, we propose the first CLSP method based on the inconsistency between the representation space and the candidate label space. Specifically, for each candidate label of a training instance, if it is not a candidate label of the instance's nearest neighbors in the representation space, then it has a high probability of being a false label. Based on this intuition, we employ a per-example pruning scheme that filters out a specific proportion of high-probability false candidate labels. Theoretically, we prove an upper bound of the pruning error rate and analyze how the quality of representations affects our proposed method. Empirically, extensive experiments on both benchmark-simulated and real-world PLL datasets validate the great value of CLSP to significantly improve many state-of-the-art deep PLL methods.
****
Noa Moriel,Matt Ricci,Mor Nitzan
Let's do the time-warp-attend: Learning topological invariants of dynamical syst

ems
Dynamical systems across the sciences, from electrical circuits to ecological ne tworks, undergo qualitative and often catastrophic changes in behavior, called b ifurcations, when their underlying parameters cross a threshold. Existing method s predict oncoming catastrophes in individual systems but are primarily time-ser ies-based and struggle both to categorize qualitative dynamical regimes across d iverse systems and to generalize to real data. To address this challenge, we pro pose a data-driven, physically-informed deep-learning framework for classifying dynamical regimes and characterizing bifurcation boundaries based on the extract ion of topologically invariant features. We focus on the paradigmatic case of th e supercritical Hopf bifurcation, which is used to model periodic dynamics acros s a wide range of applications. Our convolutional attention method is trained wi th data augmentations that encourage the learning of topological invariants whic h can be used to detect bifurcation boundaries in unseen systems and to design m odels of biological systems like oscillatory gene regulatory networks. We furthe r demonstrate our method's use in analyzing real data by recovering distinct pro liferation and differentiation dynamics along pancreatic endocrinogenesis trajec tory in gene expression space based on single-cell data. Our method provides val uable insights into the qualitative, long-term behavior of a wide range of dynam ical systems, and can detect bifurcations or catastrophic transitions in large-s cale physical and biological systems.
****
Xinshuai Dong,Biwei Huang,Ignavier Ng,Xiangchen Song,Yujia Zheng,Songyao Jin,Rob erto Legaspi,Peter Spirtes,Kun Zhang
A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variable s
Most existing causal discovery methods rely on the assumption of no latent confo unders, limiting their applicability in solving real-life problems. In this pape r, we introduce a novel, versatile framework for causal discovery that accommoda tes the presence of causally-related hidden variables almost everywhere in the c ausal network (for instance, they can be effects of measured variables), based o n rank information of covariance matrix over measured variables. We start by inv estigating the efficacy of rank in comparison to conditional independence and, t heoretically, establish necessary and sufficient conditions for the identifiabil ity of certain latent structural patterns. Furthermore, we develop a Rank-based Latent Causal Discovery algorithm, RLCD, that can efficiently locate hidden var iables, determine their cardinalities, and discover the entire causal structure over both measured and hidden ones. We also show that, under certain graphical c onditions, RLCD correctly identifies the Markov Equivalence Class of the whole l atent causal graph asymptotically. Experimental results on both synthetic and re al-world personality data sets demonstrate the efficacy of the proposed approach in finite-sample cases. Our code will be publicly available.
****
Haoxuan Li,Yanghao Xiao,Chunyuan Zheng,Peng Wu,Zhi Geng,Xu Chen,Peng Cui
Debiased Collaborative Filtering with Kernel-based Causal Balancing
Collaborative filtering builds personalized models from the collected user feedb ack. However, the collected data is observational rather than experimental, lead ing to various biases in the data, which can significantly affect the learned mo del. To address this issue, many studies have focused on propensity-based method s to combat the selection bias by reweighting the sample loss, and demonstrate t hat
balancing is important for debiasing both theoretically and empirically. However , there are two questions that still need to be addressed: which function class should be balanced and how to effectively balance that function class? In this p aper, we first perform theoretical analysis to show the effect of balancing fini te-dimensional function classes on the bias of IPS and DR methods, and based on this, we propose a universal kernel-based balancing method to balance functions on the reproducing kernel Hilbert space. In addition, we propose a novel adaptiv e causal balancing method during the alternating update between unbiased evaluat ion and training of the prediction model. Specifically, the prediction loss of t

he model is projected in the kernel-based covariate function space, and the projection coefficients are used to determine which functions should be prioritized for balancing to reduce the estimation bias. We conduct extensive experiments on three real-world datasets to demonstrate the effectiveness of the proposed approach.
****
Vishakh Padmakumar,He He
Does Writing with Language Models Reduce Content Diversity?
Large language models (LLMs) have led to a surge in collaborative writing with model assistance. As different users incorporate suggestions from the same model, there is a risk of decreased diversity in the produced content, potentially limiting diverse perspectives in public discourse. In this work, we measure the impact of co-writing on diversity via a controlled experiment, where users write argumentative essays in three setups---using a base LLM (GPT3), a feedback-tuned LLM (InstructGPT), and writing without model help. We develop a set of diversity metrics and find that writing with InstructGPT (but not the GPT3) results in a statistically significant reduction in diversity. Specifically, it increases the similarity between the writings of different authors and reduces the overall lexical and content diversity. We additionally find that this effect is mainly attributable to InstructGPT contributing less diverse text to co-written essays. In contrast, the user-contributed text remains unaffected by model collaboration. This suggests that the recent improvement in generation quality from adapting models to human feedback might come at the cost of more homogeneous and less diverse content.
****
Matthias Lanzinger,Pablo Barcelo
On the Power of the Weisfeiler-Leman Test for Graph Motif Parameters
Seminal research in the field of graph neural networks (GNNs) has revealed a direct correspondence between the expressive capabilities of GNNs and the k-dimensional
Weisfeiler-Leman (kWL) test, a widely-recognized method for verifying graph isomorphism. This connection has reignited interest in comprehending the specific graph properties effectively distinguishable by the kWL test.
A central focus of research in this field revolves around determining the least dimensionality k, for which kWL can discern graphs with different number of occurrences of a pattern graph p. We refer to such a least k as the WL-dimension of this pattern counting problem. This inquiry traditionally delves into two distinct counting problems related to patterns: subgraph counting and induced subgraph counting. Intriguingly, despite their initial appearance as separate challenges with seemingly divergent approaches, both of these problems are interconnected components of a more comprehensive problem: "graph motif parameters". In this paper, we provide a precise characterization of the WL-dimension of labeled graph motif parameters. As specific instances of this result, we obtain characterizations of the WL-dimension of the subgraph counting and induced subgraph counting problem for every labeled pattern p. Particularly noteworthy is our resolution of a problem left open in previous work
concerning induced copies.
We additionally demonstrate that in cases where the kWL test distinguishes between graphs with varying occurrences of a pattern p, the exact number of occurrences of p can be computed uniformly using only local information of the last layer of a corresponding GNN.
We finally delve into the challenge of recognizing the WL-dimension of various graph parameters. We give a polynomial time algorithm for determining the WL-dimension of the subgraph counting problem for given pattern p, answering an open question from previous work.
We additionally show how to utilize deep results from the field of graph motif parameters, together with our characterization, to determine the WL-dimension of induced subgraph counting and counting k-graphlets.
****
Lichang Chen,Shiyang Li,Jun Yan,Hai Wang,Kalpa Gunaratna,Vikas Yadav,Zheng Tang,

Vijay Srinivasan,Tianyi Zhou,Heng Huang,Hongxia Jin
AlpaGasus: Training a Better Alpaca with Fewer Data
Large language models~(LLMs) strengthen instruction-following capability through instruction-finetuning (IFT) on supervised instruction/response data. However, widely used IFT datasets (e.g., Alpaca's 52k data) surprisingly contain many low-quality instances with incorrect or irrelevant responses, which are misleading and detrimental to IFT. In this paper, we propose a simple and effective data selection strategy that automatically identifies and removes low-quality data using a strong LLM (e.g., ChatGPT). To this end, we introduce Alpagasus, which is finetuned on only 9k high-quality data filtered from the 52k Alpaca data. Alpagasus significantly outperforms the original Alpaca as evaluated by GPT-4 on multiple test sets and the controlled human study. Its 13B variant matches $>90\%$ performance of its teacher LLM (i.e., Text-Davinci-003) on test tasks. It also provides 5.7x faster training, reducing the training time for a 7B variant from 80 minutes (for Alpaca) to 14 minutes \footnote{We apply IFT for the same number of epochs as Alpaca(7B) but on fewer data, using 4$\times$NVIDIA A100 (80GB) GPUs and following the original Alpaca setting and hyperparameters.}. In the experiment, we also demonstrate that our method can work not only for machine-generated datasets but also for human-written datasets. Overall, Alpagasus demonstrates a novel data-centric IFT paradigm that can be generally applied to instruction-tuning data, leading to faster training and better instruction-following models.
****
Wei-Cheng Huang,Chun-Fu Chen,Hsiang Hsu
OVOR: OnePrompt with Virtual Outlier Regularization for Rehearsal-Free Class-Incremental Learning
Recent works have shown that by using large pre-trained models along with learnable prompts, rehearsal-free methods
for class-incremental learning (CIL) settings can achieve superior performance to prominent rehearsal-based ones.
Rehearsal-free CIL methods struggle with distinguishing classes from different tasks, as those are not trained together.
In this work we propose a regularization method based on virtual outliers to tighten decision boundaries of the classifier,
such that confusion of classes among different tasks is mitigated.
Recent prompt-based methods often require a pool of task-specific prompts, in order to prevent overwriting knowledge
of previous tasks with that of the new task, leading to extra computation in querying and composing an
appropriate prompt from the pool.
This additional cost can be eliminated, without sacrificing accuracy, as we reveal in the paper.
We illustrate that a simplified prompt-based method can achieve results comparable to
previous state-of-the-art (SOTA) methods equipped with a prompt pool, using much less learnable parameters and lower inference cost.
Our regularization method has demonstrated its compatibility with different prompt-based methods, boosting
those previous SOTA rehearsal-free CIL methods' accuracy on the ImageNet-R and CIFAR-100 benchmarks. Our source code is available at https://github.com/jpmorganchase/ovor.
****
Ziming Hong,Zhenyi Wang,Li Shen,Yu Yao,Zhuo Huang,Shiming Chen,Chuanwu Yang,Mingming Gong,Tongliang Liu
Improving Non-Transferable Representation Learning by Harnessing Content and Style
Non-transferable learning (NTL) aims to restrict the generalization of models toward the target domain(s). To this end, existing works learn non-transferable representations by reducing statistical dependence between the source and target domain. However, such statistical methods essentially neglect to distinguish between *styles* and *contents*, leading them to inadvertently fit (i) spurious corr

elation between *styles* and *labels*, and (ii) fake independence between *conte nts* and *labels*. Consequently, their performance will be limited when natural distribution shifts occur or malicious intervention is imposed. In this paper, w e propose a novel method (dubbed as H-NTL) to understand and advance the NTL pro blem by introducing a causal model to separately model *content* and *style* as two latent factors, based on which we disentangle and harness them as guidances for learning non-transferable representations with intrinsically causal relation ships. Speci■cally, to avoid fitting spurious correlation and fake independence, we propose a variational inference framework to disentangle the naturally mixed *content factors* and *style factors* under our causal model. Subsequently, bas ed on dual-path knowledge distillation, we harness the disentangled two *factors * as guidances for non-transferable representation learning: (i) we constraint t he source domain representations to fit *content factors* (which are the intrins ic cause of *labels*), and (ii) we enforce that the target domain representation s fit *style factors* which barely can predict labels. As a result, the learned feature representations follow optimal untransferability toward the target domai n and minimal negative influence on the source domain, thus enabling better NTL performance. Empirically, the proposed H-NTL signi■cantly outperforms competing methods by a large margin.
****
Qing Li,Yixin Zhu,Yitao Liang,Ying Nian Wu,Song-Chun Zhu,Siyuan Huang
Neural-Symbolic Recursive Machine for Systematic Generalization
Current learning models often struggle with human-like systematic generalization , particularly in learning compositional rules from limited data and extrapolati ng them to novel combinations. We introduce the Neural-Symbolic Recursive Ma- ch ine (NSR), whose core is a Grounded Symbol System (GSS), allowing for the emer gence of combinatorial syntax and semantics directly from training data. The NSR employs a modular design that integrates neural perception, syntactic parsing, and semantic reasoning. These components are synergistically trained through a n ovel deduction-abduction algorithm. Our findings demonstrate that NSR’s design, imbued with the inductive biases of equivariance and compositionality, grants it the expressiveness to adeptly handle diverse sequence-to-sequence tasks and ach ieve unparalleled systematic generalization. We evaluate NSR’s efficacy across f our challenging benchmarks designed to probe systematic generalization capabilit ies: SCAN for semantic parsing, PCFG for string manipulation, HINT for arithmeti c reasoning, and a compositional machine translation task. The results affirm NS R ’s superiority over contemporary neural and hybrid models in terms of generali zation and transferability.
****
Yang He,Lingao Xiao,Joey Tianyi Zhou,Ivor Tsang
Multisize Dataset Condensation
While dataset condensation effectively enhances training efficiency, its applica tion in on-device scenarios brings unique challenges. 1) Due to the fluctuating computational resources of these devices, there's a demand for a flexible datase t size that diverges from a predefined size. 2) The limited computational power on devices often prevents additional condensation operations. These two challeng es connect to the "subset degradation problem" in traditional dataset condensati on: a subset from a larger condensed dataset is often unrepresentative compared to directly condensing the whole dataset to that smaller size. In this paper, we propose Multisize Dataset Condensation (MDC) by **compressing N condensation processes into a single condensation process to obtain datasets with multiple si zes.** Specifically, we introduce an "adaptive subset loss" on top of the basic condensation loss to mitigate the "subset degradation problem". Our MDC method o ffers several benefits: 1) No additional condensation process is required; 2) re duced storage requirement by reusing condensed images. Experiments validate our findings on networks including ConvNet, ResNet and DenseNet, and datasets includ ing SVHN, CIFAR-10, CIFAR-100 and ImageNet. For example, we achieved 6.40% aver age accuracy gains on condensing CIFAR-10 to ten images per class. Code is avail able at: [https://github.com/he-y/Multisize-Dataset-Condensation](https://github .com/he-y/Multisize-Dataset-Condensation).

****
Yichun Shi,Peng Wang,Jianglong Ye,Long Mai,Kejie Li,Xiao Yang
MVDream: Multi-view Diffusion for 3D Generation
We introduce MVDream, a diffusion model that is able to generate consistent mult
i-view images from a given text prompt. Learning from both 2D and 3D data, a mul
ti-view diffusion model can achieve the generalizability of 2D diffusion models
and the consistency of 3D renderings. We demonstrate that such a multi-view diff
usion model is implicitly a generalizable 3D prior agnostic to 3D representation
s. It can be applied to 3D generation via Score Distillation Sampling, significa
ntly enhancing the consistency and stability of existing 2D-lifting methods. It
can also learn new concepts from a few 2D examples, akin to DreamBooth, but for
3D generation.
****
Kyle Vedder,Neehar Peri,Nathaniel Eliot Chodosh,Ishan Khatri,ERIC EATON,Dinesh J
ayaraman,Yang Liu,Deva Ramanan,James Hays
ZeroFlow: Scalable Scene Flow via Distillation
Scene flow estimation is the task of describing the 3D motion field between temp
orally successive point clouds. State-of-the-art methods use strong priors and t
est-time optimization techniques, but require on the order of tens of seconds to
 process full-size point clouds, making them unusable as computer vision primiti
ves for real-time applications such as open world object detection. Feedforward
methods are considerably faster, running on the order of tens to hundreds of mil
liseconds for full-size point clouds, but require expensive human supervision. T
o address both limitations, we propose _Scene Flow via Distillation_, a simple,
scalable distillation framework that uses a label-free optimization method to pr
oduce pseudo-labels to supervise a feedforward model. Our instantiation of this
framework, _ZeroFlow_, achieves **state-of-the-art** performance on the _Argover
se 2 Self-Supervised Scene Flow Challenge_ while using zero human labels by simp
ly training on large-scale, diverse unlabeled data. At test-time, ZeroFlow is ov
er 1000$\times$ faster than label-free state-of-the-art optimization-based metho
ds on full-size point clouds (34 FPS vs 0.028 FPS) and over 1000$\times$ cheaper
 to train on unlabeled data compared to the cost of human annotation (\\$394 vs
~\\$750,000). To facilitate further research, we will release our code, trained
model weights, and high quality pseudo-labels for the Argoverse 2 and Waymo Open
 datasets.
****
Chi-Min Chan,Weize Chen,Yusheng Su,Jianxuan Yu,Wei Xue,Shanghang Zhang,Jie Fu,Zh
iyuan Liu
ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate
Text evaluation has historically posed significant challenges, often demanding s
ubstantial labor and time cost. With the emergence of large language models (LLM
s), researchers have explored LLMs' potential as alternatives for human evaluati
on. While these single-agent-based approaches show promise, experimental results
 suggest that further advancements are needed to bridge the gap between their cu
rrent effectiveness and human-level evaluation quality.
Recognizing that best practices of human evaluation processes often involve mult
iple human annotators collaborating in the evaluation, we resort to a multi-agen
t debate framework, moving beyond single-agent prompting strategies.
In this paper, we construct a multi-agent referee team called $\textbf{ChatEval}
$ to autonomously discuss and evaluate the quality of different texts.
Our experiments on two benchmarks illustrate that ChatEval delivers superior acc
uracy and correlation in alignment with human assessment. Furthermore, we find t
hat the diverse role prompts (different personas) are essential in the multi-age
nt debate process; that is, utilizing the same role description in the prompts c
an lead to a degradation in performance. Our qualitative analysis also shows tha
t ChatEval transcends mere textual scoring, offering a human-mimicking evaluatio
n process for reliable assessments.
****
Olivier Laurent,Emanuel Aldea,Gianni Franchi
A Symmetry-Aware Exploration of Bayesian Neural Network Posteriors

The distribution of modern deep neural networks (DNNs) weights -- crucial for uncertainty quantification and robustness -- is an eminently complex object due to its extremely high dimensionality. This paper presents one of the first large-scale explorations of the posterior distribution of deep Bayesian Neural Networks (BNNs), expanding its study to real-world vision tasks and architectures. Specifically, we investigate the optimal approach for approximating the posterior, analyze the connection between posterior quality and uncertainty quantification, delve into the impact of modes on the posterior, and explore methods for visualizing the posterior. Moreover, we uncover weight-space symmetries as a critical aspect for understanding the posterior. To this extent, we develop an in-depth assessment of the impact of both permutation and scaling symmetries that tend to obfuscate the Bayesian posterior. While the first type of transformation is known for duplicating modes, we explore the relationship between the latter and L2 regularization, challenging previous misconceptions. Finally, to help the community improve our understanding of the Bayesian posterior, we release the first large-scale checkpoint dataset, including thousands of real-world models, along with our code.

****

Xinyuan Chen,Yaohui Wang,Lingjun Zhang,Shaobin Zhuang,Xin Ma,Jiashuo Yu,Yali Wang,Dahua Lin,Yu Qiao,Ziwei Liu

SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction

Recently video generation has achieved substantial progress with realistic results. Nevertheless, existing AI-generated videos are usually very short clips ("shot-level'') depicting a single scene. To deliver a coherent long video ("story-level''), it is desirable to have creative transition and prediction effects across different clips. This paper presents a short-to-long video diffusion model, SEINE, that focuses on generative transition and prediction. The goal is to generate high-quality long videos with smooth and creative transitions between scenes and varying lengths of shot-level videos. Specifically, we propose a random-mask video diffusion model to automatically generate transitions based on textual descriptions. By providing the images of different scenes as inputs, combined with text-based control, our model generates transition videos that ensure coherence and visual quality. Furthermore, the model can be readily extended to various tasks such as image-to-video animation and autoregressive video prediction. To conduct a comprehensive evaluation of this new generative task, we propose three assessing criteria for smooth and creative transition: temporal consistency, semantic similarity, and video-text semantic alignment. Extensive experiments validate the effectiveness of our approach over existing methods for generative transition and prediction, enabling the creation of story-level long videos.

****

Artem Tsypin,Leonid Anatolievich Ugadiarov,Kuzma Khrabrov,Alexander Telepov,Egor Rumiantsev,Alexey Skrynnik,Aleksandr Panov,Dmitry P. Vetrov,Elena Tutubalina,Artur Kadurin

Gradual Optimization Learning for Conformational Energy Minimization

Molecular conformation optimization is crucial to computer-aided drug discovery and materials design.

Traditional energy minimization techniques rely on iterative optimization methods that use molecular forces calculated by a physical simulator (oracle) as anti-gradients.

However, this is a computationally expensive approach that requires many interactions with a physical simulator.

One way to accelerate this procedure is to replace the physical simulator with a neural network.

Despite recent progress in neural networks for molecular conformation energy prediction, such models are prone to errors due to distribution shift, leading to inaccurate energy minimization.

We find that the quality of energy minimization with neural networks can be improved by providing optimization trajectories as additional training data.

Still, obtaining complete optimization trajectories demands a lot of additional

computations.
To reduce the required additional data, we present the Gradual Optimization Lear ning Framework (GOLF) for energy minimization with neural networks.
The framework consists of an efficient data-collecting scheme and an external op timizer.
The external optimizer utilizes gradients from the energy prediction model to ge nerate optimization trajectories, and the data-collecting scheme selects additio nal training data to be processed by the physical simulator.
Our results demonstrate that the neural network trained with GOLF performs \text it{on par} with the oracle on a benchmark of diverse drug-like molecules using s ignificantly less additional data.
****

Wenlong Chen,Yegor Klochkov,Yang Liu
Post-hoc bias scoring is optimal for fair classification
We consider a binary classification problem under group fairness constraints, wh ich can be one of Demographic Parity (DP), Equalized Opportunity (EOp), or Equal ized Odds (EO). We propose an explicit characterization of Bayes optimal classif ier under the fairness constraints, which turns out to be a simple modification rule of the unconstrained classifier. Namely, we introduce a novel instance-leve l measure of bias, which we call bias score, and the modification rule is a simp le linear rule on top of the finite amount of bias scores. Based on this charact erization, we develop a post-hoc approach that allows us to adapt to fairness co nstraints while maintaining high accuracy. In the case of DP and EOp constraints , the modification rule is thresholding a single bias score, while in the case o f EO constraints we are required to fit a linear modification rule with 2 parame ters. The method can also be applied for composite group-fairness criteria, such as ones involving several sensitive attributes. We achieve competitive or bette r performance compared to both in-processing and post-processing methods across three datasets: Adult, COMPAS, and CelebA. Unlike most post-processing methods, we do not require access to sensitive attributes during the inference time.
****

Linqi Zhou,Aaron Lou,Samar Khanna,Stefano Ermon
Denoising Diffusion Bridge Models
Diffusion models are powerful generative models that map noise to data using sto chastic processes. However, for many applications such as image editing, the mod el input comes from a distribution that is not random noise. As such, diffusion models must rely on cumbersome methods like guidance or projected sampling to in corporate this information in the generative process. In our work, we propose De noising Diffusion Bridge Models (DDBMs), a natural alternative to this paradigm based on *diffusion bridges*, a family of processes that interpolate between two paired distributions given as endpoints. Our method learns the score of the dif fusion bridge from data and maps from one endpoint distribution to the other by solving a (stochastic) differential equation based on the learned score. Our met hod naturally unifies several classes of generative models, such as score-based diffusion models and OT-Flow-Matching, allowing us to adapt existing design and architectural choices to our more general problem. Empirically, we apply DDBMs t o challenging image datasets in both pixel and latent space. On standard image t ranslation problems, DDBMs achieve significant improvement over baseline methods , and, when we reduce the problem to image generation by setting the source dist ribution to random noise, DDBMs achieve comparable FID scores to state-of-the-ar t methods despite being built for a more general task.
****

Adam X. Yang,Maxime Robeyns,Xi Wang,Laurence Aitchison
Bayesian Low-rank Adaptation for Large Language Models
Parameter-efficient fine-tuning (PEFT) has emerged as a new paradigm for cost-ef ficient fine-tuning of large language models (LLMs), with low-rank adaptation (L oRA) being a widely adopted choice. However, fine-tuned LLMs often become overco nfident especially when fine-tuned on small datasets. Bayesian methods, with the ir inherent ability to estimate uncertainty, serve as potent tools to mitigate o verconfidence and enhance calibration. In this work, we introduce Laplace-LoRA,

a straightforward yet effective Bayesian method, which applies the Laplace approximation to the LoRA parameters and, considerably boosts the calibration of fine-tuned LLMs.
****

Chris Cundy,Stefano Ermon
SequenceMatch: Imitation Learning for Autoregressive Sequence Modelling with Backtracking
In many domains, autoregressive models can attain high likelihood on the task of predicting the next observation. However, this maximum-likelihood (MLE) objective does not necessarily match a downstream use-case of autoregressively generating high-quality sequences. The MLE objective weights sequences proportionally to their frequency under the data distribution, with no guidance for the model's behaviour out of distribution (OOD): leading to compounding error during autoregressive generation. In order to address this compounding error problem, we formulate sequence generation as an imitation learning (IL) problem. This allows us to minimize a variety of divergences between the distribution of sequences generated by an autoregressive model and sequences from a dataset, including divergences with weight on OOD generated sequences. The IL framework also allows us to incorporate backtracking by introducing a backspace action into the generation process. This further mitigates the compounding error problem by allowing the model to revert a sampled token if it takes the sequence OOD. Our resulting method, SequenceMatch, can be implemented without adversarial training or major architectural changes. We identify the SequenceMatch-χ2 divergence as a more suitable training objective for autoregressive models which are used for generation. We show that empirically, SequenceMatch training leads to improvements over MLE on text generation with language models and arithmetic
****

Jing Liu,Ruihao Gong,Xiuying Wei,Zhiwei Dong,Jianfei Cai,Bohan Zhuang
QLLM: Accurate and Efficient Low-Bitwidth Quantization for Large Language Models
Large Language Models (LLMs) have demonstrated unparalleled efficacy in natural language processing. However, their high computational demands and memory overheads hinder their broad deployment. To address this, two quantization strategies emerge, including Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ). For LLMs, the billions of parameters make the QAT impractical due to the prohibitive training cost and thus PTQ becomes more prevalent. In existing studies, activation outliers in particular channels are identified as the biggest challenge to PTQ accuracy. They propose to transform the magnitudes from activations to weights, which however offers limited alleviation or suffers from unstable gradients, resulting in a severe performance drop at low-bitwidth. In this paper, we propose QLLM, an accurate and efficient low-bitwidth PTQ method designed for LLMs. QLLM introduces an adaptive channel reassembly technique that reallocates the magnitude of outliers to other channels, thereby mitigating their impact on the quantization range. This is achieved by channel disassembly and channel assembly, which first breaks down the outlier channels into several sub-channels to ensure a more balanced distribution of activation magnitudes. Then similar channels are merged to maintain the original channel number for efficiency. Additionally, an adaptive strategy is designed to autonomously determine the optimal number of sub-channels for channel disassembly. To further compensate for the performance loss caused by quantization, we propose an efficient tuning method that only learns a small number of low-rank weights while freezing the pre-trained quantized model. After training, these low-rank parameters can be fused into the frozen weights without affecting inference. Extensive experiments on LLaMA-1 and LLaMA-2 show that QLLM is able to obtain accurate quantized models efficiently. For example, QLLM quantizes the 4-bit LLaMA-2-70B within 10 hours on a single A100-80G GPU, outperforming the previous state-of-the-art method by 7.89% on the average accuracy across five zero-shot tasks. Code is available at [ZIP Lab](https://github.com/ziplab/QLLM) and [ModelTC](https://github.com/ModelTC/QLLM).
****

Ibrahim Alabdulmohsin,Xiao Wang,Andreas Peter Steiner,Priya Goyal,Alexander D'Amour,Xiaohua Zhai

CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?
We study data-balancing for mitigating biases in contrastive language-image pret raining (CLIP), identifying areas of strength and limitation. First, we reaffirm prior conclusions that CLIP can inadvertently absorb stereotypes. To counter th is, we present a novel algorithm, called Multi-Modal Moment Matching (M4), desig ned to reduce both representation and association biases in multimodal data. We use M4 to conduct an in-depth analysis taking into account various factors, such as the model, representation, and data size. Our study also explores the dynami c nature of how CLIP learns/unlearns biases. In particular, we find that fine-tu ning is effective in countering representation biases, though its impact diminis hes for association biases. Also, data balancing has a mixed impact on quality: it tends to improve classification but can hurt retrieval. Interestingly, data a nd architectural improvements seem to mitigate the negative impact of data balan cing on performance; e.g. applying M4 to SigLIP-B/16 with data quality filters i mproves COCO image-to-text retrieval @5 from 86% (without data balancing) to 87% and ImageNet 0-shot classification from 77% to 77.5%! Finally, we conclude with recommendations for improving the efficacy of data balancing in multimodal syst ems.
****
Michael Samuel Albergo,Nicholas Matthew Boffi,Michael Lindsey,Eric Vanden-Eijnde n
Multimarginal Generative Modeling with Stochastic Interpolants
Given a set of K probability densities, we consider the multimarginal generati ve modeling problem of learning a joint distribution that recovers these densiti es as marginals. The structure of this joint distribution should identify multi-way correspondences among the prescribed marginals. We formalize an approach to this task within a generalization of the stochastic interpolant framework, leadi ng to efficient learning algorithms built upon dynamical transport of measure. O ur generative models are defined by velocity and score fields that can be charac terized as the minimizers of simple quadratic objectives, and they are defined o n a simplex that generalizes the time variable in the usual dynamical transport framework. The resulting transport on the simplex is influenced by all marginals , and we show that multi-way correspondences can be extracted. The identificatio n of such correspondences has applications to style transfer, algorithmic fairne ss, and data decorruption. In addition, the multimarginal perspective enables a n efficient algorithm for optimizing the dynamical transport cost in the ordinar y two-marginal setting. We demonstrate these capacities with several numerical e xamples.
****
Hengjia Li,Yang Liu,Linxuan Xia,Yuqi Lin,Wenxiao Wang,Tu Zheng,Zheng Yang,Xiaohu i Zhong,Xiaobo Ren,Xiaofei He
Few-shot Hybrid Domain Adaptation of Image Generator
Can a pre-trained generator be adapted to the hybrid of multiple target domains and generate images with integrated attributes of them? In this work, we introdu ce a new task -- Few-shot $\textit{Hybrid Domain Adaptation}$ (HDA). Given a sou rce generator and several target domains, HDA aims to acquire an adapted generat or that preserves the integrated attributes of all target domains, without overr iding the source domain's characteristics. Compared with $\textit{Domain Adaptat ion}$ (DA), HDA offers greater flexibility and versatility to adapt generators t o more composite and expansive domains. Simultaneously, HDA also presents more c hallenges than DA as we have access only to images from individual target domain s and lack authentic images from the hybrid domain. To address this issue, we in troduce a discriminator-free framework that directly encodes different domains' images into well-separable subspaces. To achieve HDA, we propose a novel directi onal subspace loss comprised of a distance loss and a direction loss. Concretely , the distance loss blends the attributes of all target domains by reducing the distances from generated images to all target subspaces. The direction loss pres erves the characteristics from the source domain by guiding the adaptation along the perpendicular to subspaces. Experiments show that our method can obtain num erous domain-specific attributes in a single adapted generator, which surpasses

the baseline methods in semantic similarity, image fidelity, and cross-domain co nsistency.
****
Taewon Park,Inchul Choi,Minho Lee
Attention-based Iterative Decomposition for Tensor Product Representation
In recent research, Tensor Product Representation (TPR) is applied for the syste matic generalization task of deep neural networks by learning the compositional structure of data. However, such prior works show limited performance in discove ring and representing the symbolic structure from unseen test data because their decomposition to the structural representations was incomplete. In this work, w e propose an Attention-based Iterative Decomposition (AID) module designed to en hance the decomposition operations for the structured representations encoded fr om the sequential input data with TPR. Our AID can be easily adapted to any TPR-based model and provides enhanced systematic decomposition through a competitive attention mechanism between input features and structured representations. In o ur experiments, AID shows effectiveness by significantly improving the performan ce of TPR-based prior works on the series of systematic generalization tasks. Mo reover, in the quantitative and qualitative evaluations, AID produces more compo sitional and well-bound structural representations than other works.
****
Milad Aghajohari,Juan Agustin Duque,Tim Cooijmans,Aaron Courville
LOQA: Learning with Opponent Q-Learning Awareness
In various real-world scenarios, interactions among agents often resemble the dy namics of general-sum games, where each agent strives to optimize its own utilit y. Despite the ubiquitous relevance of such settings, decentralized machine lear ning algorithms have struggled to find equilibria that maximize individual utili ty while preserving social welfare. In this paper we introduce Learning with Opp onent Q-Learning Awareness (LOQA) , a novel reinforcement learning algorithm tai lored to optimizing an agent's individual utility while fostering cooperation am ong adversaries in partially competitive environments. LOQA assumes that each ag ent samples actions proportionally to their action-value function Q. Experimenta l results demonstrate the effectiveness of LOQA at achieving state-of-the-art pe rformance in benchmark scenarios such as the Iterated Prisoner's Dilemma and the Coin Game. LOQA achieves these outcomes with a significantly reduced computatio nal footprint compared to previous works, making it a promising approach for pra ctical multi-agent applications.
****
Maximilian Fleissner,Leena Chennuru Vankadara,Debarghya Ghoshdastidar
Explaining Kernel Clustering via Decision Trees
Despite the growing popularity of explainable and interpretable machine learning , there is still surprisingly limited work on inherently interpretable clusterin g methods. Recently, there has been a surge of interest in explaining the classi c k-means algorithm, leading to efficient algorithms that approximate k-means cl usters using axis-aligned decision trees. However, interpretable variants of k-m eans have limited applicability in practice, where more flexible clustering meth ods are often needed to obtain useful partitions of the data. In this work, we i nvestigate interpretable kernel clustering, and propose algorithms that construc t decision trees to approximate the partitions induced by kernel k-means, a nonl inear extension of k-means. We further build on previous work on explainable k-m eans and demonstrate how a suitable choice of features allows preserving interpr etability without sacrificing approximation guarantees on the interpretable mode l.
****
Guikun Xu,Yongquan Jiang,PengChuan Lei,Yan Yang,Jim Chen
GTMGC: Using Graph Transformer to Predict Molecule’s Ground-State Conformation
The ground-state conformation of a molecule is often decisive for its properties . However, experimental or computational methods, such as density functional the ory (DFT), are time-consuming and labor-intensive for obtaining this conformatio n. Deep learning (DL) based molecular representation learning (MRL) has made sig nificant advancements in molecular modeling and has achieved remarkable results

in various tasks. Consequently, it has emerged as a promising approach for direc
tly predicting the ground-state conformation of molecules. In this regard, we in
troduce GTMGC, a novel network based on Graph-Transformer (GT) that seamlessly p
redicts the spatial configuration of molecules in a 3D space from their 2D topol
ogical architecture in an end-to-end manner. Moreover, we propose a novel self-a
ttention mechanism called Molecule Structural Residual Self-Attention (MSRSA) fo
r molecular structure modeling. This mechanism not only guarantees high model pe
rformance and easy implementation but also lends itself well to other molecular
modeling tasks. Our method has been evaluated on the Molecule3D benchmark datase
t and the QM9 dataset. Experimental results demonstrate that our approach achiev
es remarkable performance and outperforms current state-of-the-art methods as we
ll as the widely used open-source software RDkit.
****
Robert Huben,Hoagy Cunningham,Logan Riggs Smith,Aidan Ewart,Lee Sharkey
Sparse Autoencoders Find Highly Interpretable Features in Language Models
One of the roadblocks to a better understanding of neural networks' internals is
 \textit{polysemanticity}, where neurons appear to activate in multiple, semanti
cally distinct contexts. Polysemanticity prevents us from identifying concise, h
uman-understandable explanations for what neural networks are doing internally.
One hypothesised cause of polysemanticity is \textit{superposition}, where neura
l networks represent more features than they have neurons by assigning features
to an overcomplete set of directions in activation space, rather than to individ
ual neurons. Here, we attempt to identify those directions, using sparse autoenc
oders to reconstruct the internal activations of a language model. These autoenc
oders learn sets of sparsely activating features that are more interpretable and
 monosemantic than directions identified by alternative approaches, where interp
retability is measured by automated methods. Moreover, we show that with our lea
rned set of features, we can pinpoint the features that are causally responsible
 for counterfactual behaviour on the indirect object identification task \citep{
wang2022interpretability} to a finer degree than previous decompositions. This w
ork indicates that it is possible to resolve superposition in language models us
ing a scalable, unsupervised method. Our method may serve as a foundation for fu
ture mechanistic interpretability work, which we hope will enable greater model
transparency and steerability.
****
Tim Franzmeyer,Stephen Marcus McAleer,Joao F. Henriques,Jakob Nicolaus Foerster,
Philip Torr,Adel Bibi,Christian Schroeder de Witt
Illusory Attacks: Information-theoretic detectability matters in adversarial att
acks
Autonomous agents deployed in the real world need to be robust against adversari
al attacks on sensory inputs.
Robustifying agent policies requires anticipating the strongest attacks possible
.
We demonstrate that existing observation-space attacks on reinforcement learning
 agents have a common weakness: while effective, their lack of information-theor
etic detectability constraints makes them \textit{detectable} using automated me
ans or human inspection.
Detectability is undesirable to adversaries as it may trigger security escalatio
ns.
We introduce \textit{\eattacks{}}, a novel form of adversarial attack on sequent
ial decision-makers that is both effective and of $\epsilon-$bounded statistical
 detectability.
We propose a novel dual ascent algorithm to learn such attacks end-to-end.
Compared to existing attacks, we empirically find \eattacks{} to be significantl
y harder to detect with automated methods, and a small study with human particip
ants\footnote{IRB approval under reference R84123/RE001} suggests they are simil
arly harder to detect for humans.
Our findings suggest the need for better anomaly detectors, as well as effective
 hardware- and system-level defenses. The project website can be found at https:
//tinyurl.com/illusory-attacks.

****
Hoyong Kim,Kangil Kim
Fixed Non-negative Orthogonal Classifier: Inducing Zero-mean Neural Collapse wit
h Feature Dimension Separation
Fixed classifiers in neural networks for classification problems have demonstrat
ed cost efficiency and even outperformed learnable classifiers in some popular b
enchmarks when incorporating orthogonality. Despite these advantages, prior rese
arch has yet to investigate the training dynamics of fixed orthogonal classifier
s on neural collapse, a recently clarified phenomenon that last-layer features c
onverge to a specific form, called simplex ETF, in training classification model
s involving the post-zero-error phase. Ensuring this phenomenon is critical for
obtaining global optimality in a layer-peeled model, potentially leading to enha
nced performance in practice. However, fixed orthogonal classifiers cannot invok
e neural collapse due to their geometric limitations. To overcome the limits, we
 analyze a $\textit{zero-mean neural collapse}$ considering the orthogonality in
 non-negative Euclidean space. Then, we propose a $\textit{fixed non-negative or
thogonal classifier}$ that induces the optimal solution and maximizes the margin
 of an orthogonal layer-peeled model by satisfying the properties of zero-mean n
eural collapse. Building on this foundation, we exploit a $\textit{feature dimen
sion separation}$ effect inherent in our classifier for further purposes: (1) en
hances softmax masking by mitigating feature interference in continual learning
and (2) tackles the limitations of mixup on the hypersphere in imbalanced learni
ng. We conducted comprehensive experiments on various datasets and architectures
 and demonstrated significant performance improvements.
****
Jiayuan Gu,Sean Kirmani,Paul Wohlhart,Yao Lu,Montserrat Gonzalez Arenas,Kanishka
 Rao,Wenhao Yu,Chuyuan Fu,Keerthana Gopalakrishnan,Zhuo Xu,Priya Sundaresan,Peng
 Xu,Hao Su,Karol Hausman,Chelsea Finn,Quan Vuong,Ted Xiao
RT-Trajectory: Robotic Task Generalization via Hindsight Trajectory Sketches
Generalization remains one of the most important desiderata for robust robot lea
rning systems. While recently proposed approaches show promise in generalization
 to novel objects, semantic concepts, or visual distribution shifts, generalizat
ion to new tasks remains challenging. For example, a language-conditioned policy
 trained on pick-and-place tasks will not be able to generalize to a folding tas
k, even if the arm trajectory of folding is similar to pick-and-place. Our key i
nsight is that this kind of generalization becomes feasible if we represent the
task through rough trajectory sketches. We propose a policy conditioning method
using such rough trajectory sketches, which we call RT-Trajectory, that is pract
ical, easy to specify, and allows the policy to effectively perform new tasks th
at would otherwise be challenging to perform. We find that trajectory sketches s
trike a balance between being detailed enough to express low-level motion-centri
c guidance while being coarse enough to allow the learned policy to interpret th
e trajectory sketch in the context of situational visual observations. In additi
on, we show how trajectory sketches can provide a useful interface to communicat
e with robotic policies -- they can be specified through simple human inputs lik
e drawings or videos, or through automated methods such as modern image-generati
ng or waypoint-generating methods. We evaluate RT-Trajectory at scale on a varie
ty of real-world robotic tasks, and find that RT-Trajectory is able to perform a
 wider range of tasks compared to language-conditioned and goal-conditioned poli
cies, when provided the same training data.
****
Maryam Toloubidokhti,Yubo Ye,Ryan Missel,Xiajun Jiang,Nilesh Kumar,Ruby Shrestha
,Linwei Wang
DATS: Difficulty-Aware Task Sampler for Meta-Learning Physics-Informed Neural Ne
tworks
Advancements in deep learning have led to the development of physics-informed ne
ural networks (PINNs) for solving partial differential equations (PDEs) without
being supervised by PDE solutions. While vanilla PINNs require training one netw
ork per PDE configuration, recent works have showed the potential to meta-learn
PINNs across a range of PDE configurations. It is however known that PINN traini

ng is associated with different levels of difficulty, depending on the underlyin
g PDE configurations or the number of residual sampling points available. Existi
ng meta-learning approaches, however, treat all PINN tasks equally. We address t
his gap by introducing a novel difficulty-aware task sampler (DATS) for meta-lea
rning of PINNs. We derive an optimal analytical solution to optimize the probabi
lity for sampling individual PINN tasks in order to minimize their validation lo
ss across tasks. We further present two alternative strategies to utilize this s
ampling probability to either adaptively weigh PINN tasks, or dynamically alloca
te optimal residual points across tasks. We evaluated DATS against uniform and s
elf-paced task-sampling baselines on two representative meta-PINN models, across
 four benchmark PDEs as well as three different residual point sampling strategi
es. The results demonstrated that DATS was able to improve the accuracy of meta-
learned PINN solutions when reducing performance disparity across PDE configurat
ions, at only a fraction of residual sampling budgets required by its baselines.
****
Ted Zadouri,Ahmet Üstün,Arash Ahmadian,Beyza Ermis,Acyr Locatelli,Sara Hooker
Pushing Mixture of Experts to the Limit: Extremely Parameter Efficient MoE for I
nstruction Tuning
The Mixture of Experts (MoE) is a widely known neural architecture where an ens
emble of specialized sub-models optimizes overall performance with a constant co
mputational cost. However, conventional MoEs pose challenges at scale due to the
 need to store all experts in memory. In this paper, we push MoE to the limit. W
e propose extremely parameter-efficient MoE by uniquely combining MoE architectu
re with lightweight experts.Our MoE architecture outperforms standard parameter-
efficient fine-tuning (PEFT) methods and is on par with full fine-tuning by only
 updating the lightweight experts -- less than 1\% of an 11B parameters model. F
urthermore, our method generalizes to unseen tasks as it does not depend on any
prior task knowledge. Our research underscores the versatility of the mixture of
 experts architecture, showcasing its ability to deliver robust performance even
 when subjected to rigorous parameter constraints.
****
Debangshu Banerjee,Avaljot Singh,Gagandeep Singh
Dissecting Neural Network Robustness Proofs
In recent years numerous methods have been developed to formally verify the robu
stness of deep neural networks (DNNs).
Though the proposed techniques are effective in providing mathematical guarantee
s about the DNNs' behavior, it is not clear whether the proofs generated by thes
e methods are human understandable.
In this paper, we bridge this gap by developing new concepts, algorithms, and re
presentations to generate human understandable insights into the internal workin
gs of DNN robustness proofs.
Leveraging the proposed method, we show that the robustness proofs of standard D
NNs rely more on spurious input features as compared to the proofs of DNNs train
ed to be robust.
Robustness proofs of the provably robust DNNs filter out a larger number of spur
ious input features as compared to adversarially trained DNNs, sometimes even le
ading to the pruning of semantically meaningful input features.
The proofs for the DNNs combining adversarial and provably robust training tend
to achieve the middle ground.
****
Neehal Tumma,Mathias Lechner,Noel Loo,Ramin Hasani,Daniela Rus
Leveraging Low-Rank and Sparse Recurrent Connectivity for Robust Closed-Loop Con
trol
Developing autonomous agents that can interact with changing environments is an
open challenge in machine learning. Robustness is particularly important in thes
e settings as agents are often fit offline on expert demonstrations but deployed
 online where they must generalize to the closed feedback loop within the enviro
nment. In this work, we explore the application of recurrent neural networks to
tasks of this nature and understand how a parameterization of their recurrent co
nnectivity influences robustness in closed-loop settings. Specifically, we repre

sent the recurrent connectivity as a function of rank and sparsity and show both theoretically and empirically that modulating these two variables has desirable effects on network dynamics. The proposed low-rank, sparse connectivity induces an interpretable prior on the network that proves to be most amenable for a class of models known as closed-form continuous-time neural networks (CfCs). We find that CfCs with fewer parameters can outperform their full-rank, fully-connected counterparts in the online setting under distribution shift. This yields memory-efficient and robust agents while opening a new perspective on how we can modulate network dynamics through connectivity.
****
Yi Sui,Tongzi Wu,Jesse C. Cresswell,Ga Wu,George Stein,Xiao Shi Huang,Xiaochen Zhang,Maksims Volkovs
Self-supervised Representation Learning from Random Data Projectors
Self-supervised representation learning (SSRL) has advanced considerably by exploiting the transformation invariance assumption under artificially designed data augmentations. While augmentation-based SSRL algorithms push the boundaries of performance in computer vision and natural language processing, they are often not directly applicable to other data modalities, and can conflict with application-specific data augmentation constraints. This paper presents an SSRL approach that can be applied to any data modality and network architecture because it does not rely on augmentations or masking. Specifically, we show that high-quality data representations can be learned by reconstructing random data projections. We evaluate the proposed approach on a wide range of representation learning tasks that span diverse modalities and real-world applications. We show that it outperforms multiple state-of-the-art SSRL baselines.
Due to its wide applicability and strong empirical results, we argue that learning from randomness is a fruitful research direction worthy of attention and further study.
****
Edward S. Hu,James Springer,Oleh Rybkin,Dinesh Jayaraman
Privileged Sensing Scaffolds Reinforcement Learning
We need to look at our shoelaces as we first learn to tie them but having mastered this skill, can do it from touch alone. We call this phenomenon “sensory scaffolding”: observation streams that are not needed by a master might yet aid a novice learner. We consider such sensory scaffolding setups for training artificial agents. For example, a robot arm may need to be deployed with just a low-cost, robust, general-purpose camera; yet its performance may improve by having privileged training-time-only access to informative albeit expensive and unwieldy motion capture rigs or fragile tactile sensors. For these settings, we propose “Scaffolder”, a reinforcement learning approach which effectively exploits privileged sensing in critics, world models, reward estimators, and other such auxiliary components that are only used at training time, to improve the target policy. For evaluating sensory scaffolding agents, we design a new “S3” suite of ten diverse simulated robotic tasks that explore a wide range of practical sensor setups. Agents must use privileged camera sensing to train blind hurdlers, privileged active visual perception to help robot arms overcome visual occlusions, privileged touch sensors to train robot hands, and more. Scaffolder easily outperforms relevant prior baselines and frequently performs comparably even to policies that have test-time access to the privileged sensors. Website: https://penn-pal-lab.github.io/scaffolder/
****
Zhibin Gou,Zhihong Shao,Yeyun Gong,yelong shen,Yujiu Yang,Minlie Huang,Nan Duan,Weizhu Chen
ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving
Large language models have made significant progress in various language tasks, yet they still struggle with complex mathematics. In this paper, we propose ToRA a series of Tool-integrated Reasoning Agents designed to solve challenging mathematical problems by seamlessly integrating natural language reasoning with the utilization of external tools (e.g., computation libraries and symbolic solvers), thereby amalgamating the analytical prowess of language and the computational

efficiency of tools. To train ToRA, we curate interactive tool-use trajectories
on mathematical datasets, apply imitation learning on the annotations, and propo
se output space shaping to further refine models' reasoning behavior. As a resul
t, ToRA models significantly outperform open-source models on 10 mathematical re
asoning datasets across all scales with 13%-19% absolute improvements on average
. Notably, ToRA-7B reaches 44.6% on the competition-level dataset MATH, surpassi
ng the best open-source model WizardMath-70B by 22% absolute. ToRA-34B is also t
he first open-source model that achieves an accuracy exceeding 50% on MATH, whic
h significantly outperforms GPT-4's CoT result, and is competitive with GPT-4 so
lving problems with programs. Additionally, we conduct a comprehensive analysis
of the benefits and remaining challenges of tool interaction for mathematical re
asoning, providing valuable insights for future research.
****
Eric Mitchell,Rafael Rafailov,Archit Sharma,Chelsea Finn,Christopher D Manning
An Emulator for Fine-tuning Large Language Models using Small Language Models
Widely used language models (LMs) are typically built by scaling up a two-stage
training pipeline: a pre-training stage that uses a very large, diverse dataset
of text and a fine-tuning (sometimes, 'alignment') stage using more targeted exa
mples of specific behaviors and/or human preferences. While it has been hypothes
ized that knowledge and skills come from pre-training, and fine-tuning mostly fi
lters this knowledge and skillset, this intuition has not been rigorously tested
. In this paper, we test this hypothesis with a novel methodology for scaling th
ese two stages independently, essentially asking, *What would happen if we combi
ned the knowledge learned by a large model during pre-training with the knowledg
e learned by a small model during fine-tuning (or vice versa)?* Using an RL-base
d framework derived from recent developments in learning from human preferences,
 we introduce *emulated fine-tuning (EFT)*, a principled and practical method fo
r sampling from a distribution that approximates the result of pre-training and
fine-tuning at different scales. Our experiments with EFT show that scaling up f
ine-tuning tends to improve helpfulness, while scaling up pre-training tends to
improve factuality. Further, we show that EFT enables test-time adjustment of co
mpeting behavioral factors like helpfulness and harmlessness without additional
training. Finally, we find that a special case of emulated fine-tuning, which we
 call LM *up-scaling*, avoids resource-intensive fine-tuning of large pre-traine
d models by ensembling small fine-tuned models with large pre-trained models, es
sentially 'emulating' the result of fine-tuning the large pre-trained model. Up-
scaling consistently improves helpfulness and factuality of widely used pre-trai
ned models like Llama, Llama-2, and Falcon, without additional hyperparameters o
r training.
****
Hongxin Zhang,Weihua Du,Jiaming Shan,Qinhong Zhou,Yilun Du,Joshua B. Tenenbaum,T
ianmin Shu,Chuang Gan
Building Cooperative Embodied Agents Modularly with Large Language Models
In this work, we address challenging multi-agent cooperation problems with decen
tralized control, raw sensory observations, costly communication, and multi-obje
ctive tasks instantiated in various embodied environments. While previous resear
ch either presupposes a cost-free communication channel or relies on a centraliz
ed controller with shared observations, we harness the commonsense knowledge, re
asoning ability, language comprehension, and text generation prowess of LLMs and
 seamlessly incorporate them into a cognitive-inspired modular framework that in
tegrates with perception, memory, and execution. Thus building a Cooperative Emb
odied Language Agent CoELA, who can plan, communicate, and cooperate with others
 to accomplish long-horizon tasks efficiently. Our experiments on C-WAH and TDW-
MAT demonstrate that CoELA driven by GPT-4 can surpass strong planning-based met
hods and exhibit emergent effective communication. Though current Open LMs like
LLAMA-2 still underperform, we fine-tune a CoELA with data collected with our ag
ents and show how they can achieve promising performance. We also conducted a us
er study for human-agent interaction and discovered that CoELA communicating in
natural language can earn more trust and cooperate more effectively with humans.
 Our research underscores the potential of LLMs for future research in multi-age

nt cooperation. Videos can be found on the project website https://vis-www.cs.um
ass.edu/Co-LLM-Agents/.
****
Niels Mündler,Jingxuan He,Slobodan Jenko,Martin Vechev
Self-contradictory Hallucinations of Large Language Models: Evaluation, Detectio
n and Mitigation
Large language models (large LMs) are susceptible to producing text that contain
s hallucinated content. An important instance of this problem is self-contradict
ion, where the LM generates two contradictory sentences within the same context.
 In this work, we present a comprehensive investigation into self-contradiction
for various instruction-tuned LMs, covering evaluation, detection, and mitigatio
n. Our primary evaluation task is open-domain text generation, but we also demon
strate the applicability of our approach to shorter question answering. Our anal
ysis reveals the prevalence of self-contradictions, e.g., in 17.7% of all senten
ces produced by ChatGPT. We then propose a novel prompting-based framework desig
ned to effectively detect and mitigate self-contradictions. Our detector achieve
s high accuracy, e.g., around 80% F1 score when prompting ChatGPT. The mitigatio
n algorithm iteratively refines the generated text to remove contradictory infor
mation while preserving text fluency and informativeness. Importantly, our entir
e framework is applicable to black-box LMs and does not require retrieval of ext
ernal knowledge. Rather, our method complements retrieval-based methods, as a la
rge portion of self-contradictions (e.g., 35.2% for ChatGPT) cannot be verified
using online text. Our approach is practically effective and has been released a
s a push-button tool to benefit the public at https://chatprotect.ai/.
****
Quoc Phong Nguyen,Bryan Kian Hsiang Low,Patrick Jaillet
Leveraging Previous Tasks in Optimizing Risk Measures with Gaussian Processes
Research on optimizing the risk measure of a blackbox function using Gaussian pr
ocesses, especially Bayesian optimization (BO) of risk measures, has become incr
easingly important due to the inevitable presence of uncontrollable variables in
 real-world applications. Nevertheless, existing works on BO of risk measures st
art the optimization from scratch for every new task without considering the res
ults of previous tasks. In contrast, its vanilla BO counterpart has received a t
horough investigation on utilizing previous tasks to speed up the current task t
hrough the body of works on meta-BO which, however, have not considered risk mea
sures. To bridge this gap, this paper presents the first algorithm for meta-BO o
f risk measures (i.e., value-at-risk (VaR) and the conditional VaR) by introduci
ng a novel adjustment to the upper confidence bound acquisition function. Our pr
oposed algorithm exhibits two desirable properties: (i) invariance to scaling an
d vertical shifting of the blackbox function and (ii) robustness to previous har
mful tasks. We provide a theoretical performance guarantee for our algorithm and
 empirically demonstrate its performance using several synthetic function benchm
arks and real-world objective functions.
****
Samuel Holt,Max Ruiz Luyten,Mihaela van der Schaar
L2MAC: Large Language Model Automatic Computer for Extensive Code Generation
Transformer-based large language models (LLMs) are constrained by the fixed cont
ext window of the underlying transformer architecture, hindering their ability t
o produce long and coherent outputs. Memory-augmented LLMs are a promising solut
ion, but current approaches cannot handle long output generation tasks since the
y (1) only focus on reading memory and reduce its evolution to the concatenation
 of new memories or (2) use very specialized memories that cannot adapt to other
 domains. This paper presents L2MAC, the first practical LLM-based stored-progra
m automatic computer (von Neumann architecture) framework, an LLM-based multi-ag
ent system, for long and consistent output generation. Its memory has two compon
ents: the instruction registry, which is populated with a prompt program to solv
e the user-given task, and a file store, which will contain the final and interm
ediate outputs. Each instruction in turn is executed by a separate LLM agent, wh
ose context is managed by a control unit capable of precise memory reading and w
riting to ensure effective interaction with the file store. These components ena

ble L2MAC to generate extensive outputs, bypassing the constraints of the finite
 context window while producing outputs that fulfill a complex user-specified ta
sk. We empirically demonstrate that L2MAC achieves state-of-the-art performance
in generating large codebases for system design tasks, significantly outperformi
ng other coding methods in implementing the detailed user-specified task, and we
 provide valuable insights into the reasons for this performance gap.
****
Christian Koke,Daniel Cremers
HoloNets: Spectral Convolutions do extend to Directed Graphs
Within the graph learning community, conventional wisdom dictates that spectral
convolutional networks may only be deployed on undirected graphs: Only there cou
ld the existence of a well-defined graph Fourier transform be guaranteed, so tha
t information may be translated between spatial- and spectral domains. Here we s
how this traditional reliance on the graph Fourier transform to be superfluous a
nd -- making use of certain advanced tools from complex analysis and spectral t
heory -- extend spectral convolutions to directed graphs.
We provide a frequency-response interpretation of newly developed filters, inves
tigate the influence of the basis used to express filters and discuss the interp
lay with characteristic operators on which networks are based. In order to thoro
ughly test the developed theory, we conduct experiments in real world settings,
 showcasing that directed spectral convolutional networks provide new state of
the art results for heterophilic node classification on many datasets and -- a
s opposed to baselines -- may be rendered stable to resolution-scale varying top
ological perturbations.
****
Insu Han,Rajesh Jayaram,Amin Karbasi,Vahab Mirrokni,David Woodruff,Amir Zandieh
HyperAttention: Long-context Attention in Near-Linear Time
We present an approximate attention mechanism named `HyperAttention` to address
the computational challenges posed by the growing complexity of long contexts us
ed in Large Language Models (LLMs).
Recent work suggests that in the worst-case scenario, the quadratic time is nece
ssary unless the entries of the attention matrix are bounded or the matrix has l
ow stable rank.
We introduce two parameters which measure: (1) the max column norm in the normal
ized attention matrix, and (2) the ratio of row norms in the unnormalized attent
ion matrix after detecting and removing large entries. We use these fine-grained
 parameters to capture the hardness of the problem.
Despite previous lower bounds, we are able to achieve a linear time sampling alg
orithm even when the matrix has unbounded entries or a large stable rank, provid
ed the above parameters are small.
HyperAttention features a modular design that easily accommodates integration of
 other fast low-level implementations, particularly FlashAttention.
Empirically, employing Locality Sensitive Hashing (LSH) to identify large entrie
s, HyperAttention outperforms existing methods, giving significant speed improve
ments compared to state-of-the-art solutions like FlashAttention.
This development presents substantial implications for enabling LLMs to handle s
ignificantly larger contexts.
****
Wei Huang,Ye Shi,Zhongyi Cai,Taiji Suzuki
Understanding Convergence and Generalization in Federated Learning through Featu
re Learning Theory
Federated Learning (FL) has attracted significant attention as an efficient priv
acy-preserving approach to distributed learning across multiple clients. Despite
 extensive empirical research and practical applications, a systematic way to th
eoretically understand the convergence and generalization properties in FL remai
ns limited. This work aims to establish a unified theoretical foundation for und
erstanding FL through feature learning theory. We focus on a scenario where each
 client employs a two-layer convolutional neural network (CNN) for local trainin
g on their own data. Many existing works analyze the convergence of Federated Av
eraging (FedAvg) under lazy training with linearizing assumptions in weight spac

e. In contrast, our approach tracks the trajectory of signal learning and noise memorization in FL, eliminating the need for these assumptions. We further show that FedAvg can achieve near-zero test error by effectively increasing signal-to-noise ratio (SNR) in feature learning, while local training without communication achieves a large constant test error. This finding highlights the benefits of communication for generalization in FL. Moreover, our theoretical results suggest that a weighted FedAvg method, based on the similarity of input features across clients, can effectively tackle data heterogeneity issues in FL. Experimental results on both synthetic and real-world datasets verify our theoretical conclusions and emphasize the effectiveness of the weighted FedAvg approach.
****
Aleksandar Makelov,Georg Lange,Atticus Geiger,Neel Nanda
Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching
Mechanistic interpretability aims to attribute high-level model behaviors to specific, interpretable learned features. It is hypothesized that these features manifest as directions or low-dimensional subspaces within activation space. Accordingly, recent studies have explored the identification and manipulation of such subspaces to reverse-engineer computations, employing methods such as activation patching. In this work, we demonstrate that naïve approaches to subspace interventions can give rise to interpretability illusions.

Specifically, even if patching along a subspace has the intended end-to-end causal effect on model behavior, this effect may be achieved by activating \emph{a dormant parallel pathway} using a component that is \textit{causally disconnected} from the model output.
We demonstrate this in a mathematical example, realize the example empirically in two different settings (the Indirect Object Identification (IOI) task and factual recall), and argue that activating dormant pathways ought to be prevalent in practice.
In the context of factual recall, we further show that the illusion is related to rank-1 fact editing, providing a mechanistic explanation for previous work observing an inconsistency between fact editing performance and fact localisation.

However, this does not imply that activation patching of subspaces is intrinsically unfit for interpretability.
To contextualize our findings, we also show what a success case looks like in a task (IOI) where prior manual circuit analysis allows an understanding of the location of the ground truth feature. We explore the additional evidence needed to argue that a patched subspace is faithful.
****
Jason Y. Zhang,Amy Lin,Moneish Kumar,Tzu-Hsuan Yang,Deva Ramanan,Shubham Tulsiani
Cameras as Rays: Pose Estimation via Ray Diffusion
Estimating camera poses is a fundamental task for 3D reconstruction and remains challenging given sparsely sampled views (<10). In contrast to existing approaches that pursue top-down prediction of global parametrizations of camera extrinsics, we propose a distributed representation of camera pose that treats a camera as a bundle of rays. This representation allows for a tight coupling with spatial image features improving pose precision. We observe that this representation is naturally suited for set-level transformers and develop a regression-based approach that maps image patches to corresponding rays. To capture the inherent uncertainties in sparse-view pose inference, we adapt this approach to learn a denoising diffusion model which allows us to sample plausible modes while improving performance. Our proposed methods, both regression- and diffusion-based, demonstrate state-of-the-art performance on camera pose estimation on CO3D while generalizing to unseen object categories and in-the-wild captures.
****
Simon Ging,Maria Alejandra Bravo,Thomas Brox
Open-ended VQA benchmarking of Vision-Language models by exploiting Classificati

on datasets and their semantic hierarchy
The evaluation of text-generative vision-language models is a challenging yet cr
ucial endeavor. By addressing the limitations of existing Visual Question Answer
ing (VQA) benchmarks and proposing innovative evaluation methodologies, our rese
arch seeks to advance our understanding of these models' capabilities. We propos
e a novel VQA benchmark based on well-known visual classification datasets which
 allows a granular evaluation of text-generative vision-language models and thei
r comparison with discriminative vision-language models. To improve the assessme
nt of coarse answers on fine-grained classification tasks, we suggest using the
semantic hierarchy of the label space to ask automatically generated follow-up q
uestions about the ground-truth category. Finally, we compare traditional NLP an
d LLM-based metrics for the problem of evaluating model predictions given ground
-truth answers. We perform a human evaluation study upon which we base our decis
ion on the final metric. We apply our benchmark to a suite of vision-language mo
dels and show a detailed comparison of their abilities on object, action, and at
tribute classification. Our contributions aim to lay the foundation for more pre
cise and meaningful assessments, facilitating targeted progress in the exciting
field of vision-language modeling.
****
Kai Lagemann,Christian Lagemann,Sach Mukherjee
Invariance-based Learning of Latent Dynamics
We propose a new model class aimed at predicting dynamical trajectories from hig
h-dimensional empirical data. This is done by combining variational autoencoders
 and (spatio-)temporal transformers within a framework designed to enforce cert
ain scientifically-motivated invariances. The models allow inference of system b
ehavior at any continuous time and generalization well beyond the data distribut
ions seen during training. Furthermore, the models do not require an explicit ne
ural ODE formulation, making them efficient and highly scalable in practice. We
study behavior through simple theoretical analyses and extensive empirical exp
eriments. The latter investigate the ability to predict the trajectories of com
plicated systems based on finite data and show that the proposed approaches can
 outperform existing neural-dynamical models. We study also more general inducti
ve bias in the context of transfer to data obtained under entirely novel system
interventions. Overall, our results provide a new framework for efficiently lear
ning complicated dynamics in a data-driven manner, with potential applications i
n a wide range of fields including physics, biology, and engineering.
****
Yuto Nishimura,Taiji Suzuki
Minimax optimality of convolutional neural networks for infinite dimensional inp
ut-output problems and separation from kernel methods
Recent deep learning applications, exemplified by text-to-image tasks, often inv
olve high-dimensional inputs and outputs. While several studies have investigate
d the function estimation capabilities of deep learning, research on dilated con
volutional neural networks (CNNs) has mainly focused on cases where input dimens
ions are infinite but output dimensions are one-dimensional, similar to many oth
er studies. However, many practical deep learning tasks involve high-dimensional
 (or even infinite dimensional) inputs and outputs.
In this paper, we investigate the optimality of dilated CNNs for estimating a ma
p between infinite-dimensional input and output spaces
by analyzing their approximation and estimation abilities.
For that purpose, we first show that approximation and estimation errors depend
only on the smoothness and decay rate with respect to the infinity norm of the o
utput, and their estimation accuracy actually achieve the {\it minimax optimal}
rate of convergence.
Second, we demonstrate that the dilated CNNs outperform {\it any} linear estimat
ors including kernel ridge regression and k-NN estimators in a minimax error s
ense, highlighting the usefulness of feature learning realized by deep neural ne
tworks.
Our theoretical analysis particularly explains the success of deep learning in r
ecent high-dimensional input-output tasks.

****
Sichao Li,Rong Wang,Quanling Deng,Amanda S Barnard
Exploring the cloud of feature interaction scores in a Rashomon set
Interactions among features are central to understanding the behavior of machine learning models. Recent research has made significant strides in detecting and quantifying feature interactions in single predictive models. However, we argue that the feature interactions extracted from a single pre-specified model may not be trustworthy since: *a well-trained predictive model may not preserve the true feature interactions and there exist multiple well-performing predictive models that differ in feature interaction strengths*. Thus, we recommend exploring feature interaction strengths in a model class of approximately equally accurate predictive models. In this work, we introduce the feature interaction score (FIS) in the context of a Rashomon set, representing a collection of models that achieve similar accuracy on a given task. We propose a general and practical algorithm to calculate the FIS in the model class. We demonstrate the properties of the FIS via synthetic data and draw connections to other areas of statistics. Additionally, we introduce a Halo plot for visualizing the feature interaction variance in high-dimensional space and a swarm plot for analyzing FIS in a Rashomon set. Experiments with recidivism prediction and image classification illustrate how feature interactions can vary dramatically in importance for similarly accurate predictive models. Our results suggest that the proposed FIS can provide valuable insights into the nature of feature interactions in machine learning models.
****
Yumeng Li,Margret Keuper,Dan Zhang,Anna Khoreva
Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive
Despite the recent advances in large-scale diffusion models, little progress has been made on the layout-to-image (L2I) synthesis task. Current L2I models either suffer from poor editability via text or weak alignment between the generated image and the input layout. This limits their usability in practice. To mitigate this, we propose to integrate adversarial supervision into the conventional training pipeline of L2I diffusion models (ALDM). Specifically, we employ a segmentation-based discriminator which provides explicit feedback to the diffusion generator on the pixel-level alignment between the denoised image and the input layout. To encourage consistent adherence to the input layout over the sampling steps, we further introduce the multistep unrolling strategy. Instead of looking at a single timestep, we unroll a few steps recursively to imitate the inference process, and ask the discriminator to assess the alignment of denoised images with the layout over a certain time window. Our experiments show that ALDM enables layout faithfulness of the generated images, while allowing broad editability via text prompts. Moreover, we showcase its usefulness for practical applications: by synthesizing target distribution samples via text control, we improve domain generalization of semantic segmentation models by a large margin (~12 mIoU points).
****
Xu Zheng,Tianchun Wang,Wei Cheng,Aitian Ma,Haifeng Chen,Mo Sha,Dongsheng Luo
Parametric Augmentation for Time Series Contrastive Learning
Modern techniques like contrastive learning have been effectively used in many areas, including computer vision, natural language processing, and graph-structured data. Creating positive examples that assist the model in learning robust and discriminative representations is a crucial stage in contrastive learning approaches. Usually, preset human intuition directs the selection of relevant data augmentations. Due to patterns that are easily recognized by humans, this rule of thumb works well in the vision and language domains. However, it is impractical to visually inspect the temporal structures in time series. The diversity of time series augmentations at both the dataset and instance levels makes it difficult to choose meaningful augmentations on the fly. Thus, although prevalent, contrastive learning with data augmentation has been less studied in the time series domain. In this study, we address this gap by analyzing time series data augmentation using information theory and summarizing the most commonly adopted augment

ations in a unified format. We then propose a parametric augmentation method, AutoTCL, which can be adaptively employed to support time series representation learning. The proposed approach is encoder-agnostic, allowing it to be seamlessly integrated with different backbone encoders. Experiments on univariate forecasting tasks demonstrate the highly competitive results of our method, with an average 6.5\% reduction in MSE and 4.7\% in MAE over the leading baselines. In classification tasks, AutoTCL achieves a 1.2% increase in average accuracy.
****
Marko Mihajlovic,Sergey Prokudin,Marc Pollefeys,Siyu Tang
ResFields: Residual Neural Fields for Spatiotemporal Signals
Neural fields, a category of neural networks trained to represent high-frequency signals, have gained significant attention in recent years due to their impressive performance in modeling complex 3D data, such as signed distance (SDFs) or radiance fields (NeRFs), via a single multi-layer perceptron (MLP). However, despite the power and simplicity of representing signals with an MLP, these methods still face challenges when modeling large and complex temporal signals due to the limited capacity of MLPs. In this paper, we propose an effective approach to address this limitation by incorporating temporal residual layers into neural fields, dubbed ResFields. It is a novel class of networks specifically designed to effectively represent complex temporal signals. We conduct a comprehensive analysis of the properties of ResFields and propose a matrix factorization technique to reduce the number of trainable parameters and enhance generalization capabilities. Importantly, our formulation seamlessly integrates with existing MLP-based neural fields and consistently improves results across various challenging tasks: 2D video approximation, dynamic shape modeling via temporal SDFs, and dynamic NeRF reconstruction. Lastly, we demonstrate the practical utility of ResFields by showcasing its effectiveness in capturing dynamic 3D scenes from sparse RGBD cameras of a lightweight capture system.
****
Weize Chen,Yusheng Su,Jingwei Zuo,Cheng Yang,Chenfei Yuan,Chi-Min Chan,Heyang Yu,Yaxi Lu,Yi-Hsin Hung,Chen Qian,Yujia Qin,Xin Cong,Ruobing Xie,Zhiyuan Liu,Maosong Sun,Jie Zhou
AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors
Autonomous agents empowered by Large Language Models (LLMs) have undergone significant improvements, enabling them to generalize across a broad spectrum of tasks. However, in real-world scenarios, cooperation among individuals is often required to enhance the efficiency and effectiveness of task accomplishment. Hence, inspired by human group dynamics, we propose a multi-agent framework AgentVerse that can effectively orchestrate a collaborative group of expert agents as a greater-than-the-sum-of-its-parts system. Our experiments demonstrate that AgentVerse can proficiently deploy multi-agent groups that outperform a single agent. Extensive experiments on text understanding, reasoning, coding, tool utilization, and embodied AI confirm the effectiveness of AgentVerse. Moreover, our analysis of agent interactions within AgentVerse reveals the emergence of specific collaborative behaviors, contributing to heightened group efficiency. We will release our codebase, AgentVerse, to further facilitate multi-agent research.
****
Jingyang Qiao,zhizhong zhang,Xin Tan,Chengwei Chen,Yanyun Qu,Yong Peng,Yuan Xie
Prompt Gradient Projection for Continual Learning
Prompt-tuning has demonstrated impressive performance in continual learning by querying relevant prompts for each input instance, which can avoid the introduction of task identifier. Its forgetting is therefore reduced as this instance-wise query mechanism enables us to select and update only relevant prompts. In this paper, we further integrate prompt-tuning with gradient projection approach. Our observation is: prompt-tuning releases the necessity of task identifier for gradient projection method; and gradient projection provides theoretical guarantees against forgetting for prompt-tuning. This inspires a new prompt gradient projection approach (PGP) for continual learning. In PGP, we deduce that reaching the orthogonal condition for prompt gradient can effectively prevent forgetting via

the self-attention mechanism in vision-transformer. The condition equations are then realized by conducting Singular Value Decomposition (SVD) on an element-wise sum space between input space and prompt space. We validate our method on diverse datasets and experiments demonstrate the efficiency of reducing forgetting both in class incremental, online class incremental, and task incremental settings. The code is available at https://github.com/JingyangQiao/prompt-gradient-projection.
****
Jianshu Hu,Yunpeng Jiang,Paul Weng
Revisiting Data Augmentation in Deep Reinforcement Learning
Various data augmentation techniques have been recently proposed in image-based deep reinforcement learning (DRL).
Although they empirically demonstrate the effectiveness of data augmentation for improving sample efficiency or generalization, which technique should be preferred is not always clear.
To tackle this question, we analyze existing methods to better understand them and to uncover how they are connected.
Notably, by expressing the variance of the Q-targets and that of the empirical actor/critic losses of these methods, we can analyze the effects of their different components and compare them.
We furthermore formulate an explanation about how these methods may be affected by choosing different data augmentation transformations in calculating the target Q-values.
This analysis suggests recommendations on how to exploit data augmentation in a more principled way.
In addition, we include a regularization term called tangent prop, previously proposed in computer vision, but whose adaptation to DRL is novel to the best of our knowledge.
We evaluate our proposition and validate our analysis in several domains.
Compared to different relevant baselines, we demonstrate that it achieves state-of-the-art performance in most environments and shows higher sample efficiency and better generalization ability in some complex environments.
****
Zhilong Zhang,Yihao Sun,Junyin Ye,Tian-Shuo Liu,Jiaji Zhang,Yang Yu
Flow to Better: Offline Preference-based Reinforcement Learning via Preferred Trajectory Generation
Offline preference-based reinforcement learning (PbRL) offers an effective solution to overcome the challenges associated with designing rewards and the high costs of online interactions. In offline PbRL, agents are provided with a fixed dataset containing human preferences between pairs of trajectories. Previous studies mainly focus on recovering the rewards from the preferences, followed by policy optimization with an off-the-shelf offline RL algorithm. However, given that preference label in PbRL is inherently trajectory-based, accurately learning transition-wise rewards from such label can be challenging, potentially leading to misguidance during subsequent offline RL training. To address this issue, we introduce our method named $\textit{Flow-to-Better (FTB)}$, which leverages the pairwise preference relationship to guide a generative model in producing preferred trajectories, avoiding Temporal Difference (TD) learning with inaccurate rewards. Conditioning on a low-preference trajectory, $\textit{FTB}$ uses a diffusion model to generate a better one with a higher preference, achieving high-fidelity full-horizon trajectory improvement. During diffusion training, we propose a technique called $\textit{Preference Augmentation}$ to alleviate the problem of insufficient preference data. As a result, we surprisingly find that the model-generated trajectories not only exhibit increased preference and consistency with the real transition but also introduce elements of $\textit{novelty}$ and $\textit{diversity}$, from which we can derive a desirable policy through imitation learning. Experimental results on D4RL benchmarks demonstrate that FTB achieves a remarkable improvement compared to state-of-the-art offline PbRL methods. Furthermore, we show that FTB can also serve as an effective data augmentation method for offline RL.

****
Sheng-Jun Huang,Yi Li,Yiming Sun,Ying-Peng Tang
One-shot Active Learning Based on Lewis Weight Sampling for Multiple Deep Models
Active learning (AL) for multiple target models aims to reduce labeled data quer ying while effectively training multiple models concurrently. Existing AL algori thms often rely on iterative model training, which can be computationally expens ive, particularly for deep models. In this paper, we propose a one-shot AL metho d to address this challenge, which performs all label queries without repeated m odel training. Specifically, we extract different representations of the same da taset using distinct network backbones, and actively learn the linear prediction layer on each representation via an ℓ_p-regression formulation. The regres sion problems are solved approximately by
sampling and reweighting the unlabeled instances based on their maximum Lewis we ights across the representations. An upper bound on the number of samples needed is provided with a rigorous analysis for $p\in [1, +\infty)$. Experimental resu lts on 11 benchmarks show that our one-shot approach achieves competitive perfor mances with the state-of-the-art AL methods for multiple target models.
****
Shyamgopal Karthik,Karsten Roth,Massimiliano Mancini,Zeynep Akata
Vision-by-Language for Training-Free Compositional Image Retrieval
Given an image and a target modification (e.g an image of the Eiffel tower and t he text “without people and at night-time”), Compositional Image Retrieval (CIR) aims to retrieve the relevant target image in a database. While supervised appr oaches rely on annotating triplets that is costly (i.e. query image, textual mod ification, and target image), recent research sidesteps this need by using large -scale vision-language models (VLMs), performing Zero-Shot CIR (ZS-CIR). However , state-of-the-art approaches in ZS-CIR still require training task-specific, cu stomized models over large amounts of image-text pairs. In this work, we propose to tackle CIR in a training-free manner via our Compositional Image Retrieval th rough Vision-by-Language (CIReVL), a simple, yet human-understandable and scalab le pipeline that effectively recombines large-scale VLMs with large language mod els (LLMs). By captioning the reference image using a pre-trained generative VLM and asking a LLM to recompose the caption based on the textual target modificat ion for subsequent retrieval via e.g. CLIP, we achieve modular language reasonin g. In four ZS-CIR benchmarks, we find competitive, in-part state-of-the-art perf ormance - improving over supervised methods Moreover, the modularity of CIReVL o ffers simple scalability without re-training, allowing us to both investigate sc aling laws and bottlenecks for ZS-CIR while easily scaling up to in parts more t han double of previously reported results. Finally, we show that CIReVL makes CI R human-understandable by composing image and text in a modular fashion in the l anguage domain, thereby making it intervenable, allowing to post-hoc re-align fa ilure cases. Code will be released upon acceptance.
****
Chenyu Liu,XINLIANG ZHOU,Zhengri Zhu,Liming Zhai,Ziyu Jia,Yang Liu
VBH-GNN: Variational Bayesian Heterogeneous Graph Neural Networks for Cross-subj ect Emotion Recognition
The research on human emotion under electroencephalogram (EEG) is an emerging fi eld in which cross-subject emotion recognition (ER) is a promising but challengi ng task. Many approaches attempt to find emotionally relevant domain-invariant f eatures using domain adaptation (DA) to improve the accuracy of cross-subject ER . However, two problems still exist with these methods. First, only single-modal data (EEG) is utilized, ignoring the complementarity between multi-modal physio logical signals. Second, these methods aim to completely match the signal featur es between different domains, which is difficult due to the extreme individual d ifferences of EEG. To solve these problems, we introduce the complementarity of multi-modal physiological signals and propose a new method for cross-subject ER that does not align the distribution of signal features but rather the distribut ion of spatio-temporal relationships between features. We design a Variational B ayesian Heterogeneous Graph Neural Network (VBH-GNN) with Relationship Distribut ion Adaptation (RDA). The RDA first aligns the domains by expressing the model s

pace as a posterior distribution of a heterogeneous graph for a given source dom ain. Then, the RDA transforms the heterogeneous graph into an emotion-specific g raph to further align the domains for the downstream ER task. Extensive experime nts on two public datasets, DEAP and Dreamer, show that our VBH-GNN outperforms state-of-the-art methods in cross-subject scenarios.
****
Chunsan Hong,ByungHee Cha,Tae-Hyun Oh
CAS: A Probability-Based Approach for Universal Condition Alignment Score
Recent conditional diffusion models have shown remarkable advancements and have been widely applied in fascinating real-world applications. However, samples gen erated by these models often do not strictly comply with user-provided condition s. Due to this, there have been few attempts to evaluate this alignment via pre-trained scoring models to select well-generated samples. Nonetheless, current st udies are confined to the text-to-image domain and require large training datase ts. This suggests that crafting alignment scores for various conditions will dem and considerable resources in the future. In this context, we introduce a univer sal condition alignment score that leverages the conditional probability measura ble through the diffusion process. Our technique operates across all conditions and requires no additional models beyond the diffusion model used for generation , effectively enabling self-rejection. Our experiments validate that our met- ri c effectively applies in diverse conditional generations, such as text-to-image, {instruction, image}-to-image, edge-/scribble-to-image, and text-to-audio.
****
Adel Javanmard,Lin Chen,Vahab Mirrokni,Ashwinkumar Badanidiyuru,Gang Fu
Learning from Aggregate responses: Instance Level versus Bag Level Loss Function s
Due to the rise of privacy concerns, in many practical applications, the trainin g data is aggregated before being shared with the learner to protect the privacy of users' sensitive responses. In an aggregate learning framework, the dataset is grouped into bags of samples, where each bag is available only with an aggreg ate response, providing a summary of individuals' responses in that bag. In this paper, we study two natural loss functions for learning from aggregate response s: the bag-level loss and the instance-level loss. In the former, the model is l earned by minimizing a loss between the aggregate responses and aggregate model predictions, while in the latter, the model aims to fit individual predictions t o the aggregate responses. In this work, we show that the instance-level loss ca n be perceived as a regularized form of the bag-level loss. This observation all ows us to compare the two approaches with respect to the bias and variance of th e resulting estimators and to introduce a novel interpolating estimator that com bines the two approaches. For linear regression tasks, we provide a precise char acterization of the risk of the interpolating estimator in an asymptotic regime where the size of the training set grows in proportion to the feature dimension. Our analysis enables us to theoretically understand the effect of different fac tors, such as bag size, on the model's prediction risk. Additionally, we propose a mechanism for differentially private learning from aggregate responses and de rive the optimal bag size in terms of the prediction risk-privacy trade-off. We also carry out thorough experiments to corroborate our theory and show the effic acy of the interpolating estimator.
****
Yi Heng Lim,Qi Zhu,Joshua Selfridge,Muhammad Firmansyah Kasim
Parallelizing non-linear sequential models over the sequence length
Sequential models, such as Recurrent Neural Networks and Neural Ordinary Differe ntial Equations, have long suffered from slow training due to their inherent seq uential nature.
For many years this bottleneck has persisted, as many thought sequential models could not be parallelized.
We challenge this long-held belief with our parallel algorithm that accelerates GPU evaluation of sequential models by up to 3 orders of magnitude faster withou t compromising output accuracy.
The algorithm does not need any special structure in the sequential models' arch

itecture, making it applicable to a wide range of architectures.
Using our method, training sequential models can be more than 10 times faster than the common sequential method without any meaningful difference in the training results.
Leveraging this accelerated training, we discovered the efficacy of the Gated Recurrent Unit in a long time series classification problem with 17k time samples. By overcoming the training bottleneck, our work serves as the first step to unlock the potential of non-linear sequential models for long sequence problems.
****
Somnath Basu Roy Chowdhury,Nicholas Monath,Ahmad Beirami,Rahul Kidambi,Kumar Avinava Dubey,Amr Ahmed,Snigdha Chaturvedi
Enhancing Group Fairness in Online Settings Using Oblique Decision Forests
Fairness, especially group fairness, is an important consideration in the context of machine learning systems. The most commonly adopted group fairness-enhancing techniques are in-processing methods that rely on a mixture of a fairness objective (e.g., demographic parity) and a task-specific objective (e.g., cross-entropy) during the training process. However, when data arrives in an online fashion – one instance at a time – optimizing such fairness objectives poses several challenges. In particular, group fairness objectives are defined using expectations of predictions across different demographic groups. In the online setting, where the algorithm has access to a single instance at a time, estimating the group fairness objective requires additional storage and significantly more computation (e.g., forward/backward passes) than the task-specific objective at every time step. In this paper, we propose Aranyani, an ensemble of oblique decision trees, to make fair decisions in online settings. The hierarchical tree structure of Aranyani enables parameter isolation and allows us to efficiently compute the fairness gradients using aggregate statistics of previous decisions, eliminating the need for additional storage and forward/backward passes. We also present an efficient framework to train Aranyani and theoretically analyze several of its properties. We conduct empirical evaluations on 5 publicly available benchmarks (including vision and language datasets) to show that Aranyani achieves a better accuracy-fairness trade-off compared to baseline approaches.
****
Jonathan Daniel Chang,Dhruv Sreenivas,Yingbing Huang,Kianté Brantley,Wen Sun
Adversarial Imitation Learning via Boosting
Adversarial imitation learning (AIL) has stood out as a dominant framework across various imitation learning (IL) applications, with Discriminator Actor Critic (DAC) demonstrating the effectiveness of off-policy learning algorithms in improving sample efficiency and scalability to higher-dimensional observations. Despite DAC’s empirical success, the original AIL objective is on-policy and DAC’s ad-hoc application of off-policy training does not guarantee successful imitation. Follow-up work such as ValueDICE tackles this issue by deriving a fully off-policy AIL objective. Instead in this work, we develop a novel and principled AIL algorithm via the framework of boosting. Like boosting, our new algorithm, AILBoost, maintains an ensemble of weighted weak learners (i.e., policies) and trains a discriminator that witnesses the maximum discrepancy between the distributions of the ensemble and the expert policy. We maintain a weighted replay buffer to represent the state-action distribution induced by the ensemble, allowing us to train discriminators using the entire data collected so far. Empirically, we evaluate our algorithm on both controller state-based and pixel-based environments from the DeepMind Control Suite. AILBoost outperforms DAC on both types of environments, demonstrating the benefit of properly weighting replay buffer data for off-policy training. On state-based environments, AILBoost outperforms ValueDICE and IQ-Learn, achieving state-of-the-art performance with as little as one expert trajectory.
****
Hyungjin Chung,Suhyeon Lee,Jong Chul Ye
Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems
Krylov subspace, which is generated by multiplying a given vector by the matrix of a linear transformation and its successive powers, has been extensively stud

ied in classical optimization literature to design algorithms that converge quickly for large linear inverse problems. For example, the conjugate gradient method (CG), one of the most popular Krylov subspace methods, is based on the idea of minimizing the residual error in the Krylov subspace. However, with the recent advancement of high-performance diffusion solvers for inverse problems, it is not clear how classical wisdom can be synergistically combined with modern diffusion models. In this study, we propose a novel and efficient diffusion sampling strategy that synergistically combines the diffusion sampling and Krylov subspace methods. Specifically, we prove that if the tangent space at a denoised sample by Tweedie's formula forms a Krylov subspace, then the CG initialized with the denoised data ensures the data consistency update to remain in the tangent space. This negates the need to compute the manifold-constrained gradient (MCG), leading to a more efficient diffusion sampling method. Our method is applicable regardless of the parametrization and setting (i.e., VE, VP). Notably, we achieve state-of-the-art reconstruction quality on challenging real-world medical inverse imaging problems, including multi-coil MRI reconstruction and 3D CT reconstruction. Moreover, our proposed method achieves more than 80 times faster inference time than the previous state-of-the-art method. Code is available at https://github.com/HJ-harry/DDS
****
Sigal Raab,Inbal Leibovitch,Guy Tevet,Moab Arar,Amit Haim Bermano,Daniel Cohen-Or
Single Motion Diffusion
Synthesizing realistic animations of humans, animals, and even imaginary creatures, has long been a goal for artists and computer graphics professionals. Compared to the imaging domain, which is rich with large available datasets, the number of data instances for the motion domain is limited, particularly for the animation of animals and exotic creatures (e.g., dragons), which have unique skeletons and motion patterns. In this work, we introduce SinMDM, a Single Motion Diffusion Model. It is designed to learn the internal motifs of a single motion sequence with arbitrary topology and synthesize a variety of motions of arbitrary length that remain faithful to the learned motifs. We harness the power of diffusion models and present a denoising network explicitly designed for the task of learning from a single input motion. SinMDM is crafted as a lightweight architecture, which avoids overfitting by using a shallow network with local attention layers that narrow the receptive field and encourage motion diversity. Our work applies to multiple contexts, including spatial and temporal in-betweening, motion expansion, style transfer, and crowd animation. Our results show that SinMDM outperforms existing methods both qualitatively and quantitatively. Moreover, while prior network-based approaches require additional training for different applications, SinMDM supports these applications during inference. Our project page, which includes links to the code and trained models, is accessible at https://sinmdm.github.io/SinMDM-page.
****
Jadie Adams,Shireen Elhabian
Point2SSM: Learning Morphological Variations of Anatomies from Point Clouds
We present Point2SSM, a novel unsupervised learning approach for constructing correspondence-based statistical shape models (SSMs) directly from raw point clouds. SSM is crucial in clinical research, enabling population-level analysis of morphological variation in bones and organs. Traditional methods of SSM construction have limitations, including the requirement of noise-free surface meshes or binary volumes, reliance on assumptions or templates, and prolonged inference times due to simultaneous optimization of the entire cohort. Point2SSM overcomes these barriers by providing a data-driven solution that infers SSMs directly from raw point clouds, reducing inference burdens and increasing applicability as point clouds are more easily acquired. While deep learning on 3D point clouds has seen success in unsupervised representation learning and shape correspondence, its application to anatomical SSM construction is largely unexplored. We conduct a benchmark of state-of-the-art point cloud deep networks on the SSM task, revealing their limited robustness to clinical challenges such as noisy, sparse, or in

complete input and limited training data. Point2SSM addresses these issues throu
gh an attention-based module, providing effective correspondence mappings from l
earned point features. Our results demonstrate that the proposed method signific
antly outperforms existing networks in terms of accurate surface sampling and co
rrespondence, better capturing population-level statistics. The source code is p
rovided at https://github.com/jadie1/Point2SSM.
****
Maresa Schröder,Dennis Frauen,Stefan Feuerriegel
Causal Fairness under Unobserved Confounding: A Neural Sensitivity Framework
Fairness for machine learning predictions is widely required in practice for leg
al, ethical, and societal reasons. Existing work typically focuses on settings w
ithout unobserved confounding, even though unobserved confounding can lead to se
vere violations of causal fairness and, thus, unfair predictions. In this work,
we analyze the sensitivity of causal fairness to unobserved confounding. Our con
tributions are three-fold. First, we derive bounds for causal fairness metrics u
nder different sources of observed confounding. This enables practitioners to au
dit the sensitivity of their machine learning models to unobserved confounding i
n fairness-critical applications. Second, we propose a novel neural framework fo
r learning fair predictions, which allows us to offer worst-case guarantees of t
he extent to which causal fairness can be violated due to unobserved confounding
. Third, we demonstrate the effectiveness of our framework in a series of experi
ments, including a real-world case study about predicting prison sentences. To t
he best of our knowledge, ours is the first work to study causal fairness under
observed confounding. To this end, our work is of direct practical value for aud
iting and ensuring the fairness of predictions in high-stakes applications.
****
Siyan Zhao,John Dang,Aditya Grover
Group Preference Optimization: Few-Shot Alignment of Large Language Models
Many applications of large language models (LLMs), ranging from chatbots to
creative writing, require nuanced subjective judgments that can differ significa
ntly
across different groups. Existing alignment algorithms can be expensive to align
for each group, requiring prohibitive amounts of group-specific preference data
and computation for real-world use cases. We introduce Group Preference Optimiza
tion (GPO), an alignment framework that steers language models to preferences of
individual groups in a few-shot manner. In GPO, we augment the base
LLM with an independent transformer module trained to predict the preferences
of a group for the LLM generations. For few-shot learning, we parameterize this
module as an in-context autoregressive transformer and train it via meta-learnin
g
on several groups. We empirically validate the efficacy of GPO through rigorous
evaluations using LLMs with varied sizes on three human opinion adaptation tasks
. These tasks involve adapting to the preferences of US demographic
groups, global countries, and individual users. Our results demonstrate that GPO
not only aligns models more accurately but also requires fewer group-specific
preferences and less training and inference computing resources, outperforming
existing strategies such as in-context steering and fine-tuning methods.
****
Driton Salihu,Adam Misik,Yuankai Wu,Constantin Patsch,Fabian Esteban Seguel,Ecke
hard Steinbach
DeepSPF: Spherical SO(3)-Equivariant Patches for Scan-to-CAD Estimation
Recently, SO(3)-equivariant methods have been explored for 3D reconstruction via
Scan-to-CAD.
Despite significant advancements attributed to the unique characteristics of 3D
data, existing SO(3)-equivariant approaches often fall short in seamlessly integ
rating local and global contextual information in a widely generalizable manner.
Our contributions in this paper are threefold.
First, we introduce Spherical Patch Fields, a representation technique designed
for patch-wise, SO(3)-equivariant 3D point clouds, anchored theoretically on the
principles of Spherical Gaussians.

Second, we present the Patch Gaussian Layer, designed for the adaptive extractio
n of local and global contextual information from resizable point cloud patches.
Culminating our contributions, we present Learnable Spherical Patch Fields (Deep
SPF) – a versatile and easily integrable backbone suitable for instance-based po
int networks.
Through rigorous evaluations, we demonstrate significant enhancements in Scan-to
-CAD performance for point cloud registration, retrieval, and completion: a sign
ificant reduction in the rotation error of existing registration methods, an imp
rovement of up to 17\% in the Top-1 error for retrieval tasks, and a notable red
uction of up to 30\% in the Chamfer Distance for completion models, all attribut
able to the incorporation of DeepSPF.
****
Jiangmeng Li,Fei Song,Yifan Jin,Wenwen Qiang,Changwen Zheng,Fuchun Sun,Hui Xiong
BayesPrompt: Prompting Large-Scale Pre-Trained Language Models on Few-shot Infer
ence via Debiased Domain Abstraction
As a novel and effective fine-tuning paradigm based on large-scale pre-trained l
anguage models (PLMs), prompt-tuning aims to reduce the gap between downstream t
asks and pre-training objectives. While prompt-tuning has yielded continuous adv
ancements in various tasks, such an approach still remains a persistent defect:
prompt-tuning methods fail to generalize to specific few-shot patterns. From the
 perspective of distribution analyses, we disclose that the intrinsic issues beh
ind the phenomenon are the over-multitudinous conceptual knowledge contained in
PLMs and the abridged knowledge for target downstream domains, which jointly res
ult in that PLMs mis-locate the knowledge distributions corresponding to the tar
get domains in the universal knowledge embedding space. To this end, we intuitiv
ely explore to approximate the unabridged target domains of downstream tasks in
a debiased manner, and then abstract such domains to generate discriminative pro
mpts, thereby providing the de-ambiguous guidance for PLMs. Guided by such an in
tuition, we propose a simple yet effective approach, namely BayesPrompt, to lear
n prompts that contain the domain discriminative information against the interfe
rence from domain-irrelevant knowledge. BayesPrompt primitively leverages known
distributions to approximate the debiased factual distributions of target domain
s and further uniformly samples certain representative features from the approxi
mated distributions to generate the ultimate prompts for PLMs. We provide theore
tical insights with the connection to domain adaptation. Empirically, our method
 achieves state-of-the-art performance on benchmarks.
****
Size Wu,Wenwei Zhang,Lumin Xu,Sheng Jin,Xiangtai Li,Wentao Liu,Chen Change Loy
CLIPSelf: Vision Transformer Distills Itself for Open-Vocabulary Dense Predictio
n
Open-vocabulary dense prediction tasks including object detection and image segm
entation have been advanced by the success of Contrastive Language-Image Pre-tra
ining (CLIP). CLIP models, particularly those incorporating vision transformers
(ViTs), have exhibited remarkable generalization ability in zero-shot image clas
sification. However, when transferring the vision-language alignment of CLIP fro
m global image representation to local region representation for the open-vocabu
lary dense prediction tasks, CLIP ViTs suffer from the domain shift from full im
ages to local image regions. In this paper, we embark on an in-depth analysis of
 the region-language alignment in CLIP models, which is essential for downstream
 open-vocabulary dense prediction tasks. Subsequently, we propose an approach na
med CLIPSelf, which adapts the image-level recognition ability of CLIP ViT to lo
cal image regions without needing any region-text pairs. CLIPSelf empowers ViTs
to distill itself by aligning a region representation extracted from its dense f
eature map with the image-level representation of the corresponding image crop.
With the enhanced CLIP ViTs, we achieve new state-of-the-art performance on open
-vocabulary object detection, semantic segmentation, and panoptic segmentation a
cross various benchmarks. Models and code are released at https://github.com/wus
ize/CLIPSelf.
****
Kelly Maggs,Celia Hacker,Bastian Rieck

Simplicial Representation Learning with Neural k-Forms
Geometric deep learning extends deep learning to incorporate information about t
he geometry and topology data, especially in complex domains like graphs. Despit
e the popularity of message passing in this field, it has limitations such as th
e need for graph rewiring, ambiguity in interpreting data, and over-smoothing. I
n this paper, we take a different approach, focusing on leveraging geometric inf
ormation from simplicial complexes embedded in $\mathbb{R}^n$ using node coordin
ates. We use differential k-forms in $\mathbb{R}^n$ to create representations
of simplices, offering interpretability and geometric consistency without messag
e passing. This approach also enables us to apply differential geometry tools an
d achieve universal approximation. Our method is efficient, versatile, and appli
cable to various input complexes, including graphs, simplicial complexes, and ce
ll complexes. It outperforms existing message passing neural networks in harness
ing information from geometrical graphs with node features serving as coordinate
s.
****
Annan Yu,Arnur Nigmetov,Dmitriy Morozov,Michael W. Mahoney,N. Benjamin Erichson
Robustifying State-space Models for Long Sequences via Approximate Diagonalizati
on
State-space models (SSMs) have recently emerged as a framework for learning long
-range sequence tasks. An example is the structured state-space sequence (S4) la
yer, which uses the diagonal-plus-low-rank structure of the HiPPO initialization
 framework. However, the complicated structure of the S4 layer poses challenges;
 and, in an effort to address these challenges, models such as S4D and S5 have c
onsidered a purely diagonal structure. This choice simplifies the implementation
, improves computational efficiency, and allows channel communication. However,
diagonalizing the HiPPO framework is itself an ill-posed problem. In this paper,
 we propose a general solution for this and related ill-posed diagonalization pr
oblems in machine learning. We introduce a generic, backward-stable ``perturb-th
en-diagonalize'' (PTD) methodology, which is based on the pseudospectral theory
of non-normal operators, and which may be interpreted as the approximate diagona
lization of the non-normal matrices defining SSMs. Based on this, we introduce t
he S4-PTD and S5-PTD models. Through theoretical analysis of the transfer functi
ons of different initialization schemes, we demonstrate that the S4-PTD/S5-PTD i
nitialization strongly converges to the HiPPO framework, while the S4D/S5 initia
lization only achieves weak convergences. As a result, our new models show resil
ience to Fourier-mode noise-perturbed inputs, a crucial property not achieved by
 the S4D/S5 models. In addition to improved robustness, our S5-PTD model average
s 87.6% accuracy on the Long-Range Arena benchmark, demonstrating that the PTD m
ethodology helps to improve the accuracy of deep learning models.
****
Hyunho Kim,Jong-Seok Lee
Scalable Monotonic Neural Networks
In this research, we focus on the problem of learning monotonic neural networks,
 as preserving the monotonicity of a model with respect to a subset of inputs is
 crucial for practical applications across various domains. Although several met
hods have recently been proposed to address this problem, they have limitations
such as not guaranteeing monotonicity in certain cases, requiring additional inf
erence time, lacking scalability with increasing network size and number of mono
tonic inputs, and manipulating network weights during training. To overcome thes
e limitations, we introduce a simple but novel architecture of the partially con
nected network which incorporates a 'scalable monotonic hidden layer' comprising
 three units: the exponentiated unit, ReLU unit, and confluence unit. This allow
s for the repetitive integration of the scalable monotonic hidden layers without
 other structural constraints. Consequently, our method offers ease of implement
ation and rapid training through the conventional error-backpropagation algorith
m. We accordingly term this method as Scalable Monotonic Neural Networks (SMNN).
 Numerical experiments demonstrated that our method achieved comparable predicti
on accuracy to the state-of-the-art approaches while effectively addressing the
aforementioned weaknesses.

****
Shengyi Huang,Jiayi Weng,Rujikorn Charakorn,Min Lin,Zhongwen Xu,Santiago Ontanon
Cleanba: A Reproducible and Efficient Distributed Reinforcement Learning Platform
Distributed Deep Reinforcement Learning (DRL) aims to leverage more computational resources to train autonomous agents with less training time. Despite recent progress in the field, reproducibility issues have not been sufficiently explored. This paper first shows that the typical actor-learner framework can have reproducibility issues even if hyperparameters are controlled. We then introduce Cleanba, a new open-source platform for distributed DRL that proposes a highly reproducible architecture. Cleanba implements highly optimized distributed variants of PPO and IMPALA. Our Atari experiments show that these variants can obtain equivalent or higher scores than strong IMPALA baselines in moolib and torchbeast and PPO baseline in CleanRL. However, Cleanba variants present 1) shorter training time and 2) more reproducible learning curves in different hardware settings.
****
Min Zhang,Haoxuan Li,Fei Wu,Kun Kuang
MetaCoCo: A New Few-Shot Classification Benchmark with Spurious Correlation
Out-of-distribution (OOD) problems in few-shot classification (FSC) occur when novel classes sampled from testing distributions differ from base classes drawn from training distributions, which considerably degrades the performance of deep learning models deployed in real-world applications. Recent studies suggest that the OOD problems in FSC mainly including: (a) cross-domain few-shot classification (CD-FSC) and (b) spurious-correlation few-shot classification (SC-FSC). Specifically, CD-FSC occurs when a classifier learns transferring knowledge from base classes drawn from \underline{seen} training distributions but recognizes novel classes sampled from unseen testing distributions. In contrast, SC-FSC arises when a classifier relies on non-causal features (or contexts) that happen to be correlated with the labels (or concepts) in base classes but such relationships no longer hold during the model deployment. Despite CD-FSC has been extensively studied, SC-FSC remains understudied due to lack of the corresponding evaluation benchmarks. To this end, we present Meta Concept Context (MetaCoCo), a benchmark with spurious-correlation shifts collected from real-world scenarios. Moreover, to quantify the extent of spurious-correlation shifts of the presented MetaCoCo, we further propose a metric by using CLIP as a pre-trained vision-language model. Extensive experiments on the proposed benchmark are performed to evaluate the state-of-the-art methods in FSC, cross-domain shifts, and self-supervised learning. The experimental results show that the performance of the existing methods degrades significantly in the presence of spurious-correlation shifts. We open-source all codes of our benchmark and hope that the proposed MetaCoCo can facilitate future research on spurious-correlation shifts problems in FSC.
****
Yassine ABBAHADDOU,Sofiane ENNADIR,Johannes F. Lutzeyer,Michalis Vazirgiannis,Henrik Boström
Bounding the Expected Robustness of Graph Neural Networks Subject to Node Feature Attacks
Graph Neural Networks (GNNs) have demonstrated state-of-the-art performance in various graph representation learning tasks. Recently, studies revealed their vulnerability to adversarial attacks. In this work, we theoretically define the concept of expected robustness in the context of attributed graphs and relate it to the classical definition of adversarial robustness in the graph representation learning literature. Our definition allows us to derive an upper bound of the expected robustness of Graph Convolutional Networks (GCNs) and Graph Isomorphism Networks subject to node feature attacks. Building on these findings, we connect the expected robustness of GNNs to the orthogonality of their weight matrices and consequently propose an attack-independent, more robust variant of the GCN, called the Graph Convolutional Orthogonal Robust Networks (GCORNs). We further introduce a probabilistic method to estimate the expected robustness, which allows us to evaluate the effectiveness of GCORN on several real-world datasets. Experimental experiments showed that GCORN outperforms available defense methods.

****
Shen Nie,Hanzhong Allan Guo,Cheng Lu,Yuhao Zhou,Chenyu Zheng,Chongxuan Li
The Blessing of Randomness: SDE Beats ODE in General Diffusion-based Image Editi
ng
We present a unified probabilistic formulation for diffusion-based image editing
, where a latent variable is edited in a task-specific manner and generally devi
ates from the corresponding marginal distribution induced by the original stocha
stic or ordinary differential equation (SDE or ODE). Instead, it defines a corre
sponding SDE or ODE for editing. In the formulation, we prove that the Kullback-
Leibler divergence between the marginal distributions of the two SDEs gradually
decreases while that for the ODEs remains as the time approaches zero, which sho
ws the promise of SDE in image editing. Inspired by it, we provide the SDE count
erparts for widely used ODE baselines in various tasks including inpainting and
image-to-image translation, where SDE shows a consistent and substantial improve
ment. Moreover, we propose \emph{SDE-Drag} -- a simple yet effective method bui
lt upon the SDE formulation for point-based content dragging. We build a challen
ging benchmark (termed \emph{DragBench}) with open-set natural, art, and AI-gene
rated images for evaluation. A user study on DragBench indicates that SDE-Drag s
ignificantly outperforms our ODE baseline, existing diffusion-based methods, and
 the renowned DragGAN. Our results demonstrate the superiority and versatility o
f SDE in image editing and push the boundary of diffusion-based editing methods.
  See the project page \url{https://ml-gsai.github.io/SDE-Drag-demo/} for the co
de and DragBench dataset.
****
Huaxiu Yao,Xinyu Yang,Xinyi Pan,Shengchao Liu,Pang Wei Koh,Chelsea Finn
Improving Domain Generalization with Domain Relations
Distribution shift presents a significant challenge in machine learning, where m
odels often underperform during the test stage when faced with a different distr
ibution than the one they were trained on. In this paper, we focus on domain shi
fts, which occur when the model is applied to new domains that are different fro
m the ones it was trained on, and propose a new approach called DG. Unlike previ
ous approaches that aim to learn a single model that is domain invariant, DG lev
erages domain similarities based on domain metadata to learn domain-specific mod
els. Concretely, DG learns a set of training-domain-specific functions during th
e training stage and reweights them based on domain relations during the test st
age. These domain relations can be directly obtained and learned from domain met
adata. Under mild assumptions, we theoretically prove that using domain relation
s to reweight training-domain-specific functions achieves stronger out-of-domain
 generalization compared to the conventional averaging approach. Empirically, we
 evaluate the effectiveness of DG using both toy and real-world datasets for tas
ks such as temperature regression, land use classification, and molecule-protein
 binding affinity prediction. Our results show that DG consistently outperforms
state-of-the-art methods.
****
Duong Minh Le,Yang Chen,Alan Ritter,Wei Xu
Constrained Decoding for Cross-lingual Label Projection
Zero-shot cross-lingual transfer utilizing multilingual LLMs has become a popula
r learning paradigm for low-resource languages with no labeled training data. Ho
wever, for NLP tasks that involve fine-grained predictions on words and phrases,
 the performance of zero-shot cross-lingual transfer learning lags far behind su
pervised fine-tuning methods. Therefore, it is common to exploit translation and
 label projection to further improve the performance by (1) translating training
 data that is available in a high-resource language (e.g., English) together wit
h the gold labels into low-resource languages, and/or (2) translating test data
in low-resource languages to a high-source language to run inference on, then pr
ojecting the predicted span-level labels back onto the original test data. Howev
er, state-of-the-art marker-based label projection methods suffer from translati
on quality degradation due to the extra label markers injected in the input to t
he translation model. In this work, we explore a new direction that leverages co
nstrained decoding for label projection to overcome the aforementioned issues. O

ur new method not only can preserve the quality of translated texts but also has the versatility of being applicable to both translating training and translating test data strategies. This versatility is crucial as our experiments reveal that translating test data can lead to a considerable boost in performance compared to translating only training data. We evaluate on two cross-lingual transfer tasks, namely Named Entity Recognition and Event Argument Extraction, spanning 20 languages. The results demonstrate that our approach outperforms the state-of-the-art marker-based method by a large margin and also shows better performance than other label projection methods that rely on external word alignment.
****
Kiarash Shamsi,Farimah Poursafaei,Shenyang Huang,Bao Tran Gia Ngo,Baris Coskunuzer,Cuneyt Gurcan Akcora
GraphPulse: Topological representations for temporal graph property prediction
Many real-world networks evolve over time, and predicting the evolution of such networks remains a challenging task. Graph Neural Networks (GNNs) have shown empirical success for learning on static graphs, but they lack the ability to effectively learn from nodes and edges with different timestamps. Consequently, the prediction of future properties in temporal graphs remains a relatively under-explored area.
In this paper, we aim to bridge this gap by introducing a principled framework, named GraphPulse. The framework combines two important techniques for the analysis of temporal graphs within a Newtonian framework. First, we employ the Mapper method, a key tool in topological data analysis, to extract essential clustering information from graph nodes. Next, we harness the sequential modeling capabilities of Recurrent Neural Networks (RNNs) for temporal reasoning regarding the graph's evolution. Through extensive experimentation, we demonstrate that our model enhances the ROC-AUC metric by 10.2\% in comparison to the top-performing state-of-the-art method across various temporal networks. We provide the implementation of GraphPulse at https://github.com/kiarashamsi/GraphPulse.
****
Sunwoo Kim,Shinhwan Kang,Fanchen Bu,Soo Yong Lee,Jaemin Yoo,Kijung Shin
HypeBoy: Generative Self-Supervised Representation Learning on Hypergraphs
Hypergraphs are marked by complex topology, expressing higher-order interactions among multiple nodes with hyperedges, and better capturing the topology is essential for effective representation learning. Recent advances in generative self-supervised learning (SSL) suggest that hypergraph neural networks (HNNs) learned from generative self-supervision have the potential to effectively encode the complex hypergraph topology. Designing a generative SSL strategy for hypergraphs, however, is not straightforward. Questions remain with regard to its generative SSL task, connection to downstream tasks, and empirical properties of learned representations. In light of the promises and challenges, we propose a novel generative SSL strategy for hypergraphs. We first formulate a generative SSL task on hypergraphs, hyperedge filling, and highlight its theoretical connection to node classification. Based on the generative SSL task, we propose a hypergraph SSL method, HYPEBOY. HYPEBOY learns effective general-purpose hypergraph representations, outperforming 15 baseline methods across 11 benchmark datasets. To our knowledge, this is the first study on generative SSL on hypergraphs, and we demonstrate its theoretical and empirical strengths for hypergraph representation learning.
****
Tycho F. A. van der Ouderaa,Markus Nagel,Mart Van Baalen,Tijmen Blankevoort
The LLM Surgeon
State-of-the-art language models are becoming increasingly large in an effort to achieve the highest performance on large corpora of available textual data. However, the sheer size of the Transformer architectures makes it difficult to deploy models within computational, environmental or device-specific constraints. We explore data-driven compression of existing pretrained models as an alternative to training smaller models from scratch. To do so, we scale Kronecker-factored curvature approximations of the target loss landscape to large language models. In doing so, we can compute both the dynamic allocation of structures that can b

e removed as well as updates of remaining weights that account for the removal. We provide a general framework for unstructured, semi-structured and structured pruning and improve upon weight updates to capture more correlations between weights, while remaining computationally efficient. Experimentally, our method can prune rows and columns from a range of OPT models and Llamav2-7B by 20\%-30\%, with a negligible loss in performance, and achieve state-of-the-art results in unstructured and semi-structured pruning of large language models. We will open source our code on GitHub upon acceptance.
****
Mingkun Yang,Ran Zhu,Qing Wang,Jie Yang
FedTrans: Client-Transparent Utility Estimation for Robust Federated Learning
Federated Learning (FL) is an important privacy-preserving learning paradigm that plays an important role in the Intelligent Internet of Things. Training a global model in FL, however, is vulnerable to the noise in the heterogeneous data across the clients. In this paper, we introduce **FedTrans**, a novel client-transparent client utility estimation method designed to guide client selection for noisy scenarios, mitigating performance degradation problems. To estimate the client utility, we propose a Bayesian framework that models client utility and its relationships with the weight parameters and the performance of local models. We then introduce a variational inference algorithm to effectively infer client utility, given only a small amount of auxiliary data. Our evaluation demonstrates that leveraging FedTrans as a guide for client selection can lead to a better accuracy performance (up to 7.8\%), ensuring robustness in noisy scenarios.
****
Zhiyuan Zhao,Xueying Ding,B. Aditya Prakash
PINNsFormer: A Transformer-Based Framework For Physics-Informed Neural Networks
Physics-Informed Neural Networks (PINNs) have emerged as a promising deep learning framework for approximating numerical solutions to partial differential equations (PDEs). However, conventional PINNs, relying on multilayer perceptrons (MLP), neglect the crucial temporal dependencies inherent in practical physics systems and thus fail to propagate the initial condition constraints globally and accurately capture the true solutions under various scenarios. In this paper, we introduce a novel Transformer-based framework, termed PINNsFormer, designed to address this limitation. PINNsFormer can accurately approximate PDE solutions by utilizing multi-head attention mechanisms to capture temporal dependencies. PINNsFormer transforms point-wise inputs into pseudo sequences and replaces point-wise PINNs loss with a sequential loss. Additionally, it incorporates a novel activation function, \texttt{Wavelet}, which anticipates Fourier decomposition through deep neural networks. Empirical results demonstrate that PINNsFormer achieves superior generalization ability and accuracy across various scenarios, including PINNs failure modes and high-dimensional PDEs. Moreover, PINNsFormer offers flexibility in integrating existing learning schemes for PINNs, further enhancing its performance.
****
Vladislav Lialin,Sherin Muckatira,Namrata Shivagunde,Anna Rumshisky
ReLoRA: High-Rank Training Through Low-Rank Updates
Despite the dominance and effectiveness of scaling, resulting in large networks with hundreds of billions of parameters, the necessity to train overparameterized models remains poorly understood, while training costs grow exponentially. In this paper, we explore parameter-efficient training techniques as an approach to training large neural networks. We introduce a novel method called ReLoRA, which utilizes low-rank updates to train high-rank networks. We apply ReLoRA to training transformer language models with up to 1.3B parameters and demonstrate comparable performance to regular neural network training. ReLoRA saves up to 5.5Gb of RAM per GPU and improves training speed by 9-40% depending on the model size and hardware setup. Our findings show the potential of parameter- efficient techniques for large-scale pre-training. Our code is available on GitHub.
****
Micha■ Zaj■c,Tinne Tuytelaars,Gido M van de Ven
Prediction Error-based Classification for Class-Incremental Learning

Class-incremental learning (CIL) is a particularly challenging variant of contin
ual learning, where the goal is to learn to discriminate between all classes pre
sented in an incremental fashion. Existing approaches often suffer from excessiv
e forgetting and imbalance of the scores assigned to classes that have not been
seen together during training. In this study, we introduce a novel approach, Pre
diction Error-based Classification (PEC), which differs from traditional discrim
inative and generative classification paradigms. PEC computes a class score by m
easuring the prediction error of a model trained to replicate the outputs of a f
rozen random neural network on data from that class. The method can be interpret
ed as approximating a classification rule based on Gaussian Process posterior va
riance. PEC offers several practical advantages, including sample efficiency, ea
se of tuning, and effectiveness even when data are presented one class at a time
. Our empirical results show that PEC performs strongly in single-pass-through-d
ata CIL, outperforming other rehearsal-free baselines in all cases and rehearsal
-based methods with moderate replay buffer size in most cases across multiple be
nchmarks.
****
Junhao Hu,Weijie Gan,Zhixin Sun,Hongyu An,Ulugbek Kamilov
A Plug-and-Play Image Registration Network
Deformable image registration (DIR) is an active research topic in biomedical im
aging. There is a growing interest in developing DIR methods based on deep learn
ing (DL). A traditional DL approach to DIR is based on training a convolutional
neural network (CNN) to estimate the registration field between two input images
. While conceptually simple, this approach comes with a limitation that it exclu
sively relies on a pre-trained CNN without explicitly enforcing fidelity between
 the registered image and the reference. We present plug-and-play image registra
tion network (PIRATE) as a new DIR method that addresses this issue by integrati
ng an explicit data-fidelity penalty and a CNN prior. PIRATE pre-trains a CNN de
noiser on the registration field and "plugs" it into an iterative method as a re
gularizer. We additionally present PIRATE+ that fine-tunes the CNN prior in PIRA
TE using deep equilibrium models (DEQ). PIRATE+ interprets the fixed-point itera
tion of PIRATE as a network with effectively infinite layers and then trains the
 resulting network end-to-end, enabling it to learn more task-specific informati
on and boosting its performance. Our numerical results on OASIS and CANDI datase
ts show that our methods achieve state-of-the-art performance on DIR.
****
Xiangyu Liu,Chenghao Deng,Yanchao Sun,Yongyuan Liang,Furong Huang
Beyond Worst-case Attacks: Robust RL with Adaptive Defense via Non-dominated Pol
icies
In light of the burgeoning success of reinforcement learning (RL) in diverse rea
l-world applications, considerable focus has been directed towards ensuring RL p
olicies are robust to adversarial attacks during test time. Current approaches l
argely revolve around solving a minimax problem to prepare for potential worst-c
ase scenarios. While effective against strong attacks, these methods often compr
omise performance in the absence of attacks or the presence of only weak attacks
. To address this, we study policy robustness under the well-accepted state-adve
rsarial attack model, extending our focus beyond merely worst-case attacks. We f
irst formalize this task at test time as a regret minimization problem and estab
lish its intrinsic difficulty in achieving sublinear regret when the baseline po
licy is from a general continuous policy class, Π. This finding prompts us t
o \textit{refine} the baseline policy class Π prior to test time, aiming for
 efficient adaptation within a compact, finite policy class $\tilde{\Pi}$, which
 can resort to an adversarial bandit subroutine. In light of the importance of a
 finite and compact $\tilde{\Pi}$, we propose a novel training-time algorithm to
 iteratively discover \textit{non-dominated policies}, forming a near-optimal an
d minimal $\tilde{\Pi}$, thereby ensuring both robustness and test-time efficien
cy. Empirical validation on the Mujoco corroborates the superiority of our appro
ach in terms of natural and robust performance, as well as adaptability to vario
us attack scenarios.
****

John Kirchenbauer,Jonas Geiping,Yuxin Wen,Manli Shu,Khalid Saifullah,Kezhi Kong, Kasun Fernando,Aniruddha Saha,Micah Goldblum,Tom Goldstein
On the Reliability of Watermarks for Large Language Models
As LLMs become commonplace, machine-generated text has the potential to flood the internet with spam, social media bots, and valueless content. _Watermarking_ is a simple and effective strategy for mitigating such harms by enabling the detection and documentation of LLM-generated text. Yet a crucial question remains: How reliable is watermarking in realistic settings in the wild? There, watermarked text may be modified to suit a user's needs, or entirely rewritten to avoid detection. We study the robustness of watermarked text after it is re-written by humans, paraphrased by a non-watermarked LLM, or mixed into a longer hand-written document. We find that watermarks remain detectable even after human and machine paraphrasing. While these attacks dilute the strength of the watermark, paraphrases are statistically likely to leak n-grams or even longer fragments of the original text, resulting in high-confidence detections when enough tokens are observed. For example, after strong human paraphrasing the watermark is detectable after observing 800 tokens on average, when setting a $1\mathrm{e}{-5}$ false positive rate. We also consider a range of new detection schemes that are sensitive to short spans of watermarked text embedded inside a large document, and we compare the robustness of watermarking to other kinds of detectors.
****
Jingxiang Sun,Bo Zhang,Ruizhi Shao,Lizhen Wang,Wen Liu,Zhenda Xie,Yebin Liu
DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior
We present DreamCraft3D, a hierarchical 3D content generation method that produces high-fidelity and coherent 3D objects. We tackle the problem by leveraging a 2D reference image to guide the stages of geometry sculpting and texture boosting. A central focus of this work is to address the consistency issue that existing works encounter. To sculpt geometries that render coherently, we perform score distillation sampling via a view-dependent diffusion model. This 3D prior, alongside several training strategies, prioritizes the geometry consistency but compromises the texture fidelity. We further propose bootstrapped score distillation to specifically boost the texture. We train a personalized diffusion model, Dreambooth, on the augmented renderings of the scene, imbuing it with 3D knowledge of the scene being optimized. The score distillation from this 3D-aware diffusion prior provides view-consistent guidance for the scene. Notably, through an alternating optimization of the diffusion prior and 3D scene representation, we achieve mutually reinforcing improvements: the optimized 3D scene aids in training the scene-specific diffusion model, which offers increasingly view-consistent guidance for 3D optimization. The optimization is thus bootstrapped and leads to substantial texture boosting. With tailored 3D priors throughout the hierarchical generation, DreamCraft3D generates coherent 3D objects with photorealistic renderings, advancing the state-of-the-art in 3D content generation.
****
Kuan Li,YiWen Chen,Yang Liu,Jin Wang,Qing He,Minhao Cheng,Xiang Ao
Boosting the Adversarial Robustness of Graph Neural Networks: An OOD Perspective
Current defenses against graph attacks often rely on certain properties to eliminate structural perturbations by identifying adversarial edges from normal edges. However, this dependence makes defenses vulnerable to adaptive (white-box) attacks from adversaries with the same knowledge. Adversarial training seems to be a feasible way to enhance robustness without reliance on artificially designed properties. However, in this paper, we show that it can lead to models learning incorrect information. To solve this issue, we re-examine graph attacks from the out-of-distribution (OOD) perspective for poisoning and evasion attacks and introduce a novel adversarial training paradigm incorporating OOD detection. This approach strengthens the robustness of Graph Neural Networks (GNNs) without reliance on prior knowledge. To further evaluate adaptive robustness, we develop adaptive attacks against our methods, revealing a trade-off between graph attack efficacy and defensibility. Through extensive experiments over 25,000 perturbed graphs, our method could still maintain good robustness against both adaptive and non-adaptive attacks. The code is provided at https://github.com/likuanppd/GOOD-AT

.
****
Harikrishna Narasimhan,Aditya Krishna Menon,Wittawat Jitkrittum,Sanjiv Kumar
Plugin estimators for selective classification with out-of-distribution detection
Real-world classifiers can benefit from the option of abstaining from predicting on samples where they have low confidence. Such abstention is particularly useful on samples which are close to the learned decision boundary, or which are outliers with respect to the training sample. These settings have been the subject of extensive but disjoint study in the selective classification (SC) and out-of-distribution (OOD) detection literature. Recent work on selective classification with OOD detection (SCOD) has argued for the unified study of these problems; however, the formal underpinnings of this problem are still nascent, and existing techniques are heuristic in nature. In this paper, we propose new plugin estimators for SCOD that are theoretically grounded, effective, and generalise existing approaches from the SC and OOD detection literature. In the course of our analysis, we formally explicate how naïve use of existing SC and OOD detection baselines may be inadequate for SCOD. We empirically demonstrate that our approaches yields competitive SC and OOD detection trade-offs compared to common baselines.
****
Mauricio Tec,Ana Trisovic,Michelle Audirac,Sophie Mirabai Woodward,Jie Kate Hu,Naeem Khoshnevis,Francesca Dominici
SpaCE: The Spatial Confounding Environment
Spatial confounding poses a significant challenge in scientific studies involving spatial data, where unobserved spatial variables can influence both treatment and outcome, possibly leading to spurious associations. To address this problem, we introduce SpaCE: The Spatial Confounding Environment, the first toolkit to provide realistic benchmark datasets and tools for systematically evaluating causal inference methods designed to alleviate spatial confounding. Each dataset includes training data, true counterfactuals, a spatial graph with coordinates, and smoothness and confounding scores characterizing the effect of a missing spatial confounder. It also includes realistic semi-synthetic outcomes and counterfactuals, generated using state-of-the-art machine learning ensembles, following best practices for causal inference benchmarks. The datasets cover real treatment and covariates from diverse domains, including climate, health and social sciences. SpaCE facilitates an automated end-to-end pipeline, simplifying data loading, experimental setup, and evaluating machine learning and causal inference models. The SpaCE project provides several dozens of datasets of diverse sizes and spatial complexity. It is publicly available as a Python package, encouraging community feedback and contributions.
****
Nico Daheim,Thomas Möllenhoff,Edoardo Ponti,Iryna Gurevych,Mohammad Emtiyaz Khan
Model Merging by Uncertainty-Based Gradient Matching
Models trained on different datasets can be merged by a weighted-averaging of their parameters, but why does it work and when can it fail? Here, we connect the inaccuracy of weighted-averaging to mismatches in the gradients and propose a new uncertainty-based scheme to improve the performance by reducing the mismatch. The connection also reveals implicit assumptions in other schemes such as averaging, task arithmetic, and Fisher-weighted averaging. Our new method gives consistent improvements for large language models and vision transformers, both in terms of performance and robustness to hyperparameters.
****
Yufeng Zhang,Hang Yu,Jianguo Li,Weiyao Lin
Finite-State Autoregressive Entropy Coding for Efficient Learned Lossless Compression
Learned lossless data compression has garnered significant attention recently due to its superior compression ratios compared to traditional compressors. However, the computational efficiency of these models jeopardizes their practicality. This paper proposes a novel system for improving the compression ratio while maintaining computational efficiency for learned lossless data compression. Our app

roach incorporates two essential innovations. First, we propose the Finite-State AutoRegressive (FSAR) entropy coder, an efficient autoregressive Markov model based entropy coder that utilizes a lookup table to expedite autoregressive entropy coding. Next, we present a Straight-Through Hardmax Quantization (STHQ) scheme to enhance the optimization of discrete latent space. Our experiments show that the proposed lossless compression method could improve the compression ratio by up to 6\% compared to the baseline, with negligible extra computational time. Our work provides valuable insights into enhancing the computational efficiency of learned lossless data compression, which can have practical applications in various fields. Code is available at https://github.com/alipay/Finite_State_Autoregressive_Entropy_Coding.
****
Quoc Phong Nguyen,Wan Theng Ruth Chew,Le Song,Bryan Kian Hsiang Low,Patrick Jaillet
Optimistic Bayesian Optimization with Unknown Constraints
Though some research efforts have been dedicated to constrained Bayesian optimization (BO), there remains a notable absence of a principled approach with a theoretical performance guarantee in the decoupled setting. Such a setting involves independent evaluations of the objective function and constraints at different inputs, and is hence a relaxation of the commonly-studied coupled setting where functions must be evaluated together. As a result, the decoupled setting requires an adaptive selection between evaluating either the objective function or a constraint, in addition to selecting an input (in the coupled setting). This paper presents a novel constrained BO algorithm with a provable performance guarantee that can address the above relaxed setting. Specifically, it considers the fundamental trade-off between exploration and exploitation in constrained BO, and, interestingly, affords a noteworthy connection to active learning. The performance of our proposed algorithms is also empirically evaluated using several synthetic and real-world optimization problems.
****
Chenyu Zhang,Han Wang,Aritra Mitra,James Anderson
Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning
Federated reinforcement learning (FRL) has emerged as a promising paradigm for reducing the sample complexity of reinforcement learning tasks by exploiting information from different agents. However, when each agent interacts with a potentially different environment, little to nothing is known theoretically about the non-asymptotic performance of FRL algorithms. The lack of such results can be attributed to various technical challenges and their intricate interplay: Markovian sampling, linear function approximation, multiple local updates to save communication, heterogeneity in the reward functions and transition kernels of the agents' MDPs, and continuous state-action spaces. Moreover, in the on-policy setting, the behavior policies vary with time, further complicating the analysis. In response, we introduce FedSARSA, a novel federated on-policy reinforcement learning scheme, equipped with linear function approximation, to address these challenges and provide a comprehensive finite-time error analysis. Notably, we establish that FedSARSA converges to a policy that is near-optimal for all agents, with the extent of near-optimality proportional to the level of heterogeneity. Furthermore, we prove that FedSARSA leverages agent collaboration to enable linear speedups as the number of agents increases, which holds for both fixed and adaptive step-size configurations.
****
Tongda Xu,Ziran Zhu,Dailan He,Yanghao Li,Lina Guo,Yuanyuan Wang,Zhe Wang,Hongwei Qin,Yan Wang,Jingjing Liu,Ya-Qin Zhang
Idempotence and Perceptual Image Compression
Idempotence is the stability of image codec to re-compression. At the first glance, it is unrelated to perceptual image compression. However, we find that theoretically: 1) Conditional generative model-based perceptual codec satisfies idempotence; 2) Unconditional generative model with idempotence constraint is equivalent to conditional generative codec. Based on this newfound equivalence, we propose a new paradigm of perceptual image codec by inverting unconditional generati

ve model with idempotence constraints. Our codec is theoretically equivalent to conditional generative codec, and it does not require training new models. Instead, it only requires a pre-trained mean-square-error codec and unconditional generative model. Empirically, we show that our proposed approach outperforms state-of-the-art methods such as HiFiC and ILLM, in terms of Fréchet Inception Distance (FID). The source code is provided in https://github.com/tongdaxu/Idempotence-and-Perceptual-Image-Compression.
****
Jeongyeol Kwon,Dohyun Kwon,Stephen Wright,Robert D Nowak
On Penalty Methods for Nonconvex Bilevel Optimization and First-Order Stochastic Approximation
In this work, we study first-order algorithms for solving Bilevel Optimization (BO) where the objective functions are smooth but possibly nonconvex in both levels and the variables are restricted to closed convex sets. As a first step, we study the landscape of BO through the lens of penalty methods, in which the upper- and lower-level objectives are combined in a weighted sum with penalty parameter $\sigma > 0$. In particular, we establish a strong connection between the penalty function and the hyper-objective by explicitly characterizing the conditions under which the values and derivatives of the two must be $O(\sigma)$-close. A by-product of our analysis is the explicit formula for the gradient of hyper-objective when the lower-level problem has multiple solutions under minimal conditions, which could be of independent interest. Next, viewing the penalty formulation as $O(\sigma)$-approximation of the original BO, we propose first-order algorithms that find an ϵ-stationary solution by optimizing the penalty formulation with $\sigma = O(\epsilon)$. When the perturbed lower-level problem uniformly satisfies the {\it small-error} proximal error-bound (EB) condition, we propose a first-order algorithm that converges to an ϵ-stationary point of the penalty function using in total $O(\epsilon^{-7})$ accesses to first-order stochastic gradient oracles. Under an additional assumption on stochastic oracles, we show that the algorithm can be implemented in a fully {\it single-loop} manner, {\it i.e.,} with $O(1)$ samples per iteration, and achieves the improved oracle-complexity of $O(\epsilon^{-5})$.
****
Guangyi Chen,Yuke Li,Xiao Liu,Zijian Li,Eman Al Suradi,Donglai Wei,Kun Zhang
LLCP: Learning Latent Causal Processes for Reasoning-based Video Question Answer
Current approaches to Video Question Answering (VideoQA) primarily focus on cross-modality matching, which is limited by the requirement for extensive data annotations and the insufficient capacity for causal reasoning (e.g. attributing accidents). To address these challenges, we introduce a causal framework for video reasoning, termed Learning Latent Causal Processes (LLCP). At the heart of LLCP lies a multivariate generative model designed to analyze the spatial-temporal dynamics of objects within events. Leveraging the inherent modularity of causal mechanisms, we train the model through self-supervised local auto-regression eliminating the need for annotated question-answer pairs. During inference, the model is applied to answer two types of reasoning questions: accident attribution, which infers the cause from observed effects, and counterfactual prediction, which predicts the effects of counterfactual conditions given the factual evidence. In the first scenario, we identify variables that deviate from the established distribution by the learned model, signifying the root cause of accidents. In the second scenario, we replace embeddings of previous variables with counterfactual ones, enabling us to forecast potential developments. Once we have identified these cause/effect variables, natural language answers are derived through a combination of grammatical parsing and a pre-trained vision-language model. We assess the efficacy of LLCP on both synthetic and real-world data, demonstrating comparable performance to supervised methods despite our framework using no paired textual annotations.
****
Amro Kamal Mohamed Abbas,Evgenia Rusak,Kushal Tirumala,Wieland Brendel,Kamalika Chaudhuri,Ari S. Morcos
Effective pruning of web-scale datasets based on complexity of concept clusters

Utilizing massive web-scale datasets has led to unprecedented performance gains in machine learning models, but also imposes outlandish compute requirements for their training. In order to improve training and data efficiency, we here push the limits of pruning large-scale multimodal datasets for training CLIP-style models. Today's most effective pruning method on ImageNet clusters data samples into separate concepts according to their embedding and prunes away the most proto- typical samples. We scale this approach to LAION and improve it by noting that the pruning rate should be concept-specific and adapted to the complexity of the concept. Using a simple and intuitive complexity measure, we are able to reduce the training cost to a quarter of regular training. More specifically, we are able to outperform the LAION-trained OpenCLIP-ViT-B/32 model on ImageNet zero-shot accuracy by 1.1p.p. while only using 27.7% of the data and training compute. On the DataComp Medium benchmark, we achieve a new state-of-the-art ImageNet zero-shot accuracy and a competitive average zero-shot accuracy on 38 evaluation tasks.
****
Qihan Liu,Jianing Ye,Xiaoteng Ma,Jun Yang,Bin Liang,Chongjie Zhang
Efficient Multi-agent Reinforcement Learning by Planning
Multi-agent reinforcement learning (MARL) algorithms have accomplished remarkable breakthroughs in solving large-scale decision-making tasks. Nonetheless, most existing MARL algorithms are model-free, limiting sample efficiency and hindering their applicability in more challenging scenarios. In contrast, model-based reinforcement learning (MBRL), particularly algorithms integrating planning, such as MuZero, has demonstrated superhuman performance with limited data in many tasks. Hence, we aim to boost the sample efficiency of MARL by adopting model-based approaches. However, incorporating planning and search methods into multi-agent systems poses significant challenges. The expansive action space of multi-agent systems often necessitates leveraging the nearly-independent property of agents to accelerate learning. To tackle this issue, we propose the MAZero algorithm, which combines a centralized model with Monte Carlo Tree Search (MCTS) for policy search. We design an ingenious network structure to facilitate distributed execution and parameter sharing. To enhance search efficiency in deterministic environments with sizable action spaces, we introduce two novel techniques: Optimistic Search Lambda (OS(λ)) and Advantage-Weighted Policy Optimization (AWPO). Extensive experiments on the SMAC benchmark demonstrate that MAZero outperforms model-free approaches in terms of sample efficiency and provides comparable or better performance than existing model-based methods in terms of both sample and computational efficiency.
****
Yuan-Hong Liao,David Acuna,Rafid Mahmood,James Lucas,Viraj Uday Prabhu,Sanja Fidler
Transferring Labels to Solve Annotation Mismatches Across Object Detection Datasets
In object detection, varying annotation protocols across datasets can result in annotation mismatches, leading to inconsistent class labels and bounding regions. Addressing these mismatches typically involves manually identifying common trends and fixing the corresponding bounding boxes and class labels. To alleviate this laborious process, we introduce the label transfer problem in object detection. Here, the goal is to transfer bounding boxes from one or more source datasets to match the annotation style of a target dataset. We propose a data-centric approach, Label-Guided Pseudo-Labeling (LGPL), that improves downstream detectors in a manner agnostic to the detector learning algorithms and model architectures. Validating across four object detection scenarios, defined over seven different datasets and three different architectures, we show that transferring labels for a target task via LGPL consistently improves the downstream detection in every setting, on average by 1.88 mAP and 2.65 AP^{75}. Most importantly, we find that when training with multiple labeled datasets, carefully addressing annotation mismatches with LGPL alone can improve downstream object detection better than off-the-shelf supervised domain adaptation techniques that align instance features.

****
Peter Richtárik,Elnur Gasanov,Konstantin Pavlovich Burlachenko
Error Feedback Reloaded: From Quadratic to Arithmetic Mean of Smoothness Constan ts
Error feedback (EF) is a highly popular and immensely effective mechanism for fi xing convergence issues which arise in distributed training methods (such as dis tributed GD or SGD) when these are enhanced with greedy communication compressio n techniques such as Top-k. While EF was proposed almost a decade ago (Seide et al, 2014), and despite concentrated effort by the community to advance the theor etical understanding of this mechanism, there is still a lot to explore. In this work we study a modern form of error feedback called EF21 (Richtárik et al, 202 1) which offers the currently best-known theoretical guarantees, under the weake st assumptions, and also works well in practice. In particular, while the theore tical communication complexity of EF21 depends on the quadratic mean of certain smoothness parameters, we improve this dependence to their arithmetic mean, whic h is always smaller, and can be substantially smaller, especially in heterogeneo us data regimes. We take the reader on a journey of our discovery process. Start ing with the idea of applying EF21 to an equivalent reformulation of the underly ing problem which (unfortunately) requires (often impractical) machine cloning, we continue to the discovery of a new weighted version of EF21 which can (fortun ately) be executed without any cloning, and finally circle back to an improved a nalysis of the original EF21 method. While this development applies to the simpl est form of EF21, our approach naturally extends to more elaborate variants invo lving stochastic gradients and partial participation. Further, our technique imp roves the best-known theory of EF21 in the rare features regime (Richtárik et al , 2023). Finally, we validate our theoretical findings with suitable experiments .
****
Adam Block,Dylan J Foster,Akshay Krishnamurthy,Max Simchowitz,Cyril Zhang
Butterfly Effects of SGD Noise: Error Amplification in Behavior Cloning and Auto regression
This work studies training instabilities of behavior cloning with deep neural ne tworks. We observe that minibatch SGD updates to the policy network during train ing result in sharp oscillations in long-horizon rewards, despite negligibly aff ecting the behavior cloning loss. We empirically disentangle the statistical and computational causes of these oscillations, and find them to stem from the chao tic propagation of minibatch SGD noise through unstable closed-loop dynamics. W hile SGD noise is benign in the single-step action prediction objective, it resu lts in catastrophic error accumulation over long horizons, an effect we term *gr adient variance amplification* (GVA). We demonstrate that many standard mitigat ion techniques do not alleviate GVA, but that taking an exponential moving avera ge (EMA) of iterates is surprisingly effective at doing so. Furthermore, we ill ustrate the generality of the phenomenon by showing both the existence of GVA an d its amelioration by EMA in autoregressive language generation. Finally, we pr ovide theoretical vignettes both exhibiting the benefits of EMA in alleviating G VA and illustrating the extent to which classical convex models help in understa nding the benefits of iterate averaging in deep learning.
****
Can Xu,Qingfeng Sun,Kai Zheng,Xiubo Geng,Pu Zhao,Jiazhan Feng,Chongyang Tao,Qing wei Lin,Daxin Jiang
WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instruc tions
Training large language models (LLMs) with open-domain instruction following dat a brings colossal success. However, manually creating such instruction data is v ery time-consuming and labor-intensive. Moreover, humans may struggle to produce high-complexity instructions. In this paper, we show an avenue for creating lar ge amounts of instruction data with varying levels of complexity using LLM inste ad of humans. Starting with an initial set of instructions, we use our proposed Evol-Instruct to rewrite them step by step into more complex instructions. Then, we mix all generated instruction data to fine-tune LLaMA. We call the resulting

model WizardLM. Both automatic and human evaluations consistently indicate that WizardLM outperforms baselines such as Alpaca (trained from Self-Instruct) and Vicuna (trained from human-created instructions). The experimental results demon strate that the quality of instruction-following dataset crafted by Evol-Instruc t can significantly improve the performance of LLMs.
****
Jinsung Jeon,Hyundong Jin,Jonghyun Choi,Sanghyun Hong,Dongeun Lee,Kookjin Lee,No seong Park
PAC-FNO: Parallel-Structured All-Component Fourier Neural Operators for Recogniz ing Low-Quality Images
A standard practice in developing image recognition models is to train a model o n a specific image resolution and then deploy it. However, in real-world inferen ce, models often encounter images different from the training sets in resolution and/or subject to natural variations such as weather changes, noise types and c ompression artifacts. While traditional solutions involve training multiple mode ls for different resolutions or input variations, these methods are computationa lly expensive and thus do not scale in practice. To this end, we propose a novel neural network model, parallel-structured and all-component Fourier neural oper ator (PAC-FNO), that addresses the problem. Unlike conventional feed-forward neu ral networks, PAC-FNO operates in the frequency domain, allowing it to handle im ages of varying resolutions within a single model. We also propose a two-stage a lgorithm for training PAC-FNO with a minimal modification to the original, downs tream model. Moreover, the proposed PAC-FNO is ready to work with existing image recognition models. Extensively evaluating methods with seven image recognition benchmarks, we show that the proposed PAC-FNO improves the performance of exist ing baseline models on images with various resolutions by up to 77.1% and variou s types of natural variations in the images at inference.
****
Hongbin Huang,Minghua Chen,Xiao Qiao
Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns
Limited data availability poses a major obstacle in training deep learning model s for financial applications. Synthesizing financial time series to augment real -world data is challenging due to the irregular and scale-invariant patterns uni quely associated with financial time series - temporal dynamics that repeat with varying duration and magnitude. Such dynamics cannot be captured by existing ap proaches, which often assume regularity and uniformity in the underlying data. W e develop a novel generative framework called FTS-Diffusion to model irregular a nd scale-invariant patterns that consists of three modules. First, we develop a scale-invariant pattern recognition algorithm to extract recurring patterns that vary in duration and magnitude. Second, we construct a diffusion-based generati ve network to synthesize segments of patterns. Third, we model the temporal tran sition of patterns in order to aggregate the generated segments. Extensive exper iments show that FTS-Diffusion generates synthetic financial time series highly resembling observed data, outperforming state-of-the-art alternatives. Two downs tream experiments demonstrate that augmenting real-world data with synthetic dat a generated by FTS-Diffusion reduces the error of stock market prediction by up to 17.9%. To the best of our knowledge, this is the first work on generating int ricate time series with irregular and scale-invariant patterns, addressing data limitation issues in finance.
****
Yun-Hin Chan,Rui Zhou,Running Zhao,Zhihan JIANG,Edith C. H. Ngai
Internal Cross-layer Gradients for Extending Homogeneity to Heterogeneity in Fed erated Learning
Federated learning (FL) inevitably confronts the challenge of system heterogenei ty in practical scenarios. To enhance the capabilities of most model-homogeneous FL methods in handling system heterogeneity, we propose a training scheme that can extend their capabilities to cope with this challenge. In this paper, we com mence our study with a detailed exploration of homogeneous and heterogeneous FL settings and discover three key observations: (1) a positive correlation between

client performance and layer similarities, (2) higher similarities in the shallow layers in contrast to the deep layers, and (3) the smoother gradients distributions indicate the higher layer similarities. Building upon these observations, we propose InCo Aggregation that leverages internal cross-layer gradients, a mixture of gradients from shallow and deep layers within a server model, to augment the similarity in the deep layers without requiring additional communication between clients. Furthermore, our methods can be tailored to accommodate model-homogeneous FL methods such as FedAvg, FedProx, FedNova, Scaffold, and MOON, to expand their capabilities to handle the system heterogeneity. Copious experimental results validate the effectiveness of InCo Aggregation, spotlighting internal cross-layer gradients as a promising avenue to enhance the performance in heterogeneous FL.
****
Jingyan Chen,Guanghui Zhu,Chunfeng Yuan,Yihua Huang
Boosting Graph Anomaly Detection with Adaptive Message Passing
Unsupervised graph anomaly detection has been widely used in real-world applications. Existing methods primarily focus on local inconsistency mining (LIM), based on the intuition that establishing high similarities between abnormal nodes and their neighbors is difficult. However, the message passing employed by graph neural networks (GNNs) results in local anomaly signal loss, as GNNs tend to make connected nodes similar, which conflicts with the LIM intuition. In this paper, we propose GADAM, a novel framework that not only resolves the conflict between LIM and message passing but also leverages message passing to augment anomaly detection through a transformative approach to anomaly mining beyond LIM. Specifically, we first propose an efficient MLP-based LIM approach to obtain local anomaly scores in a conflict-free way. Next, we introduce a novel approach to capture anomaly signals from a global perspective. This involves a hybrid attention based adaptive message passing, enabling nodes to selectively absorb abnormal or normal signals from their surroundings. Extensive experiments conducted on nine benchmark datasets, including two large-scale OGB datasets, demonstrate that GADAM surpassinges existing state-of-the-art methods in terms of both effectiveness and efficiency.
****
Xinyao Fan,Yueying Wu,Chang Xu,Yuhao Huang,Weiqing Liu,Jiang Bian
MG-TSD: Multi-Granularity Time Series Diffusion Models with Guided Learning Process
Recently, diffusion probabilistic models have attracted attention in generative time series forecasting due to their remarkable capacity to generate high-fidelity samples. However, the effective utilization of their strong modeling ability in the probabilistic time series forecasting task remains an open question, partially due to the challenge of instability arising from their stochastic nature. To address this challenge, we introduce a novel Multi-Granularity Time Series Diffusion (MG-TSD) model, which achieves state-of-the-art predictive performance by leveraging the inherent granularity levels within the data as given targets at intermediate diffusion steps to guide the learning process of diffusion models. The way to construct the targets is motivated by the observation that forward process of the diffusion model, which sequentially corrupts the data distribution to a standard normal distribution, intuitively aligns with the process of smoothing fine-grained data into a coarse-grained representation, both of which result in a gradual loss of fine distribution features. In the study, we derive a novel multi-granularity guidance diffusion loss function and propose a concise implementation method to effectively utilize coarse-grained data across various granularity levels.
More importantly, our approach does not rely on additional external data, making it versatile and applicable across various domains. Extensive experiments conducted on real-world datasets demonstrate that our MG-TSD model outperforms existing time series prediction methods.
****
Seonghyeon Ye,Doyoung Kim,Sungdong Kim,Hyeonbin Hwang,Seungone Kim,Yongrae Jo,James Thorne,Juho Kim,Minjoon Seo

FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets
Evaluation of Large Language Models (LLMs) is challenging because instruction-following necessitates alignment with human values and the required set of skills varies depending on the instruction. However, previous studies have mainly focused on coarse-grained evaluation (i.e. overall preference-based evaluation), which limits interpretability since it does not consider the nature of user instructions that require instance-wise skill composition. In this paper, we introduce FLASK (Fine-grained Language Model Evaluation based on Alignment Skill Sets), a fine-grained evaluation protocol for both human-based and model-based evaluation which decomposes coarse-level scoring to a skill set-level scoring for each instruction. We experimentally observe that the fine-graininess of evaluation is crucial for attaining a holistic view of model performance and increasing the reliability of the evaluation. Using FLASK, we compare multiple open-source and proprietary LLMs and observe a high correlation between model-based and human-based evaluations.
****
Haoran Deng,Yang Yang,Jiahe Li,Cheng Chen,Weihao Jiang,Shiliang Pu
Fast Updating Truncated SVD for Representation Learning with Sparse Matrices
Updating truncated Singular Value Decomposition (SVD) has extensive applications in representation learning.
The continuous evolution of massive-scaled data matrices in practical scenarios highlights the importance of aligning SVD-based models with fast-paced updates.
Recent methods for updating truncated SVD can be recognized as Rayleigh-Ritz projection procedures where their projection matrices are augmented based on the original singular vectors.
However, the updating process in these methods densifies the update matrix and applies the projection to all singular vectors, resulting in inefficiency.
This paper presents a novel method for dynamically approximating the truncated SVD of a sparse and temporally evolving matrix.
The proposed method takes advantage of sparsity in the orthogonalization process of the augment matrices and employs an extended decomposition to store projections in the column space of singular vectors independently.
Numerical experimental results on updating truncated SVD for evolving sparse matrices show an order of magnitude improvement in the efficiency of our proposed method while maintaining precision comparing to previous methods.
****
Xuan Li,Zhanke Zhou,Jiangchao Yao,Yu Rong,Lu Zhang,Bo Han
Neural Atoms: Propagating Long-range Interaction in Molecular Graphs through Efficient Communication Channel
Graph Neural Networks (GNNs) have been widely adopted for drug discovery with molecular graphs. Nevertheless, current GNNs mainly excel in leveraging short-range interactions (SRI) but struggle to capture long-range interactions (LRI), both of which are crucial for determining molecular properties. To tackle this issue, we propose a method to abstract the collective information of atomic groups into a few $\textit{Neural Atoms}$ by implicitly projecting the atoms of a molecular.
Specifically, we explicitly exchange the information among neural atoms and project them back to the atoms’ representations as an enhancement. With this mechanism, neural atoms establish the communication channels among distant nodes, effectively reducing the interaction scope of arbitrary node pairs into a single hop.

To provide an inspection of our method from a physical perspective, we reveal its connection to the traditional LRI calculation method, Ewald Summation. The Neural Atom can enhance GNNs to capture LRI by approximating the potential LRI of the molecular.
We conduct extensive experiments on four long-range graph benchmarks, covering graph-level and link-level tasks on molecular graphs. We achieve up to a 27.32% and 38.27% improvement in the 2D and 3D scenarios, respectively.
Empirically, our method can be equipped with an arbitrary GNN to help capture LRI. Code and datasets are publicly available in https://github.com/tmlr-group/Neu

ralAtom.
****
Thomas Tian,Chenfeng Xu,Masayoshi Tomizuka,Jitendra Malik,Andrea Bajcsy
What Matters to You? Towards Visual Representation Alignment for Robot Learning
When operating in service of people, robots need to optimize rewards aligned with end-user preferences. Since robots will rely on raw perceptual inputs, their rewards will inevitably use visual representations. Recently there has been excitement in using representations from pre-trained visual models, but key to making these work in robotics is fine-tuning, which is typically done via proxy tasks like dynamics prediction or enforcing temporal cycle-consistency. However, all these proxy tasks bypass the human's input on what matters to them, exacerbating spurious correlations and ultimately leading to behaviors that are misaligned with user preferences. In this work, we propose that robots should leverage human feedback to align their visual representations with the end-user and disentangle what matters for the task. We propose Representation-Aligned Preference-based Learning (RAPL), a method for solving the visual representation alignment problem and visual reward learning problem through the lens of preference-based learning and optimal transport. Across experiments in X MAGICAL and in robotic manipulation, we find that RAPL's reward consistently generates preferred robot behaviors with high sample efficiency, and shows strong zero-shot generalization when the visual representation is learned from a different embodiment than the robot's.
****
Qi Yan,Raihan Seraj,Jiawei He,Lili Meng,Tristan Sylvain
AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval
Machine-based prediction of real-world events is garnering attention due to its potential for informed decision-making. Whereas traditional forecasting predominantly hinges on structured data like time-series, recent breakthroughs in language models enable predictions using unstructured text. In particular, (Zou et al., 2022) unveils AutoCast, a new benchmark that employs news articles for answering forecasting queries. Nevertheless, existing methods still trail behind human performance. The cornerstone of accurate forecasting, we argue, lies in identifying a concise, yet rich subset of news snippets from a vast corpus. With this motivation, we introduce AutoCast++, a zero-shot ranking-based context retrieval system, tailored to sift through expansive news document collections for event forecasting. Our approach first re-ranks articles based on zero-shot question-passage relevance, honing in on semantically pertinent news. Following this, the chosen articles are subjected to zero-shot summarization to attain succinct context. Leveraging a pre-trained language model, we conduct both the relevance evaluation and article summarization without needing domain-specific training. Notably, recent articles can sometimes be at odds with preceding ones due to new facts or unanticipated incidents, leading to fluctuating temporal dynamics. To tackle this, our re-ranking mechanism gives preference to more recent articles, and we further regularize the multi-passage representation learning to align with human forecaster responses made on different dates. Empirical results underscore marked improvements across multiple metrics, improving the performance for multiple-choice questions (MCQ) by 48% and true/false (TF) questions by up to 8%.
****
Yuyao Zhang,Lan Wei,Nikolaos Freris
Synergistic Patch Pruning for Vision Transformer: Unifying Intra- & Inter-Layer Patch Importance
The Vision Transformer (ViT) has emerged as a powerful architecture for various computer vision tasks. Nonetheless, this comes with substantially heavier computational costs than Convolutional Neural Networks (CNNs). The attention mechanism in ViTs, which integrates information from different image patches to the class token ([CLS]), renders traditional structured pruning methods used in CNNs unsuitable. To overcome this issue, we propose SynergisTic pAtch pRuning (STAR) that unifies intra-layer and inter-layer patch importance scoring. Specifically, our approach combines a) online evaluation of intra-layer importance for the [CLS] and b) offline evaluation of the inter-layer importance of each patch. The two i

mportance scores are fused by minimizing a weighted average of Kullback-Leibler (KL) Divergences and patches are successively pruned at each layer by maintainin g only the top-k most important ones. Unlike prior art that relies on manual sel ection of the pruning rates at each layer, we propose an automated method for se lecting them based on offline-derived metrics. We also propose a variant that us es these rates as weighted percentile parameters (for the layer-wise normalized scores), thus leading to an alternate adaptive rate selection technique that is input-based. Extensive experiments demonstrate the significant acceleration of t he inference with minimal performance degradation. For instance, on the ImageNet dataset, the pruned DeiT-Small reaches a throughput of 4,256 images/s, which is over 66\% higher than the much smaller (unpruned) DeiT-Tiny model, while having a substantially higher accuracy (+6.8\% Top-1 and +3.1\% Top-5).
****
Moritz Hardt,Yu Sun
Test-Time Training on Nearest Neighbors for Large Language Models
Many recent efforts augment language models with retrieval, by adding retrieved data to the input context. For this approach to succeed, the retrieved data must be added at both training and test time. Moreover, as input length grows linear ly with the size of retrieved data, cost in computation and memory grows quadrat ically for modern Transformers. To avoid these complications, we simply fine-tun e the model on retrieved data at test time, using its standard training setup. W e build a large-scale distributed index based on text embeddings of the Pile dat aset. For each test input, our system retrieves its neighbors and fine-tunes the model on their text. Surprisingly, retrieving and training on as few as 20 neig hbors, each for only one gradient iteration, drastically improves performance ac ross more than 20 language modeling tasks in the Pile. For example, test-time tr aining with nearest neighbors significantly narrows the performance gap between a small GPT-2 and a GPT-Neo model more than 10 times larger. Sufficient index qu ality and size, however, are necessary. Our work establishes a first baseline of test-time training for language modeling.
****
Suqin Yuan,Lei Feng,Tongliang Liu
Early Stopping Against Label Noise Without Validation Data
Early stopping methods in deep learning face the challenge of balancing the volu me of training and validation data, especially in the presence of label noise. C oncretely, sparing more data for validation from training data would limit the p erformance of the learned model, yet insufficient validation data could result i n a sub-optimal selection of the desired model. In this paper, we propose a nove l early stopping method called Label Wave, which does not require validation dat a for selecting the desired model in the presence of label noise. It works by tr acking the changes in the model's predictions on the training set during the tra ining process, aiming to halt training before the model unduly fits mislabeled d ata. This method is empirically supported by our observation that minimum fluctu ations in predictions typically occur at the training epoch before the model exc essively fits mislabeled data. Through extensive experiments, we show both the e ffectiveness of the Label Wave method across various settings and its capability to enhance the performance of existing methods for learning with noisy labels.
****
Yujia Bao,Srinivasan Sivanandan,Theofanis Karaletsos
Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words
Vision Transformer (ViT) has emerged as a powerful architecture in the realm of modern computer vision. However, its application in certain imaging fields, such as microscopy and satellite imaging, presents unique challenges. In these domai ns, images often contain multiple channels, each carrying semantically distinct and independent information. Furthermore, the model must demonstrate robustness to sparsity in input channels, as they may not be densely available during train ing or testing. In this paper, we propose a modification to the ViT architecture that enhances reasoning across the input channels and introduce Hierarchical Ch annel Sampling (HCS) as an additional regularization technique to ensure robustn ess when only partial channels are presented during test time. Our proposed mode

l, ChannelViT, constructs patch tokens independently from each input channel and utilizes a learnable channel embedding that is added to the patch tokens, similar to positional embeddings. We evaluate the performance of ChannelViT on ImageNet, JUMP-CP (microscopy cell imaging), and So2Sat (satellite imaging). Our results show that ChannelViT outperforms ViT on classification tasks and generalizes well, even when a subset of input channels is used during testing. Across our experiments, HCS proves to be a powerful regularizer, independent of the architecture employed, suggesting itself as a straightforward technique for robust ViT training. Lastly, we find that ChannelViT generalizes effectively even when there is limited access to all channels during training, highlighting its potential for multi-channel imaging under real-world conditions with sparse sensors. Our code is available at https://github.com/insitro/ChannelViT.
****
Rui Pan,Yuxing Liu,Xiaoyu Wang,Tong Zhang
Accelerated Convergence of Stochastic Heavy Ball Method under Anisotropic Gradient Noise
Heavy-ball momentum with decaying learning rates is widely used with SGD for optimizing deep learning models. In contrast to its empirical popularity, the understanding of its theoretical property is still quite limited, especially under the standard anisotropic gradient noise condition for quadratic regression problems. Although it is widely conjectured that heavy-ball momentum method can provide accelerated convergence and should work well in large batch settings, there is no rigorous theoretical analysis. In this paper, we fill this theoretical gap by establishing a non-asymptotic convergence bound for stochastic heavy-ball methods with step decay scheduler on quadratic objectives, under the anisotropic gradient noise condition. As a direct implication, we show that heavy-ball momentum can provide $\tilde{\mathcal{O}}(\sqrt{\kappa})$ accelerated convergence of the bias term of SGD while still achieving near-optimal convergence rate with respect to the stochastic variance term. The combined effect implies an overall convergence rate within log factors from the statistical minimax rate. This means SGD with heavy-ball momentum is useful in the large-batch settings such as distributed machine learning or federated learning, where a smaller number of iterations can significantly reduce the number of communication rounds, leading to acceleration in practice.
****
Yaoyu Zhu,Jianhao Ding,Tiejun Huang,Xiaodong Xie,Zhaofei Yu
Online Stabilization of Spiking Neural Networks
Spiking neural networks (SNNs), attributed to the binary, event-driven nature of spikes, possess heightened biological plausibility and enhanced energy efficiency on neuromorphic hardware compared to analog neural networks (ANNs). Mainstream SNN training schemes apply backpropagation-through-time (BPTT) with surrogate gradients to replace the non-differentiable spike emitting process during backpropagation. While achieving competitive performance, the requirement for storing intermediate information at all time-steps incurs higher memory consumption and fails to fulfill the online property crucial to biological brains.
Our work focuses on online training techniques, aiming for memory efficiency while preserving biological plausibility.
The limitation of not having access to future information in early time steps in online training has constrained previous efforts to incorporate advantageous modules such as batch normalization.
To address this problem, we propose Online Spiking Renormalization (OSR) to ensure consistent parameters between testing and training, and Online Threshold Stabilizer (OTS) to stabilize neuron firing rates across time steps. Furthermore, we design a novel online approach to compute the sample mean and variance over time for OSR. Experiments conducted on various datasets demonstrate the proposed method's superior performance among SNN online training algorithms.
Our code is available at https://github.com/zhuyaoyu/SNN-online-normalization.
****
Ru Peng,Heming Zou,Haobo Wang,Yawen Zeng,Zenan Huang,Junbo Zhao
Energy-based Automated Model Evaluation

The conventional evaluation protocols on machine learning models rely heavily on a labeled, i.i.d-assumed testing dataset, which is not often present in real-world applications.
The Automated Model Evaluation (AutoEval) shows an alternative to this traditional workflow, by forming a proximal prediction pipeline of the testing performance without the presence of ground-truth labels.
Despite its recent successes, the AutoEval frameworks still suffer from an overconfidence issue, substantial storage and computational cost.
In that regard, we propose a novel measure --- Meta-Distribution Energy (MDE) that allows the AutoEval framework to be both more efficient and effective.
The core of the MDE is to establish a meta-distribution statistic, on the information (energy) associated with individual samples, then offer a smoother representation enabled by energy-based learning.
We further provide our theoretical insights by connecting the MDE with the classification loss.
We provide extensive experiments across modalities, datasets and different architectural backbones to validate MDE's validity, together with its superiority compared with prior approaches.
We also prove MDE's versatility by showing its seamless integration with large-scale models, and easy adaption to learning scenarios with noisy- or imbalanced-labels.
****
Wenlong Zhang,Xiaohui Li,Xiangyu Chen,Xiaoyun Zhang,Yu Qiao,Xiao-Ming Wu,Chao Dong
SEAL: A Framework for Systematic Evaluation of Real-World Super-Resolution
Real-world Super-Resolution (Real-SR) methods focus on dealing with diverse real-world images and have attracted increasing attention in recent years. The key idea is to use a complex and high-order degradation model to mimic real-world degradations.
Although they have achieved impressive results in various scenarios, they are faced with the obstacle of evaluation. Currently, these methods are only assessed by their average performance on a small set of degradation cases randomly selected from a large space, which fails to provide a comprehensive understanding of their overall performance and often yields inconsistent and potentially misleading results.
To overcome the limitation in evaluation, we propose SEAL, a framework for systematic evaluation of real-SR. In particular, we cluster the extensive degradation space to create a set of representative degradation cases, which serves as a comprehensive test set. Next, we propose a coarse-to-fine evaluation protocol to measure the distributed and relative performance of real-SR methods on the test set. The protocol incorporates two new metrics: acceptance rate (AR) and relative performance ratio (RPR), derived from acceptance and excellence lines. Under SEAL, we benchmark existing real-SR methods, obtain new observations and insights into their performance, and develop a new strong baseline. We consider SEAL as the first step towards creating an unbiased and comprehensive real-SR evaluation platform, which can promote the development of real-SR.
****
Peter West,Ximing Lu,Nouha Dziri,Faeze Brahman,Linjie Li,Jena D. Hwang,Liwei Jiang,Jillian Fisher,Abhilasha Ravichander,Khyathi Chandu,Benjamin Newman,Pang Wei Koh,Allyson Ettinger,Yejin Choi
The Generative AI Paradox: “What It Can Create, It May Not Understand”
The recent wave of generative AI has sparked unprecedented global attention, with both excitement and concern over potentially superhuman levels of artificial intelligence: models now take only seconds to produce outputs that would challenge or exceed the capabilities even of expert humans. At the same time, models still show basic errors in understanding that would not be expected even in non-expert humans. This presents us with an apparent paradox: how do we reconcile seemingly superhuman capabilities with the persistence of errors that few humans would make? In this work, we posit that this tension reflects a divergence in the configuration of intelligence in today's generative models relative to intelligenc

e in humans. Specifically, we propose and test the **Generative AI Paradox** hypothesis: generative models, having been trained directly to reproduce expert-like outputs, acquire generative capabilities that are not contingent upon---and can therefore exceed---their ability to understand those same types of outputs. This contrasts with humans, for whom basic understanding almost always precedes the ability to
generate expert-level outputs. We test this hypothesis through controlled experiments analyzing generation vs.~understanding in generative models, across both language and image modalities. Our results show that although models can outperform humans in generation, they consistently fall short of human capabilities in measures of understanding, as well as weaker correlation between generation and understanding performance, and more brittleness to adversarial inputs. Our findings support the hypothesis that models' generative capability may not be contingent upon understanding capability, and call for caution in interpreting artificial intelligence by analogy to human intelligence.
****
Yufei Gu,Xiaoqing Zheng,Tomaso Aste
Unraveling the Enigma of Double Descent: An In-depth Analysis through the Lens of Learned Feature Space
Double descent presents a counter-intuitive aspect within the machine learning domain, and researchers have observed its manifestation in various models and tasks. While some theoretical explanations have been proposed for this phenomenon in specific contexts, an accepted theory for its occurring mechanism in deep learning remains yet to be established. In this study, we revisit the phenomenon of double descent and demonstrate that the presence of noisy data strongly influences its occurrence. By comprehensively analysing the feature space of learned representations, we unveil that double descent arises in imperfect models trained with noisy data. We argue that while small and intermediate models before the interpolation threshold follow the traditional bias-variance trade-off, over-parameterized models interpolate noisy samples among robust data thus acquiring the capability to separate the information from the noise. The source code is available at \url{https://github.com/Yufei-Gu-451/double_descent_inference.git}.
****
Giovanni De Felice,Andrea Cini,Daniele Zambon,Vladimir Gusev,Cesare Alippi
Graph-based Virtual Sensing from Sparse and Partial Multivariate Observations
Virtual sensing techniques allow for inferring signals at new unmonitored locations by exploiting spatio-temporal measurements coming from physical sensors at different locations. However, as the sensor coverage becomes sparse due to costs or other constraints, physical proximity cannot be used to support interpolation. In this paper, we overcome this challenge by leveraging dependencies between the target variable and a set of correlated variables (covariates) that can frequently be associated with each location of interest. From this viewpoint, covariates provide partial observability, and the problem consists of inferring values for unobserved channels by exploiting observations at other locations to learn how such variables can correlate. We introduce a novel graph-based methodology to exploit such relationships and design a graph deep learning architecture, named GgNet, implementing the framework. The proposed approach relies on propagating information over a nested graph structure that is used to learn dependencies between variables as well as locations. GgNet is extensively evaluated under different virtual sensing scenarios, demonstrating higher reconstruction accuracy compared to the state-of-the-art.
****
Ziheng Qin,Kai Wang,Zangwei Zheng,Jianyang Gu,Xiangyu Peng,xu Zhao Pan,Daquan Zhou,Lei Shang,Baigui Sun,Xuansong Xie,Yang You
InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning
Data pruning aims to obtain lossless performances with less overall cost. A common approach is to filter out samples that make less contribution to the training. This could lead to gradient expectation bias compared to the original data. To solve this problem, we propose InfoBatch, a novel framework aiming to achieve lossless training acceleration by unbiased dynamic data pruning. Specifically, In

foBatch
randomly prunes a portion of less informative samples based on the loss distribu tion and rescales the gradients of the remaining samples to approximate the orig inal gradient. As a plug-and-play and architecture-agnostic framework, InfoBatch consistently obtains lossless training results on classification, semantic segm entation, vision pertaining, and instruction fine-tuning tasks. On CIFAR10/100, ImageNet-
1K, and ADE20K, InfoBatch losslessly saves 40% overall cost. For pertaining MAE and diffusion model, InfoBatch can respectively save 24.8% and 27% cost. For LLa MA instruction fine-tuning, combining InfoBatch and the recent coreset selection method (DQ) can achieve 10 times acceleration. Our results encourage more explo ration on the data efficiency aspect of large model training. Code is publicly a vailable at NUS-HPC-AI-Lab/InfoBatch.
****
Clément Bonnet,Daniel Luo,Donal John Byrne,Shikha Surana,Sasha Abramowitz,Paul D uckworth,Vincent Coyette,Laurence Illing Midgley,Elshadai Tegegn,Tristan Kalloni atis,Omayma Mahjoub,Matthew Macfarlane,Andries Petrus Smit,Nathan Grinsztajn,Rap hael Boige,Cemlyn Neil Waters,Mohamed Ali Ali Mimouni,Ulrich Armel Mbou Sob,Ruan John de Kock,Siddarth Singh,Daniel Furelos-Blanco,Victor Le,Arnu Pretorius,Alex andre Laterre
Jumanji: a Diverse Suite of Scalable Reinforcement Learning Environments in JAX
Open-source reinforcement learning (RL) environments have played a crucial role in driving progress in the development of AI algorithms.
In modern RL research, there is a need for simulated environments that are perfo rmant, scalable, and modular to enable their utilization in a wider range of pot ential real-world applications.
Therefore, we present Jumanji, a suite of diverse RL environments specifically d esigned to be fast, flexible, and scalable.
Jumanji provides a suite of environments focusing on combinatorial problems freq uently encountered in industry, as well as challenging general decision-making t asks.
By leveraging the efficiency of JAX and hardware accelerators like GPUs and TPUs , Jumanji enables rapid iteration of research ideas and large-scale experimentat ion, ultimately empowering more capable agents.
Unlike existing RL environment suites, Jumanji is highly customizable, allowing users to tailor the initial state distribution and problem complexity to their n eeds.
Furthermore, we provide actor-critic baselines for each environment, accompanied by preliminary findings on scaling and generalization scenarios.
Jumanji aims to set a new standard for speed, adaptability, and scalability of R L environments.
****
Jiarui Lu,Bozitao Zhong,Zuobai Zhang,Jian Tang
Str2Str: A Score-based Framework for Zero-shot Protein Conformation Sampling
The dynamic nature of proteins is crucial for determining their biological funct ions and properties, for which Monte Carlo (MC) and molecular dynamics (MD) simu lations stand as predominant tools to study such phenomena. By utilizing empiric ally derived force fields, MC or MD simulations explore the conformational space through numerically evolving the system via Markov chain or Newtonian mechanics . However, the high-energy barrier of the force fields can hamper the exploratio n of both methods by the rare event, resulting in inadequately sampled ensemble without exhaustive running. Existing learning-based approaches perform direct sa mpling yet heavily rely on target-specific simulation data for training, which s uffers from high data acquisition cost and poor generalizability. Inspired by si mulated annealing, we propose Str2Str, a novel structure-to-structure translatio n framework capable of zero-shot conformation sampling with roto-translation equ ivariant property. Our method leverages an amortized denoising score matching ob jective trained on general crystal structures and has no reliance on simulation data during both training and inference. Experimental results across several ben chmarking protein systems demonstrate that Str2Str outperforms previous state-of

-the-art generative structure prediction models and can be orders of magnitude faster compared with long MD simulations.
****

Blaise Delattre,Alexandre Araujo,Quentin Barthélemy,Alexandre Allauzen
The Lipschitz-Variance-Margin Tradeoff for Enhanced Randomized Smoothing
Real-life applications of deep neural networks are hindered by their unsteady predictions when faced with noisy inputs and adversarial attacks. The certified radius in this context is a crucial indicator of the robustness of models. However how to design an efficient classifier with an associated certified radius? Randomized smoothing provides a promising framework by relying on noise injection into the inputs to obtain a smoothed and robust classifier. In this paper, we first show that the variance introduced by the Monte-Carlo sampling in the randomized smoothing procedure estimate closely interacts with two other important properties of the classifier, \textit{i.e.} its Lipschitz constant and margin. More precisely, our work emphasizes the dual impact of the Lipschitz constant of the base classifier, on both the smoothed classifier and the empirical variance. To increase the certified robust radius, we introduce a different way to convert logits to probability vectors for the base classifier to leverage the variance-margin trade-off. We leverage the use of Bernstein's concentration inequality along with enhanced Lipschitz bounds for randomized smoothing. Experimental results show a significant improvement in certified accuracy compared to current state-of-the-art methods. Our novel certification procedure allows us to use pre-trained models with randomized smoothing, effectively improving the current certification radius in a zero-shot manner.
****

Song Xia,Yi Yu,Xudong Jiang,Henghui Ding
Mitigating the Curse of Dimensionality for Certified Robustness via Dual Randomized Smoothing
Randomized Smoothing (RS) has been proven a promising method for endowing an arbitrary image classifier with certified robustness. However, the substantial uncertainty inherent in the high-dimensional isotropic Gaussian noise imposes the curse of dimensionality on RS. Specifically, the upper bound of ${\ell_2}$ certified robustness radius provided by RS exhibits a diminishing trend with the expansion of the input dimension d, proportionally decreasing at a rate of $1/\sqrt{d}$. This paper explores the feasibility of providing ${\ell_2}$ certified robustness for high-dimensional input through the utilization of dual smoothing in the lower-dimensional space. The proposed Dual Randomized Smoothing (DRS) down-samples the input image into two sub-images and smooths the two sub-images in lower dimensions. Theoretically, we prove that DRS guarantees a tight ${\ell_2}$ certified robustness radius for the original input and reveal that DRS attains a superior upper bound on the ${\ell_2}$ robustness radius, which decreases proportionally at a rate of $(1/\sqrt m + 1/\sqrt n)$ with $m+n=d$. Extensive experiments demonstrate the generalizability and effectiveness of DRS, which exhibits a notable capability to integrate with established methodologies, yielding substantial improvements in both accuracy and ${\ell_2}$ certified robustness baselines of RS on the CIFAR-10 and ImageNet datasets. Code is available at https://github.com/xiasong0501/DRS.
****

Souradip Chakraborty,Amrit Bedi,Alec Koppel,Huazheng Wang,Dinesh Manocha,Mengdi Wang,Furong Huang
PARL: A Unified Framework for Policy Alignment in Reinforcement Learning
We present a novel unified bilevel optimization-based framework, \textsf{PARL}, formulated to address the recently highlighted critical issue of policy alignment in reinforcement learning using utility or preference-based feedback. We identify a major gap within current algorithmic designs for solving policy alignment due to a lack of precise characterization of the dependence of the alignment objective on the data generated by policy trajectories. This shortfall contributes to the sub-optimal performance observed in contemporary algorithms.
Our framework addressed these concerns by explicitly parameterizing the distribution of the upper alignment objective (reward design) by the lower optimal vari

able (optimal policy for the designed reward). Interestingly, from an optimizati
on perspective, our formulation leads to a new class of stochastic bilevel probl
ems where the stochasticity at the upper objective depends upon the lower-level
variable.
To demonstrate the efficacy of our formulation in resolving alignment issues in
RL, we devised an algorithm named \textsf{A-PARL} to solve PARL problem, establi
shing sample complexity bounds of order $\mathcal{O}(1/T)$. Our empirical result
s substantiate that the proposed \textsf{PARL} can address the alignment concern
s in RL by showing significant improvements (up to 63\% in terms of required sam
ples) for policy alignment in large-scale environments of the Deepmind control s
uite and Meta world tasks.
****
Zhiwei Xu,Yutong Wang,Spencer Frei,Gal Vardi,Wei Hu
Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data
Neural networks trained by gradient descent (GD) have exhibited a number of surp
rising generalization behaviors. First, they can achieve a perfect fit to noisy
training data and still generalize near-optimally, showing that overfitting can
sometimes be benign. Second, they can undergo a period of classical, harmful ove
rfitting---achieving a perfect fit to training data with near-random performance
 on test data---before transitioning (''grokking'') to near-optimal generalizati
on later in training. In this work, we show that both of these phenomena provabl
y occur in two-layer ReLU networks trained by GD on XOR cluster data where a con
stant fraction of the training labels are flipped. In this setting, we show that
 after the first step of GD, the network achieves 100\% training accuracy, perfe
ctly fitting the noisy labels in the training data, but achieves near-random tes
t accuracy. At a later training step, the network achieves near-optimal test acc
uracy while still fitting the random labels in the training data, exhibiting a '
'grokking'' phenomenon. This provides the first theoretical result of benign ove
rfitting in neural network classification when the data distribution is not line
arly separable. Our proofs rely on analyzing the feature learning process under
GD, which reveals that the network implements a non-generalizable linear classif
ier after one step and gradually learns generalizable features in later steps.
****
Yujee Song,Donghyun LEE,Rui Meng,Won Hwa Kim
Decoupled Marked Temporal Point Process using Neural Ordinary Differential Equat
ions
A Marked Temporal Point Process (MTPP) is a stochastic process whose realization
 is a set of event-time data. MTPP is often used to understand complex dynamics
of asynchronous temporal events such as money transaction, social media, healthc
are, etc. Recent studies have utilized deep neural networks to capture complex t
emporal dependencies of events and generate embedding that aptly represent the o
bserved events. While most previous studies focus on the inter-event dependencie
s and their representations, how individual events influence the overall dynamic
s over time has been under-explored. In this regime, we propose a Decoupled MTPP
 framework that disentangles characterization of a stochastic process into a set
 of evolving influences from different events. Our approach employs Neural Ordin
ary Differential Equations (Neural ODEs) to learn flexible continuous dynamics o
f these influences while simultaneously addressing multiple inference problems,
such as density estimation and survival rate computation. We emphasize the signi
ficance of disentangling the influences by comparing our framework with state-of
-the-art methods on real-life datasets, and provide analysis on the model behavi
or for potential applications.
****
Kuofeng Gao,Yang Bai,Jindong Gu,Shu-Tao Xia,Philip Torr,Zhifeng Li,Wei Liu
Inducing High Energy-Latency of Large Vision-Language Models with Verbose Images
Large vision-language models (VLMs) such as GPT-4 have achieved exceptional perf
ormance across various multi-modal tasks. However, the deployment of VLMs necess
itates substantial energy consumption and computational resources. Once attacker
s maliciously induce high energy consumption and latency time (energy-latency co
st) during inference of VLMs, it will exhaust computational resources. In this p

aper, we explore this attack surface about availability of VLMs and aim to induc
e high energy-latency cost during inference of VLMs. We find that high energy-la
tency cost during inference of VLMs can be manipulated by maximizing the length
of generated sequences. To this end, we propose verbose images, with the goal of
 crafting an imperceptible perturbation to induce VLMs to generate long sentence
s during inference. Concretely, we design three loss objectives. First, a loss i
s proposed to delay the occurrence of end-of-sequence (EOS) token, where EOS tok
en is a signal for VLMs to stop generating further tokens. Moreover, an uncertai
nty loss and a token diversity loss are proposed to increase the uncertainty ove
r each generated token and the diversity among all tokens of the whole generated
 sequence, respectively, which can break output dependency at token-level and se
quence-level. Furthermore, a temporal weight adjustment algorithm is proposed, w
hich can effectively balance these losses. Extensive experiments demonstrate tha
t our verbose images can increase the length of generated sequences by 7.87× and
 8.56× compared to original images on MS-COCO and ImageNet datasets, which prese
nts potential challenges for various applications.
****

Yaxin Fang,Faming Liang
Causal-StoNet: Causal Inference for High-Dimensional Complex Data
With the advancement of data science, the collection of increasingly complex dat
asets has become commonplace. In such datasets, the data dimension can be extrem
ely high, and the underlying data generation process can be unknown and highly n
onlinear. As a result, the task of making causal inference with high-dimensional
 complex data has become a fundamental problem in many disciplines, such as medi
cine, econometrics, and social science. However, the existing methods for causal
 inference are frequently developed under the assumption that the data dimension
 is low or that the underlying data generation process is linear or approximatel
y linear. To address these challenges, this paper proposes a novel stochastic de
ep learning approach for conducting causal inference with high-dimensional compl
ex data. The proposed approach is based on some deep learning techniques, includ
ing sparse deep learning theory and stochastic neural networks, that have been d
eveloped in recent literature. By using these techniques, the proposed approach
can address both the high dimensionality and unknown data generation process in
a coherent way. Furthermore, the proposed approach can also be used when missing
 values are present in the datasets. Extensive numerical studies indicate that
the proposed approach outperforms existing ones.
****

Yiyang Ma,Huan Yang,Wenhan Yang,Jianlong Fu,Jiaying Liu
Solving Diffusion ODEs with Optimal Boundary Conditions for Better Image Super-R
esolution
Diffusion models, as a kind of powerful generative model, have given impressive
results on image super-resolution (SR) tasks. However, due to the randomness int
roduced in the reverse process of diffusion models, the performances of diffusio
n-based SR models are fluctuating at every time of sampling, especially for samp
lers with few resampled steps. This inherent randomness of diffusion models resu
lts in ineffectiveness and instability, making it challenging for users to guara
ntee the quality of SR results. However, our work takes this randomness as an op
portunity: fully analyzing and leveraging it leads to the construction of an eff
ective plug-and-play sampling method that owns the potential to benefit a series
 of diffusion-based SR methods. More in detail, we propose to steadily sample hi
gh-quality SR images from pre-trained diffusion-based SR models by solving diffu
sion ordinary differential equations (diffusion ODEs) with optimal boundary cond
itions (BCs) and analyze the characteristics between the choices of BCs and thei
r corresponding SR results. Our analysis shows the route to obtain an approximat
ely optimal BC via an efficient exploration in the whole space. The quality of S
R results sampled by the proposed method with fewer steps outperforms the qualit
y of results sampled by current methods with randomness from the same pre-traine
d diffusion-based SR model, which means that our sampling method ''boosts'' curr
ent diffusion-based SR models without any additional training.
****

Paul Pu Liang,Chun Kai Ling,Yun Cheng,Alexander Obolenskiy,Yudong Liu,Rohan Pandey,Alex Wilf,Louis-Philippe Morency,Russ Salakhutdinov
Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications
In many machine learning systems that jointly learn from multiple modalities, a core research question is to understand the nature of multimodal interactions: how modalities combine to provide new task-relevant information that was not present in either alone. We study this challenge of interaction quantification in a semi-supervised setting with only labeled unimodal data and naturally co-occurring multimodal data (e.g., unlabeled images and captions, video and corresponding audio) but when labeling them is time-consuming. Using a precise information-theoretic definition of interactions, our key contribution is the derivation of lower and upper bounds to quantify the amount of multimodal interactions in this semi-supervised setting. We propose two lower bounds: one based on the shared information between modalities and the other based on disagreement between separately trained unimodal classifiers, and derive an upper bound through connections to approximate algorithms for min-entropy couplings. We validate these estimated bounds and show how they accurately track true interactions. Finally, we show how these theoretical results can be used to estimate multimodal model performance, guide data collection, and select appropriate multimodal models for various tasks.
****
Suhyeon Lee,Won Jun Kim,Jinho Chang,Jong Chul Ye
LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation
Following the impressive development of LLMs, vision-language alignment in LLMs is actively being researched to enable multimodal reasoning and visual input/output. This direction of research is particularly relevant to medical imaging because accurate medical image analysis and generation consist of a combination of reasoning based on visual features and prior knowledge. Many recent works have focused on training adapter networks that serve as an information bridge between image processing (encoding or generating) networks and LLMs; but presumably, in order to achieve maximum reasoning potential of LLMs on visual information as well, visual and language features should be allowed to interact more freely. This is especially important in the medical domain because understanding and generating medical images such as chest X-rays (CXR) require not only accurate visual and language-based reasoning but also a more intimate mapping between the two modalities. Thus, taking inspiration from previous work on the transformer and VQ-GAN combination for bidirectional image and text generation, we build upon this approach and develop a method for instruction-tuning an LLM pre-trained only on text to gain vision-language capabilities for medical images. Specifically, we leverage a pretrained LLM’s existing question-answering and instruction-following abilities to teach it to understand visual inputs by instructing it to answer questions about image inputs and, symmetrically, output both text and image responses appropriate to a given query by tuning the LLM with diverse tasks that encompass image-based text-generation and text-based image-generation. We show that our LLM-CXR trained in this approach shows better image-text alignment in both CXR understanding and generation tasks while being smaller in size compared to previously developed models that perform a narrower range of tasks.
****
Yiming Gao,Feiyu Liu,Liang Wang,Dehua Zheng,Zhenjie Lian,Weixuan Wang,Wenjin Yang,Siqin Li,Xianliang Wang,Wenhui Chen,Jing Dai,QIANG FU,Yang Wei,Lanxiao Huang,Wei Liu
Enhancing Human Experience in Human-Agent Collaboration: A Human-Centered Modeling Approach Based on Positive Human Gain
Existing game AI research mainly focuses on enhancing agents' abilities to win games, but this does not inherently make humans have a better experience when collaborating with these agents. For example, agents may dominate the collaboration and exhibit unintended or detrimental behaviors, leading to poor experiences for their human partners. In other words, most game AI agents are modeled in a "self-centered" manner. In this paper, we propose a "human-centered" modeling scheme for collaborative agents that aims to enhance the experience of humans. Specif

ically, we model the experience of humans as the goals they expect to achieve during the task. We expect that agents should learn to enhance the extent to which humans achieve these goals while maintaining agents' original abilities (e.g., winning games). To achieve this, we propose the Reinforcement Learning from Human Gain (RLHG) approach. The RLHG approach introduces a "baseline", which corresponds to the extent to which humans primitively achieve their goals, and encourages agents to learn behaviors that can effectively enhance humans in achieving their goals better. We evaluate the RLHG agent in the popular Multi-player Online Battle Arena (MOBA) game, Honor of Kings, by conducting real-world human-agent tests. Both objective performance and subjective preference results show that the RLHG agent provides participants better gaming experience.
****
Zhongwang Zhang,Yuqing Li,Tao Luo,Zhi-Qin John Xu
Stochastic Modified Equations and Dynamics of Dropout Algorithm
Dropout is a widely utilized regularization technique in the training of neural networks, nevertheless, its underlying mechanism and impact on achieving good generalization abilities remain to be further understood. In this work, we start by undertaking a rigorous theoretical derivation of the stochastic modified equations, with the primary aim of providing an effective approximation for the discrete iterative process of dropout. Meanwhile, we experimentally verify SDE's ability to approximate dropout under a wider range of settings. Subsequently, we empirically delve into the intricate mechanisms by which dropout facilitates the identification of flatter minima. This exploration is conducted through intuitive approximations, exploiting the structural analogies inherent in the Hessian of loss landscape and the covariance of dropout. Our empirical findings substantiate the ubiquitous presence of the Hessian-variance alignment relation throughout the training process of dropout.
****
Guangsheng Bao,Yanbin Zhao,Zhiyang Teng,Linyi Yang,Yue Zhang
Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature
Large language models (LLMs) have shown the ability to produce fluent and cogent content, presenting both productivity opportunities and societal risks. To build trustworthy AI systems, it is imperative to distinguish between machine-generated and human-authored content. The leading zero-shot detector, DetectGPT, showcases commendable performance but is marred by its intensive computational costs. In this paper, we introduce the concept of **conditional probability curvature** to elucidate discrepancies in word choices between LLMs and humans within a given context. Utilizing this curvature as a foundational metric, we present **Fast-DetectGPT**, an optimized zero-shot detector, which substitutes DetectGPT's perturbation step with a more efficient sampling step. Our evaluations on various datasets, source models, and test conditions indicate that Fast-DetectGPT not only surpasses DetectGPT by a relative around 75\% in both the white-box and black-box settings but also accelerates the detection process by a factor of 340, as detailed in Table 1.
****
seyed amir hossein saberi,Amir Najafi,Alireza Heidari,Mohammad Hosein Movasaghinia,Abolfazl Motahari,Babak Khalaj
Out-Of-Domain Unlabeled Data Improves Generalization
We propose a novel framework for incorporating unlabeled data into semi-supervised classification problems, where scenarios involving the minimization of either i) adversarially robust or ii) non-robust loss functions have been considered. Notably, we allow the unlabeled samples to deviate slightly (in total variation sense) from the in-domain distribution. The core idea behind our framework is to combine Distributionally Robust Optimization (DRO) with self-supervised training. As a result, we also leverage efficient polynomial-time algorithms for the training stage. From a theoretical standpoint, we apply our framework on the classification problem of a mixture of two Gaussians in $\mathbb{R}^d$, where in addition to the m independent and labeled samples from the true distribution, a set of n (usually with $n\gg m$) out of domain and unlabeled samples are gievn a

s well. Using only the labeled data, it is known that the generalization error can be bounded by $\propto\left(d/m\right)^{1/2}$. However, using our method on both isotropic and non-isotropic Gaussian mixture models, one can derive a new set of analytically explicit and non-asymptotic bounds which show substantial improvement on the generalization error compared ERM. Our results underscore two significant insights: 1) out-of-domain samples, even when unlabeled, can be harnessed to narrow the generalization gap, provided that the true data distribution adheres to a form of the "cluster assumption", and 2) the semi-supervised learning paradigm can be regarded as a special case of our framework when there are no distributional shifts. We validate our claims through experiments conducted on a variety of synthetic and real-world datasets.
****
Keqiang Yan,Cong Fu,Xiaofeng Qian,Xiaoning Qian,Shuiwang Ji
Complete and Efficient Graph Transformers for Crystal Material Property Prediction
Crystal structures are characterized by atomic bases within a primitive unit cell that repeats along a regular lattice throughout 3D space. The periodic and infinite nature of crystals poses unique challenges for geometric graph representation learning. Specifically, constructing graphs that effectively capture the complete geometric information of crystals and handle chiral crystals remains an unsolved and challenging problem. In this paper, we introduce a novel approach that utilizes the periodic patterns of unit cells to establish the lattice-based representation for each atom, enabling efficient and expressive graph representations of crystals. Furthermore, we propose ComFormer, a SE(3) transformer designed specifically for crystalline materials. ComFormer includes two variants; namely, iComFormer that employs invariant geometric descriptors of Euclidean distances and angles, and eComFormer that utilizes equivariant vector representations. Experimental results demonstrate the state-of-the-art predictive accuracy of ComFormer variants on various tasks across three widely-used crystal benchmarks. Our code is publicly available as part of the AIRS library (https://github.com/divelab/AIRS).
****
Jae-Hong Lee,Joon-Hyuk Chang
Continual Momentum Filtering on Parameter Space for Online Test-time Adaptation
Deep neural networks (DNNs) have revolutionized tasks such as image classification and speech recognition but often falter when training and test data diverge in distribution. External factors, from weather effects on images to varied speech environments, can cause this discrepancy, compromising DNN performance. Online test-time adaptation (OTTA) methods present a promising solution, recalibrating models in real-time during the test stage without requiring historical data. However, the OTTA paradigm is imperfect, often falling prey to issues such as catastrophic forgetting due to its reliance on noisy, self-trained predictions. Although some contemporary strategies mitigate this by tying adaptations to the static source model, this restricts model flexibility. This paper introduces a continual momentum filtering (CMF) framework, leveraging the Kalman filter (KF) to strike a balance between model adaptability and information retention. The CMF intertwines optimization via stochastic gradient descent with a KF-based inference process. This methodology not only aids in averting catastrophic forgetting but also provides high adaptability to shifting data distributions. We validate our framework on various OTTA scenarios and real-world situations regarding covariate and label shifts, and the CMF consistently shows superior performance compared to state-of-the-art methods.
****
Jian-Feng CAI,Yu Long,Ruixue WEN,Jiaxi Ying
A Fast and Provable Algorithm for Sparse Phase Retrieval
We study the sparse phase retrieval problem, which seeks to recover a sparse signal from a limited set of magnitude-only measurements. In contrast to prevalent sparse phase retrieval algorithms that primarily use first-order methods, we propose an innovative second-order algorithm that employs a Newton-type method with hard thresholding. This algorithm overcomes the linear convergence limitations

of first-order methods while preserving their hallmark per-iteration computational efficiency. We provide theoretical guarantees that our algorithm converges to the s-sparse ground truth signal $\boldsymbol{x}^{\natural} \in \mathbb{R}^n$ (up to a global sign) at a quadratic convergence rate after at most $O(\log (\Vert\boldsymbol{x}^{\natural} \Vert /x_{\min}^{\natural}))$ iterations, using $\Omega(s^2\log n)$ Gaussian random samples. Numerical experiments show that our algorithm achieves a significantly faster convergence rate than state-of-the-art methods.
****
Wenting Zhao,Xiang Ren,Jack Hessel,Claire Cardie,Yejin Choi,Yuntian Deng
(InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild
Chatbots such as GPT-4 and ChatGPT are now serving millions of users. Despite their widespread use, there remains a lack of public datasets showcasing how these tools are used by a population of users in practice. To bridge this gap, we offered free access to ChatGPT for online users in exchange for their affirmative, consensual, opt-in for anonymous collection of their chat transcripts. From this, we compiled (InThe)WildChat, a corpus of 570K user-ChatGPT conversations, which consists of over 1.5 million interaction turns. We compare WildChat with other popular user-chatbot interaction datasets, and find that our dataset offers the most diverse user prompts, contains the largest number of languages, and presents the richest variety of potentially toxic use-cases for researchers to study. In particular, in WildChat we find that a majority of the potentially unsafe use is produced by users attempting to “jailbreak” the model using prompts posted on online platforms; these are successful more than 70% of the time for ChatGPT. Finally, because it captures a broad range of use cases, we demonstrate the dataset’s potential utility in fine-tuning state-of-the-art instruction following models. WildLlama, a chatbot fine-tuned on WildChat, outperforms the latest Vicuna model of the same size on MT-Bench, which shows that WildChat has a high utility in addition to being a source for toxicity study. We will release WildChat and WildLlama with a license that emphasizes on accountability, collaboration, and transparency. The clean portion of WildChat will be publicly available, and the portion that contains potentially unsafe content will be made available upon request with a justification for AI safety research.
****
Yuzhang Shang,Zhihang Yuan,Zhen Dong
PB-LLM: Partially Binarized Large Language Models
This paper explores network binarization, a radical form of quantization, compressing model weights to a single bit, specifically for Large Language Models (LLMs) compression.
Due to previous binarization methods collapsing LLMs, we propose a novel approach, Partially-Binarized LLM (PB-LLM), which can achieve extreme low-bit quantization while maintaining the linguistic reasoning capacity of quantized LLMs.
Specifically, our exploration first uncovers the ineffectiveness of naïve applications of existing binarization algorithms and highlights the imperative role of salient weights in achieving low-bit quantization.
Thus, PB-LLM filters a small ratio of salient weights during binarization, allocating them to higher-bit storage, i.e., partially-binarization.
PB-LLM is extended to recover the capacities of quantized LMMs, by analyzing from the perspective of post-training quantization (PTQ) and quantization-aware training (QAT).
Under PTQ, combining the concepts from GPTQ, we reconstruct the binarized weight matrix guided by the Hessian matrix and successfully recover the reasoning capacity of PB-LLM in low-bit.
Under QAT, we freeze the salient weights during training, explore the derivation of optimal scaling factors crucial for minimizing the quantization error, and propose a scaling mechanism based on this derived scaling strategy for residual binarized weights.
Those explorations and the developed methodologies significantly contribute to rejuvenating the performance of low-bit quantized LLMs and present substantial advancements in the field of network binarization for LLMs.

Code is available at https://github.com/hahnyuan/PB-LLM.
****
Bo Li,Xiaowen Jiang,Mikkel N. Schmidt,Tommy Sonne Alstrøm,Sebastian U Stich
An improved analysis of per-sample and per-update clipping in federated learning
Gradient clipping is key mechanism that is essential to differentially private training techniques in Federated learning. Two popular strategies are per-sample clipping, which clips the mini-batch gradient, and per-update clipping, which clips each user's model update. However, there has not been a thorough theoretical analysis of these two clipping methods.

In this work, we rigorously analyze the impact of these two clipping techniques on the convergence of a popular federated learning algorithm FedAvg under standard stochastic noise and gradient dissimilarity assumptions. We provide a convergence guarantee given any arbitrary clipping threshold. Specifically, we show that per-sample clipping is guaranteed to converge to the neighborhood of the stationary point, with the size dependent on the stochastic noise, gradient dissimilarity, and clipping threshold. In contrast, the convergence to the stationary point can be guaranteed with a sufficiently small stepsize in per-update clipping at the cost of more communication rounds. We further provide insights into understanding the impact of the improved convergence analysis in the differentially private setting.
****
Chengrun Yang,Xuezhi Wang,Yifeng Lu,Hanxiao Liu,Quoc V Le,Denny Zhou,Xinyun Chen
Large Language Models as Optimizers
Optimization is ubiquitous. While derivative-based algorithms have been powerful tools for various problems, the absence of gradient imposes challenges on many real-world applications. In this work, we propose Optimization by PROmpting (OPRO), a simple and effective approach to leverage large language models (LLMs) as optimizers, where the optimization task is described in natural language. In each optimization step, the LLM generates new solutions from the prompt that contains previously generated solutions with their values, then the new solutions are evaluated and added to the prompt for the next optimization step. We first showcase OPRO on linear regression and traveling salesman problems, then move on to our main application in prompt optimization, where the goal is to find instructions that maximize the task accuracy. With a variety of LLMs, we demonstrate that the best prompts optimized by OPRO outperform human-designed prompts by up to 8% on GSM8K, and by up to 50% on Big-Bench Hard tasks. Code at https://github.com/google-deepmind/opro.
****
Qi Zhao,Shijie Wang,Ce Zhang,Changcheng Fu,Minh Quan Do,Nakul Agarwal,Kwonjoon Lee,Chen Sun
AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos?
Can we better anticipate an actor’s future actions (e.g. mix eggs) by knowing what commonly happens after the current action (e.g. crack eggs)? What if the actor also shares the goal (e.g. make fried rice) with us? The long-term action anticipation (LTA) task aims to predict an actor’s future behavior from video observations in the form of verb and noun sequences, and it is crucial for human-machine interaction.
We propose to formulate the LTA task from two perspectives: a bottom-up approach that predicts the next actions autoregressively by modeling temporal dynamics; and a top-down approach that infers the goal of the actor and plans the needed procedure to accomplish the goal. We hypothesize that large language models (LLMs), which have been pretrained on procedure text data (e.g. recipes, how-tos), have the potential to help LTA from both perspectives. It can help provide the prior knowledge on the possible next actions, and infer the goal given the observed part of a procedure, respectively. We propose AntGPT, which represents video observations as sequences of human actions, and uses the action representation for an LLM to infer the goals and model temporal dynamics. AntGPT achieves state-of-the-art performance on Ego4D LTA v1 and v2, EPIC-Kitchens-55, as well as EGTE

A GAZE+, thanks to LLMs’ goal inference and temporal dynamics modeling capabilities. We further demonstrate that these capabilities can be effectively distilled into a compact neural network 1.3% of the original LLM model size. Code and model will be released upon acceptance.
****
Ahmed A. A. Elhag,Yuyang Wang,Joshua M. Susskind,Miguel Ángel Bautista
Manifold Diffusion Fields
We present Manifold Diffusion Fields (MDF), an approach that unlocks learning of diffusion models of data in general non-euclidean geometries. Leveraging insights from spectral geometry analysis, we define an intrinsic coordinate system on the manifold via the eigen-functions of the Laplace-Beltrami Operator. MDF represents functions using an explicit parametrization formed by a set of multiple input-output pairs. Our approach allows to sample continuous functions on manifolds and is invariant with respect to rigid and isometric transformations of the manifold. In addition, we show that MDF generalizes to the case where the training set contains functions on different manifolds. Empirical results on multiple datasets and manifolds including challenging scientific problems like weather prediction or molecular conformation show that MDF can capture distributions of such functions with better diversity and fidelity than previous approaches.
****
Juncheng Li,Kaihang Pan,Zhiqi Ge,Minghe Gao,Wei Ji,Wenqiao Zhang,Tat-Seng Chua,Siliang Tang,Hanwang Zhang,Yueting Zhuang
Fine-tuning Multimodal LLMs to Follow Zero-shot Demonstrative Instructions
Recent advancements in Multimodal Large Language Models (MLLMs) have been utilizing Visual Prompt Generators (VPGs) to convert visual features into tokens that LLMs can recognize. This is achieved by training the VPGs on millions of image-caption pairs, where the VPG-generated tokens of images are fed into a frozen LLM to generate the corresponding captions. However, this image-captioning based training objective inherently biases the VPG to concentrate solely on the primary visual contents sufficient for caption generation, often neglecting other visual details. This shortcoming results in MLLMs’ underperformance in comprehending demonstrative instructions consisting of multiple, interleaved, and multimodal instructions that demonstrate the required context to complete a task. To address this issue, we introduce a generic and lightweight Visual Prompt Generator Complete module (VPG-C), which can infer and complete the missing details essential for comprehending demonstrative instructions. Further, we propose a synthetic discriminative training strategy to fine-tune VPG-C, eliminating the need for supervised demonstrative instructions. As for evaluation, we build DEMON, a comprehensive benchmark for demonstrative instruction understanding. Synthetically trained with the proposed strategy, VPG-C achieves significantly stronger zero-shot performance across all tasks of DEMON. Further evaluation on the MME and OwlEval benchmarks also demonstrate the superiority of VPG-C. The code and models are available at https://github.com/DCDmllm/Cheetah.
****
Ge Yan,Yaniv Romano,Tsui-Wei Weng
Provably Robust Conformal Prediction with Improved Efficiency
Conformal prediction is a powerful tool to generate uncertainty sets with guaranteed coverage using any predictive model, under the assumption that the training and test data are i.i.d.. Recently, it has been shown that adversarial examples are able to manipulate conformal methods to construct prediction sets with invalid coverage rates, as the i.i.d. assumption is violated. To address this issue, a recent work, Randomized Smoothed Conformal Prediction (RSCP), was first proposed to certify the robustness of conformal prediction methods to adversarial noise. However, RSCP has two major limitations: (i) its robustness guarantee is flawed when used in practice and (ii) it tends to produce large uncertainty sets. To address these limitations, we first propose a novel framework called RSCP+ to provide provable robustness guarantee in evaluation, which fixes the issues in the original RSCP method. Next, we propose two novel methods, Post-Training Transformation (PTT) and Robust Conformal Training (RCT), to effectively reduce prediction set size with little computation overhead. Experimental results in CIFAR10

, CIFAR100, and ImageNet suggest the baseline method only yields trivial predict
ions including full label set, while our methods could boost the efficiency by u
p to $4.36\times$, $5.46\times$, and $16.9\times$ respectively and provide pract
ical robustness guarantee.
****
Jie Hu,Vishwaraj Doshi,Do Young Eun
Accelerating Distributed Stochastic Optimization via Self-Repellent Random Walks
We study a family of distributed stochastic optimization algorithms where gradie
nts are sampled by a token traversing a network of agents in random-walk fashion
. Typically, these random-walks are chosen to be Markov chains that asymptotical
ly sample from a desired target distribution, and play a critical role in the co
nvergence of the optimization iterates. In this paper, we take a novel approach
by replacing the standard *linear* Markovian token by one which follows a *non-l
inear* Markov chain - namely the Self-Repellent Radom Walk (SRRW). Defined for a
ny given 'base' Markov chain, the SRRW, parameterized by a positive scalar $\\al
pha$, is less likely to transition to states that were highly visited in the pas
t, thus the name. In the context of MCMC sampling on a graph, a recent breakthro
ugh in Doshi et al. (2023) shows that the SRRW achieves $O(1/\\alpha)$ decrease
in the asymptotic variance for sampling. We propose the use of a `generalized' v
ersion of the SRRW to drive token algorithms for distributed stochastic optimiza
tion in the form of stochastic approximation, termed SA-SRRW. We prove that the
optimization iterate errors of the resulting SA-SRRW converge to zero almost sur
ely and prove a central limit theorem, deriving the explicit form of the resulti
ng asymptotic covariance matrix corresponding to iterate errors. This asymptotic
 covariance is always smaller than that of an algorithm driven by the base Marko
v chain and decreases at rate $O(1/\\alpha^2)$ - the performance benefit of usin
g SRRW thereby *amplified* in the stochastic optimization context. Empirical res
ults support our theoretical findings.
****
Wei Liu,Weihao Zeng,Keqing He,Yong Jiang,Junxian He
What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Sele
ction in Instruction Tuning
Instruction tuning is a standard technique employed to align large language mode
ls to end tasks and user preferences after the initial pretraining phase. Recent
 research indicates the critical role of data engineering in instruction tuning
-- when appropriately selected, only limited data is necessary to achieve superi
or performance. However, we still lack a principled understanding of what makes
good instruction tuning data for alignment, and how we should select data automa
tically and effectively. In this work, we delve deeply into automatic data selec
tion strategies for alignment. We start with controlled studies to measure data
across three dimensions: complexity, quality, and diversity, along which we exam
ine existing methods and introduce novel techniques for enhanced data measuremen
t. Subsequently, we propose a simple strategy to select data samples based on th
e measurement. We present Deita (short for Data-Efficient Instruction Tuning for
 Alignment), a series of models fine-tuned from LLaMA models using data samples
automatically selected with our proposed approach. When assessed through both a
utomatic metrics and human evaluation, Deita performs better or on par with the
state-of-the-art open-source alignment models such as Vicuna and WizardLM with o
nly 6K training data samples -- 10x less than the data used in the baselines. We
 anticipate this work to provide clear guidelines and tools on automatic data se
lection, aiding researchers and practitioners in achieving data-efficient alignm
ent.
****
Hyungyu Lee,Saehyung Lee,Hyemi Jang,Junsung Park,Ho Bae,Sungroh Yoon
DAFA: Distance-Aware Fair Adversarial Training
The disparity in accuracy between classes in standard training is amplified duri
ng adversarial training, a phenomenon termed the robust fairness problem. Existi
ng methodologies aimed to enhance robust fairness by sacrificing the model's per
formance on easier classes in order to improve its performance on harder ones. H
owever, we observe that under adversarial attacks, the majority of the model's p

redictions for samples from the worst class are biased towards classes similar t
o the worst class, rather than towards the easy classes. Through theoretical and
 empirical analysis, we demonstrate that robust fairness deteriorates as the dis
tance between classes decreases. Motivated by these insights, we introduce the D
istance-Aware Fair Adversarial Training (DAFA) methodology, which addresses robu
st fairness by taking into account the similarities between classes. Specificall
y, our method assigns distinct adversarial margins and loss weights to each clas
s and adjusts them to encourage a trade-off in robustness among similar classes.
 Experimental results across various datasets demonstrate that our method not on
ly maintains average robust accuracy but also significantly improves the worst r
obust accuracy, indicating a marked improvement in robust fairness compared to e
xisting methods.
****
Haiping Wang,Yuan Liu,Bing WANG,YUJING SUN,Zhen Dong,Wenping Wang,Bisheng Yang
FreeReg: Image-to-Point Cloud Registration Leveraging Pretrained Diffusion Model
s and Monocular Depth Estimators
Matching cross-modality features between images and point clouds is a fundamenta
l problem for image-to-point cloud registration. However, due to the modality di
fference between images and points, it is difficult to learn robust and discrimi
native cross-modality features by existing metric learning methods for feature m
atching. Instead of applying metric learning on cross-modality data, we propose
to unify the modality between images and point clouds by pretrained large-scale
models first, and then establish robust correspondence within the same modality.
 We show that the intermediate features, called diffusion features, extracted by
 depth-to-image diffusion models are semantically consistent between images and
point clouds, which enables the building of coarse but robust cross-modality cor
respondences. We further extract geometric features on depth maps produced by th
e monocular depth estimator. By matching such geometric features, we significant
ly improve the accuracy of the coarse correspondences produced by diffusion feat
ures. Extensive experiments demonstrate that without any task-specific training,
 direct utilization of both features produces accurate image-to-point cloud regi
stration. On three public indoor and outdoor benchmarks, the proposed method ave
ragely achieves a 20.6 percent improvement in Inlier Ratio, a $3.0\times$ higher
 Inlier Number, and a 48.6 percent improvement in Registration Recall than exist
ing state-of-the-arts. The code and additional results are available at \url{htt
ps://whu-usi3dv.github.io/FreeReg/}.
****
Ganghua Wang,Xun Xian,Ashish Kundu,Jayanth Srinivasa,Xuan Bi,Mingyi Hong,Jie Din
g
Demystifying Poisoning Backdoor Attacks from a Statistical Perspective
Backdoor attacks pose a significant security risk to machine learning applicatio
ns due to their stealthy nature and potentially serious consequences. Such attac
ks involve embedding triggers within a learning model with the intention of caus
ing malicious behavior when an active trigger is present while maintaining regul
ar functionality without it. This paper derives a fundamental understanding of b
ackdoor attacks that applies to both discriminative and generative models, inclu
ding diffusion models and large language models. We evaluate the effectiveness o
f any backdoor attack incorporating a constant trigger, by establishing tight lo
wer and upper boundaries for the performance of the compromised model on both cl
ean and backdoor test data. The developed theory answers a series of fundamental
 but previously underexplored problems, including (1) what are the determining f
actors for a backdoor attack's success, (2) what is the direction of the most ef
fective backdoor attack, and (3) when will a human-imperceptible trigger succeed
. We demonstrate the theory by conducting experiments using benchmark datasets a
nd state-of-the-art backdoor attack scenarios. Our code is available \href{https
://github.com/KeyWgh/DemystifyBackdoor}{here}.
****
Lianmin Zheng,Wei-Lin Chiang,Ying Sheng,Tianle Li,Siyuan Zhuang,Zhanghao Wu,Yong
hao Zhuang,Zhuohan Li,Zi Lin,Eric Xing,Joseph E. Gonzalez,Ion Stoica,Hao Zhang
LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset

Studying how people interact with large language models (LLMs) in real-world sce
narios is increasingly important due to their widespread use in various applicat
ions. In this paper, we introduce LMSYS-Chat-1M, a large-scale dataset containin
g one million real-world conversations with 25 state-of-the-art LLMs. This datas
et is collected from 210K unique IP addresses in the wild on our Vicuna demo and
 Chatbot Arena website. We offer an overview of the dataset's content, including
 its curation process, basic statistics, and topic distribution, highlighting it
s diversity, originality, and scale. We demonstrate its versatility through four
 use cases: developing content moderation models that perform similarly to GPT-4
, building a safety benchmark, training instruction-following models that perfor
m similarly to Vicuna, and creating challenging benchmark questions. We believe
that this dataset will serve as a valuable resource for understanding and advanc
ing LLM capabilities. The dataset is publicly available at https://huggingface.c
o/datasets/lmsys/lmsys-chat-1m.
****
Weidi Xu,Jingwei Wang,Lele Xie,Jianshan He,Hongting Zhou,Taifeng Wang,Xiaopei Wa
n,Jingdong Chen,Chao Qu,Wei Chu
LogicMP: A Neuro-symbolic Approach for Encoding First-order Logic Constraints
Integrating first-order logic constraints (FOLCs) with neural networks is a cruc
ial but challenging problem since it involves modeling intricate correlations to
 satisfy the constraints. This paper proposes a novel neural layer, LogicMP, whi
ch performs mean-field variational inference over a Markov Logic Network (MLN).
It can be plugged into any off-the-shelf neural network to encode FOLCs while re
taining modularity and efficiency. By exploiting the structure and symmetries in
 MLNs, we theoretically demonstrate that our well-designed, efficient mean-field
 iterations greatly mitigate the difficulty of MLN inference, reducing the infer
ence from sequential calculation to a series of parallel tensor operations. Empi
rical results in three kinds of tasks over images, graphs, and text show that Lo
gicMP outperforms advanced competitors in both performance and efficiency.
****
Hongzhi Wen,Wenzhuo Tang,Xinnan Dai,Jiayuan Ding,Wei Jin,Yuying Xie,Jiliang Tang
CellPLM: Pre-training of Cell Language Model Beyond Single Cells
The current state-of-the-art single-cell pre-trained models are greatly inspired
 by the success of large language models. They trained transformers by treating
genes as tokens and cells as sentences. However, three fundamental differences b
etween single-cell data and natural language data are overlooked: (1) scRNA-seq
data are presented as bag-of-genes instead of sequences of RNAs; (2) Cell-cell r
elations are more intricate and important than inter-sentence relations; and (3)
 The quantity of single-cell data is considerably inferior to text data, and the
y are very noisy. In light of these characteristics, we propose a new pre-traine
d model, CellPLM, which takes cells as tokens and tissues as sentences. In addit
ion, we leverage spatially-resolved transcriptomic data in pre-training to facil
itate learning cell-cell relationships and introduce a Gaussian prior distributi
on as an additional inductive bias to overcome data limitations. CellPLM is the
first single-cell pre-trained transformer that encodes cell-cell relations and i
t consistently achieves state-of-the-art performance across distinct downstream
tasks.
****
Bowen Jing,Tommi S. Jaakkola,Bonnie Berger
Equivariant Scalar Fields for Molecular Docking with Fast Fourier Transforms
Molecular docking is critical to structure-based virtual screening, yet the thro
ughput of such workflows is limited by the expensive optimization of scoring fun
ctions involved in most docking algorithms. We explore how machine learning can
accelerate this process by learning a scoring function with a functional form th
at allows for more rapid optimization. Specifically, we define the scoring funct
ion to be the cross-correlation of multi-channel ligand and protein scalar field
s parameterized by equivariant graph neural networks, enabling rapid optimizatio
n over rigid-body degrees of freedom with fast Fourier transforms. The runtime o
f our approach can be amortized at several levels of abstraction, and is particu
larly favorable for virtual screening settings with a common binding pocket. We

benchmark our scoring functions on two simplified docking-related tasks: decoy pose scoring and rigid conformer docking. Our method attains similar but faster performance on crystal structures compared to the widely-used Vina and Gnina scoring functions, and is more robust on computationally predicted structures. Code is available at https://github.com/bjing2016/scalar-fields.
****

Longwei Zou,Han Zhang,Yangdong Deng
A Multi-Level Framework for Accelerating Training Transformer Models
The fast growing capabilities of large-scale deep learning models, such as Bert, GPT and ViT, are revolutionizing the landscape of NLP, CV and many other domains. Training such models, however, poses an unprecedented demand for computing power, which incurs exponentially increasing energy cost and carbon dioxide emissions. It is thus critical to develop efficient training solutions to reduce the training costs. Motivated by a set of key observations of inter- and intra-layer similarities among feature maps and attentions that can be identified from typical training processes, we propose a multi-level framework for training acceleration. Specifically, the framework is based on three basic operators, Coalescing, De-coalescing and Interpolation, which can be orchestrated to build a multi-level training framework. The framework consists of a V-cycle training process, which progressively down- and up-scales the model size and projects the parameters between adjacent levels of models via coalescing and de-coalescing. The key idea is that a smaller model that can be trained for fast convergence and the trained parameters provides high-qualities intermediate solutions for the next level larger network. The interpolation operator is designed to break the symmetry of neurons incurred by de-coalescing for better convergence performance. Our experiments on transformer-based language models (e.g. Bert, GPT) as well as a vision model (e.g. DeiT) prove that the proposed framework reduces the computational cost by about 20% on training BERT/GPT-Base models and up to 51.6% on training the BERT-Large model while preserving the performance.
****

Ethan Baron,Itamar Zimerman,Lior Wolf
A 2-Dimensional State Space Layer for Spatial Inductive Bias
A central objective in computer vision is to design models with appropriate 2-D inductive bias. Desiderata for 2-D inductive bias include two-dimensional position awareness, dynamic spatial locality, and translation and permutation invariance. To address these goals, we leverage an expressive variation of the multidimensional State Space Model (SSM). Our approach introduces efficient parameterization, accelerated computation, and a suitable normalization scheme. Empirically, we observe that incorporating our layer at the beginning of each transformer block of Vision Transformers (ViT), as well as when replacing the Conv2D filters of ConvNeXT with our proposed layers significantly enhances performance for multiple backbones and across multiple datasets. The new layer is effective even with a negligible amount of additional parameters and inference time. Ablation studies and visualizations demonstrate that the layer has a strong 2-D inductive bias. For example, vision transformers equipped with our layer exhibit effective performance even without positional encoding. Our code is attached as supplementary.
****

Louis Béthune,Thomas Massena,Thibaut Boissin,Aurélien Bellet,Franck Mamalet,Yannick Prudent,Corentin Friedrich,Mathieu Serrurier,David Vigouroux
DP-SGD Without Clipping: The Lipschitz Neural Network Way
State-of-the-art approaches for training Differentially Private (DP) Deep Neural Networks (DNN) face difficulties to estimate tight bounds on the sensitivity of the network's layers, and instead rely on a process of per-sample gradient clipping. This clipping process not only biases the direction of gradients but also proves costly both in memory consumption and in computation. To provide sensitivity bounds and bypass the drawbacks of the clipping process, we propose to rely on Lipschitz constrained networks. Our theoretical analysis reveals an unexplored link between the Lipschitz constant with respect to their input and the one with respect to their parameters. By bounding the Lipschitz constant of each layer

with respect to its parameters, we prove that we can train these networks with privacy guarantees. Our analysis not only allows the computation of the aforeme
ntioned sensitivities at scale, but also provides guidance on how to maximize th
e gradient-to-noise ratio for fixed privacy guarantees. To facilitate the applic
ation of Lipschitz networks and foster robust and certifiable learning under pri
vacy guarantees, we provide a Python package that implements building blocks all
owing the construction and private training of such networks.
****
Youhan Lee,Hasun Yu,Jaemyung Lee,Jaehoon Kim
Pre-training Sequence, Structure, and Surface Features for Comprehensive Protein
 Representation Learning
Proteins can be represented in various ways, including their sequences, 3D struc
tures, and surfaces. While recent studies have successfully employed sequence- o
r structure-based representations to address multiple tasks in protein science,
there has been significant oversight in incorporating protein surface informatio
n, a critical factor for protein function. In this paper, we present a pre-train
ing strategy that incorporates information from protein sequences, 3D structures
, and surfaces to improve protein representation learning. Specifically, we util
ize Implicit Neural Representations (INRs) for learning surface characteristics,
 and name it ProteinINR. We confirm that ProteinINR successfully reconstructs pr
otein surfaces, and integrate this surface learning into the existing pre-traini
ng strategy of sequences and structures. Our results demonstrate that our approa
ch can enhance performance in various downstream tasks, thereby underscoring the
 importance of including surface attributes in protein representation learning.
These findings underline the importance of understanding protein surfaces for ge
nerating effective protein representations.
****
Zhenyi Wang,Yan Li,Li Shen,Heng Huang
A Unified and General Framework for Continual Learning
Continual Learning (CL) focuses on learning from dynamic and changing data distr
ibutions while retaining previously acquired knowledge. Various methods have bee
n developed to address the challenge of catastrophic forgetting, including regul
arization-based, Bayesian-based, and memory-replay-based techniques. However, th
ese methods lack a unified framework and common terminology for describing their
 approaches. This research aims to bridge this gap by introducing a comprehensiv
e and overarching framework that encompasses and reconciles these existing metho
dologies. Notably, this new framework is capable of encompassing established CL
approaches as special instances within a unified and general optimization object
ive.
An intriguing finding is that despite their diverse origins, these methods share
 common mathematical structures. This observation highlights the compatibility o
f these seemingly distinct techniques, revealing their interconnectedness throug
h a shared underlying optimization objective. Moreover, the proposed general fra
mework introduces an innovative concept called *refresh learning*, specifically
designed to enhance the CL performance. This novel approach draws inspiration fr
om neuroscience, where the human brain often sheds outdated information to impro
ve the retention of crucial knowledge and facilitate the acquisition of new info
rmation. In essence, *refresh learning* operates by initially unlearning current
 data and subsequently relearning it. It serves as a versatile plug-in that seam
lessly integrates with existing CL methods, offering an adaptable and effective
enhancement to the learning process. Extensive experiments on CL benchmarks and
theoretical analysis demonstrate the effectiveness of the proposed *refresh lear
ning*.
****
Junyi An,Chao Qu,Zhipeng Zhou,Fenglei Cao,Xu Yinghui,Yuan Qi,Furao Shen
Hybrid Directional Graph Neural Network for Molecules
Equivariant message passing neural networks have emerged as the prevailing appro
ach for predicting chemical properties of molecules due to their ability to leve
rage translation and rotation symmetries, resulting in a strong inductive bias.
However, the equivariant operations in each layer can impose excessive constrain

ts on the function form and network flexibility. To address these challenges, we introduce a novel network called the Hybrid Directional Graph Neural Network (HDGNN), which effectively combines strictly equivariant operations with learnable modules. We evaluate the performance of HDGNN on the QM9 dataset and the IS2RE dataset of OC20, demonstrating its state-of-the-art performance on several tasks and competitive performance on others. Our code is anonymously released on https://github.com/ajy112/HDGNN.
****
AJAY KUMAR JAISWAL,Zhe Gan,Xianzhi Du,Bowen Zhang,Zhangyang Wang,Yinfei Yang
Compressing LLMs: The Truth is Rarely Pure and Never Simple
Despite their remarkable achievements, modern Large Language Models (LLMs) encounter exorbitant computational and memory footprints. Recently, several works have shown significant success in *training-free* and *data-free* compression (pruning and quantization) of LLMs achieving 50-60\% sparsity and reducing the bit-width down to 3 or 4 bits per weight, with negligible perplexity degradation over the uncompressed baseline. As recent research efforts are focused on developing increasingly sophisticated compression methods, our work takes a step back, and re-evaluates the effectiveness of existing SoTA compression methods, which rely on a fairly simple and widely questioned metric, perplexity (even for dense LLMs). We introduce **K**nowledge-**I**ntensive **C**ompressed LLM Benchmar**K** **(LLM-KICK)**, a collection of carefully-curated tasks to re-define the evaluation protocol for compressed LLMs, which have significant alignment with their dense counterparts, and perplexity fail to capture subtle change in their true capabilities. LLM-KICK unveils many favorable merits and unfortunate plights of current SoTA compression methods: all pruning methods suffer significant performance degradation, sometimes at trivial sparsity ratios (*e.g.*, 25-30\%), and fail for N:M sparsity on knowledge-intensive tasks; current quantization methods are more successful than pruning; yet, pruned LLMs even at ≥ 50\% sparsity are robust in-context retrieval and summarization systems; among others. LLM-KICK is designed to holistically access compressed LLMs' ability for language understanding, reasoning, generation, in-context retrieval, in-context summarization, *etc.* We hope our study can foster the development of better LLM compression methods. The reproduced codes are available at https://github.com/VITA-Group/llm-kick.
****
Yuchen Zhuang,Xiang Chen,Tong Yu,Saayan Mitra,Victor Bursztyn,Ryan A. Rossi,Somdeb Sarkhel,Chao Zhang
ToolChain*: Efficient Action Space Navigation in Large Language Models with A* Search
Large language models (LLMs) have demonstrated powerful decision-making and planning capabilities in solving complicated real-world problems. LLM-based autonomous agents can interact with diverse tools (e.g., functional APIs) and generate solution plans that execute a series of API function calls in a step-by-step manner. The multitude of candidate API function calls significantly expands the action space, amplifying the critical need for efficient action space navigation. However, existing methods either struggle with unidirectional exploration in expansive action spaces, trapped into a locally optimal solution, or suffer from exhaustively traversing all potential actions, causing inefficient navigation. To address these issues, we propose ToolChain*, an efficient tree search-based planning algorithm for LLM-based agents. It formulates the entire action space as a decision tree, where each node represents a possible API function call involved in a solution plan. By incorporating the A* search algorithm with task-specific cost function design, it efficiently prunes high-cost branches that may involve incorrect actions, identifying the most low-cost valid path as the solution. Extensive experiments on multiple tool-use and reasoning tasks demonstrate that ToolChain* efficiently balances exploration and exploitation within an expansive action space. It outperforms state-of-the-art baselines on planning and reasoning tasks by 3.1% and 3.5% on average while requiring 7.35x and 2.31x less time, respectively.
****
Zipeng Wang,Xuehui Yu,Xumeng Han,Wenwen Yu,Zhixun Huang,Jianbin Jiao,Zhenjun Han

P2Seg: Pointly-supervised Segmentation via Mutual Distillation
Point-level Supervised Instance Segmentation (PSIS) aims to enhance the applicab
ility and scalability of instance segmentation by utilizing low-cost yet instanc
e-informative annotations. Existing PSIS methods usually rely on positional info
rmation to distinguish objects, but predicting precise boundaries remains challe
nging due to the lack of contour annotations. Nevertheless, weakly supervised se
mantic segmentation methods are proficient in utilizing intra-class feature cons
istency to capture the boundary contours of the same semantic regions. In this p
aper, we design a Mutual Distillation Module (MDM) to leverage the complementary
 strengths of both instance position and semantic information and achieve accura
te instance-level object perception. The MDM consists of Semantic to Instance (S
2I) and Istance to Semantic (I2S). S2I is guided by the precise boundaries of se
mantic regions to learn the association between annotated points and instance co
ntours. I2S leverages discriminative relationships between instances to facilita
te the differentiation of various objects within the semantic map. Extensive exp
eriments substantiate the efficacy of MDM in fostering the synergy between insta
nce and semantic information, consequently improving the quality of instance-lev
el object representations. Our method achieves 55.7 mAP50 and 17.6 mAP on the PA
SCAL VOC and MS COCO datasets, significantly outperforming recent PSIS methods a
nd several box-supervised instance segmentation competitors.
****
Yaxuan Zhu,Jianwen Xie,Ying Nian Wu,Ruiqi Gao
Learning Energy-Based Models by Cooperative Diffusion Recovery Likelihood
Training energy-based models (EBMs) on high-dimensional data can be both challen
ging and time-consuming, and there exists a noticeable gap in sample quality bet
ween EBMs and other generative frameworks like GANs and diffusion models. To clo
se this gap, inspired by the recent efforts of learning EBMs by maximimizing dif
fusion recovery likelihood (DRL), we propose cooperative diffusion recovery like
lihood (CDRL), an effective approach to tractably learn and sample from a series
 of EBMs defined on increasingly noisy versons of a dataset, paired with an init
ializer model for each EBM. At each noise level, the two models are jointly esti
mated within a cooperative training framework: Samples from the initializer serv
e as starting points that are refined by a few MCMC sampling steps from the EBM.
 The EBM is then optimized by maximizing recovery likelihood, while the initiali
zer model is optimized by learning from the difference between the refined sampl
es and the initial samples. In addition, we made several practical designs for E
BM training to further improve the sample quality. Combining these advances, we
significantly boost the generation performance compared to existing EBM methods
on CIFAR-10 and ImageNet 32x32. And we have shown that CDRL has great potential
to largely reduce the sampling time. We also demonstrate the effectiveness of ou
r models for several downstream tasks, including classifier-free guided generati
on, compositional generation, image inpainting and out-of-distribution detection
.
****
Ganesh Ramachandra Kini,Vala Vakilian,Tina Behnia,Jaidev Gill,Christos Thrampoul
idis
Symmetric Neural-Collapse Representations with Supervised Contrastive Loss: The
Impact of ReLU and Batching
Supervised contrastive loss (SCL) is a competitive and often superior alternativ
e to the cross-entropy loss for classification. While prior studies have demonst
rated that both losses yield symmetric training representations under balanced d
ata, this symmetry breaks under class imbalances. This paper presents an intrigu
ing discovery: the introduction of a ReLU activation at the final layer effectiv
ely restores the symmetry in SCL-learned representations. We arrive at this find
ing analytically, by establishing that the global minimizers of an unconstrained
 features model with SCL loss and entry-wise non-negativity constraints form an
orthogonal frame. Extensive experiments conducted across various datasets, archi
tectures, and imbalance scenarios corroborate our finding. Importantly, our exp
eriments reveal that the inclusion of the ReLU activation restores symmetry with
out compromising test accuracy. This constitutes the first geometry characteriza

tion of SCL under imbalances. Additionally, our analysis and experiments undersc
ore the pivotal role of batch selection strategies in representation geometry. B
y proving necessary and sufficient conditions for mini-batch choices that ensure
 invariant symmetric representations, we introduce batch-binding as an efficient
 strategy that guarantees these conditions hold.
****
Runtian Zhai,Bingbin Liu,Andrej Risteski,J Zico Kolter,Pradeep Kumar Ravikumar
Understanding Augmentation-based Self-Supervised Representation Learning via RKH
S Approximation and Regression
Data augmentation is critical to the empirical success of modern self-supervised
 representation learning, such as contrastive learning and masked language model
ing.
However, a theoretical understanding of the exact role of the augmentation remai
ns limited.
Recent work has built the connection between self-supervised learning and the ap
proximation of the top eigenspace of a graph Laplacian operator, suggesting that
 learning a linear probe atop such representation can be connected to RKHS regre
ssion.
Building on this insight, this work delves into a statistical analysis of augmen
tation-based pretraining.
Starting from the isometry property, a geometric characterization of the target
function given by the augmentation, we disentangle the effects of the model and
the augmentation,
and prove two generalization bounds that are free of model complexity.
Our first bound works for an arbitrary encoder, and it is the sum of an estimati
on error bound incurred by fitting a linear probe, and an approximation error bo
und by RKHS approximation.
Our second bound specifically addresses the case
where the encoder extracts the top-d eigenspace of a finite-sample-based approxi
mation of the underlying RKHS.
A key ingredient in our analysis is the *augmentation complexity*,
which we use to quantitatively compare different augmentations and analyze their
 impact on downstream performance.
****
Eric Todd,Millicent Li,Arnab Sen Sharma,Aaron Mueller,Byron C Wallace,David Bau
Function Vectors in Large Language Models
We report the presence of a simple neural mechanism that represents an input-out
put function as a vector within autoregressive transformer language models (LMs)
. Using causal mediation analysis on a diverse range of in-context-learning (ICL
) tasks, we find that a small number attention heads transport a compact represe
ntation of the demonstrated task, which we call a function vector (FV). FVs are
 robust to changes in context, i.e., they trigger execution of the task on input
s such as zero-shot and natural text settings that do not resemble the ICL conte
xts from which they are collected. We test FVs across a range of tasks, models,
and layers and find strong causal effects across settings in middle layers. We i
nvestigate the internal structure of FVs and find while that they often contain
information that encodes the output space of the function, this information alon
e is not sufficient to reconstruct an FV. Finally, we test semantic vector compo
sition in FVs, and find that to some extent they can be summed to create vectors
 that trigger new complex tasks. Our findings show that compact, causal internal
 vector representations of function abstractions can be explicitly extracted fro
m LLMs.
****
Hattie Zhou,Arwen Bradley,Etai Littwin,Noam Razin,Omid Saremi,Joshua M. Susskind
,Samy Bengio,Preetum Nakkiran
What Algorithms can Transformers Learn? A Study in Length Generalization
Large language models exhibit surprising emergent generalization properties, yet
 also struggle on many simple reasoning tasks such as arithmetic and parity. In
this work, we focus on length generalization, and we propose a unifying framewor
k to understand when and how Transformers can be expected to length generalize o

n a given task. First, we show that there exist algorithmic tasks for which stan
dard
decoder-only Transformers trained from scratch naturally exhibit strong length g
eneralization. For these tasks, we leverage the RASP programming language (Weiss
 et al., 2021) to show that the correct algorithmic solution which solves the ta
sk can be represented by a simple Transformer. We thus propose the RASP-Generali
zation Conjecture: Transformers tend to learn a length-generalizing solution if
there exists a short RASP-L program that works for all input lengths. We present
 empirical evidence to support the correlation between RASP-simplicity and gener
alization. We leverage our insights to give new scratchpad formats which yield s
trong length generalization on traditionally hard tasks (such as parity and addi
tion), and we illustrate how scratchpad can hinder generalization when it increa
ses the complexity of the corresponding RASP-L program. Overall, our work provid
es a novel perspective on the mechanisms of length generalization and the algori
thmic capabilities of Transformers.
****
Raman Dutt,Ondrej Bohdal,Sotirios A. Tsaftaris,Timothy Hospedales
FairTune: Optimizing Parameter Efficient Fine Tuning for Fairness in Medical Ima
ge Analysis
Training models with robust group fairness properties is crucial in ethically se
nsitive application areas such as medical diagnosis. Despite the growing body of
 work aiming to minimise demographic bias in AI, this problem remains challengin
g. A key reason for this challenge is the fairness generalisation gap: High-capa
city deep learning models can fit all training data nearly perfectly, and thus a
lso exhibit perfect fairness during training. In this case, bias emerges only du
ring testing when generalisation performance differs across sub-groups. This mot
ivates us to take a bi-level optimisation perspective on fair learning: Optimisi
ng the learning strategy based on validation fairness. Specifically, we consider
 the highly effective workflow of adapting pre-trained models to downstream medi
cal imaging tasks using parameter-efficient fine-tuning (PEFT) techniques. There
 is a trade-off between updating more parameters, enabling a better fit to the t
ask of interest vs. fewer parameters, potentially reducing the generalisation ga
p. To manage this tradeoff, we propose FairTune, a framework to optimise the cho
ice of PEFT parameters with respect to fairness. We demonstrate empirically that
 FairTune leads to improved fairness on a range of medical imaging datasets. The
 code is available at https://github.com/Raman1121/FairTune.
****
Jifan Yu,Xiaozhi Wang,Shangqing Tu,Shulin Cao,Daniel Zhang-Li,Xin Lv,Hao Peng,Zi
jun Yao,Xiaohan Zhang,Hanming Li,Chunyang Li,Zheyuan Zhang,Yushi Bai,Yantao Liu,
Amy Xin,Kaifeng Yun,Linlu GONG,Nianyi Lin,Jianhui Chen,Zhili Wu,Yunjia Qi,Weikai
 Li,Yong Guan,Kaisheng Zeng,Ji Qi,Hailong Jin,Jinxin Liu,Yu Gu,Yuan Yao,Ning Din
g,Lei Hou,Zhiyuan Liu,Xu Bin,Jie Tang,Juanzi Li
KoLA: Carefully Benchmarking World Knowledge of Large Language Models
The unprecedented performance of large language models (LLMs) necessitates impro
vements in evaluations. Rather than merely exploring the breadth of LLM abilitie
s, we believe meticulous and thoughtful designs are essential to thorough, unbia
sed, and applicable evaluations. Given the importance of world knowledge to LLMs
, we construct a Knowledge-oriented LLM Assessment benchmark (KoLA), in which we
 carefully design three crucial factors: (1) For ability modeling, we mimic huma
n cognition to form a four-level taxonomy of knowledge-related abilities, coveri
ng 19 tasks. (2) For data, to ensure fair comparisons, we use both Wikipedia, a
corpus prevalently pre-trained by LLMs, along with continuously collected emergi
ng corpora, aiming to evaluate the capacity to handle unseen data and evolving k
nowledge. (3) For evaluation criteria, we adopt a contrastive system, including
overall standard scores for better numerical comparability across tasks and mode
ls, and a unique self-contrast metric for automatically evaluating knowledge-cre
ating ability. We evaluate 21 open-source and commercial LLMs and obtain some in
triguing findings. The KoLA dataset will be updated every three months to provid
e timely references for developing LLMs and knowledge-related systems.
****

Michael Kleinman,Alessandro Achille,Stefano Soatto
Critical Learning Periods Emerge Even in Deep Linear Networks
Critical learning periods are periods early in development where temporary sensory deficits can have a permanent effect on behavior and learned representations.

Despite the radical differences between biological and artificial networks, critical learning periods have been empirically observed in both systems. This suggests that critical periods may be fundamental to learning and not an accident of biology.
Yet, why exactly critical periods emerge in deep networks is still an open question, and in particular it is unclear whether the critical periods observed in both systems depend on particular architectural or optimization details. To isolate the key underlying factors, we focus on deep linear network models, and show that, surprisingly, such networks also display much of the behavior seen in biology and artificial networks, while being amenable to analytical treatment. We show that critical periods depend on the depth of the model and structure of the data distribution. We also show analytically and in simulations that the learning of features is tied to competition between sources. Finally, we extend our analysis to multi-task learning to show that pre-training on certain tasks can damage the transfer performance on new tasks, and show how this depends on the relationship between tasks and the duration of the pre-training stage. To the best of our knowledge, our work provides the first analytically tractable model that sheds light into why critical learning periods emerge in biological and artificial networks.
****
Wenxuan Li,Alan Yuille,Zongwei Zhou
How Well Do Supervised Models Transfer to 3D Image Segmentation?
The pre-training and fine-tuning paradigm has become prominent in transfer learning. For example, if the model is pre-trained on ImageNet and then fine-tuned to PASCAL, it can significantly outperform that trained directly on PASCAL. While ImageNet pre-training has shown enormous success, it is formed in 2D and the learned features are for classification tasks. Therefore, when transferring to more diverse tasks, like 3D image segmentation, its performance is inevitably compromised due to the deviation from the original ImageNet context. A significant challenge lies in the lack of large, annotated 3D datasets rivaling the scale of ImageNet for model pre-training. To overcome this challenge, we make two contributions. Firstly, we construct ImageNetCT-9K that comprises 9,262 three-dimensional computed tomography (CT) volumes with high-quality, per-voxel annotations. Secondly, we develop a suite of models that is supervised pre-trained on our ImageNetCT-9K. Our preliminary analyses indicate that the model trained only with 20 CT volumes, 640 masks, and 40 GPU hours has a transfer learning ability similar to the model trained with 5,050 CT volumes and 1,152 GPU hours. More importantly, the transfer learning ability of supervised models can further scale up with larger annotated datasets (i.e., SPT), achieving significantly better performance than all existing 3D models, irrespective of their pre-training methodologies or sources. We hope this study can facilitate collective efforts in constructing larger 3D vision datasets and more releases of supervised pre-trained models. Our code is attached as supplementary and will be publicly available.
****
Jianzhe Lin,Maurice Diesendruck,Liang Du,Robin Abraham
BatchPrompt: Accomplish more with less
The ever-increasing token limits of large language models (LLMs) have enabled long context as input. Many LLMs are trained/fine-tuned to perform zero-shot/few-shot inference using instruction-based prompts. Crafting prompts for these LLMs typically requires the user to provide a detailed task description, demonstrations, and single example of context for inference. This regular prompt baseline is referred to as “SinglePrompt” in this paper. However, for NLP tasks where each data point for inference is not necessarily lengthy, the token count
for instructions and few-shot examples in the prompt may be considerably larger than that of the data point, resulting in lower token-resource utilization compa

red with encoder-based models like fine-tuned BERT. This cost-efficiency issue,
affecting inference speed and compute budget, counteracts the many benefits LLMs
 have to offer. This paper aims to alleviate the preceding problem by batching m
ultiple data points into a single prompt, a prompting strategy we refer to as “B
atchPrompt”. This strategy increases the “density” of data points, which in turn
 leads to improved token utilization. Applying BatchPrompt na ■■vely, however, i
s very challenging due to significant performance degradation, as observed in ou
r experiments. We also noticed varying inference outcomes for the same data poin
ts appearing in different positions within a prompt. Based on this observation,
to address the quality issue while remain high token-resource utilization, we in
troduce Batch Permutation and Ensembling (BPE) for BatchPrompt, a simple majorit
y voting way that recovers labeling quality through repeatedly permutating data
positions in a batch at the price of more token usage. To counterbalance the add
itional token usage caused by the voting process, we further propose Self-reflec
tion-guided EArly Stopping (SEAS), which can terminate the voting process early
for data points the LLM confidently handles. Our comprehensive experimental eval
uation demonstrates that BPE +SEAS can boost the performance of BatchPrompt with
 a striking margin on a range of popular NLP tasks, including question answering
 (Boolq), textual entailment (RTE), and duplicate questions identification (QQP)
. These performances are even competitive with/higher than single-data prompting
 (SinglePrompt), while BatchPrompt requires much fewer LLM calls and input token
s (For SinglePrompt v.s. BatchPrompt+BPE +SEAS with batch size 32, using just 15
.7% the number of LLM calls, Boolq accuracy 90.6% $\rightarrow$ 90.9% with 27.4% tokens, QQP
 accuracy 87.2% $\rightarrow$ 88.4% with 18.6% tokens, RTE accuracy 91.5% $\rightarrow$ 91.1% with 30.8%
 tokens). We hope our simple yet effective approach will shed light on the futur
e research of large language models. The code will be released.
****
Jenny Zhang,Joel Lehman,Kenneth Stanley,Jeff Clune
OMNI: Open-endedness via Models of human Notions of Interestingness
Open-ended algorithms aim to learn new, interesting behaviors forever. That requ
ires a vast environment search space, but there are thus infinitely many possibl
e tasks. Even after filtering for tasks the current agent can learn (i.e., learn
ing progress), countless learnable yet uninteresting tasks remain (e.g., minor v
ariations of previously learned tasks). An Achilles Heel of open-endedness resea
rch is the inability to quantify (and thus prioritize) tasks that are not just l
earnable, but also $\textit{interesting}$ (e.g., worthwhile and novel). We propo
se solving this problem by $\textit{Open-endedness via Models of human Notions o
f Interestingness}$ (OMNI). The insight is that we can utilize foundation models
 (FMs) as a model of interestingness (MoI), because they $\textit{already}$ inte
rnalize human concepts of interestingness from training on vast amounts of human
-generated data, where humans naturally write about what they find interesting o
r boring. We show that FM-based MoIs improve open-ended learning by focusing on
tasks that are both learnable $\textit{and interesting}$, outperforming baseline
s based on uniform task sampling or learning progress alone. This approach has t
he potential to dramatically advance the ability to intelligently select which t
asks to focus on next (i.e., auto-curricula), and could be seen as AI selecting
its own next task to learn, facilitating self-improving AI and AI-Generating Alg
orithms.
****
Michihiro Yasunaga,Xinyun Chen,Yujia Li,Panupong Pasupat,Jure Leskovec,Percy Lia
ng,Ed H. Chi,Denny Zhou
Large Language Models as Analogical Reasoners
Chain-of-thought (CoT) prompting for language models demonstrates impressive per
formance across reasoning tasks, but typically needs labeled exemplars of the re
asoning process. In this work, we introduce a new prompting approach, analogical
 prompting, designed to automatically guide the reasoning process of large langu
age models. Inspired by analogical reasoning, a cognitive process in which human
s draw from relevant past experiences to tackle new problems, our approach promp
ts language models to self-generate relevant exemplars or knowledge in the conte
xt, before proceeding to solve the given problem. This method presents several a

dvantages: it obviates the need for labeling or retrieving exemplars, offering g enerality and convenience; it can also tailor the generated exemplars and knowle dge to each problem, offering adaptability. Experimental results show that our a pproach outperforms 0-shot CoT and manual few-shot CoT in a variety of reasoning tasks, including math problem solving in GSM8K and MATH, code generation in Cod eforces, and other reasoning tasks in BIG-Bench.
****
Aakash Lahoti,Stefani Karp,Ezra Winston,Aarti Singh,Yuanzhi Li
Role of Locality and Weight Sharing in Image-Based Tasks: A Sample Complexity Se paration between CNNs, LCNs, and FCNs
Vision tasks are characterized by the properties of locality and translation inv ariance.

The superior performance of convolutional neural networks (CNNs) on these ta sks is widely attributed to the inductive bias of locality and weight sharing ba ked into their architecture.

Existing attempts to quantify the statistical benefits of these biases in CN Ns over locally connected convolutional neural networks (LCNs) and fully connect ed neural networks (FCNs) fall into one of the following categories: either they disregard the optimizer and only provide uniform convergence upper bounds with no separating lower bounds,

or they consider simplistic tasks that do not truly mirror the locality and translation invariance as found in real-world vision tasks.

To address these deficiencies, we introduce the Dynamic Signal Distribution (DSD) classification task that models an image as consisting of k patches, eac h of dimension d, and the label is determined by a d-sparse signal vector th at can freely appear in any one of the k patches.

On this task, for any orthogonally equivariant algorithm like gradient desce nt, we prove that CNNs require $\tilde{O}(k+d)$ samples, whereas LCNs require $\Omega(kd)$ samples, establishing the statistical advantages of weight sharing in translation invariant tasks.

Furthermore, LCNs need $\tilde{O}(k(k+d))$ samples, compared to $\Omega(k^2d)$ samples for FCNs, showcasing the benefits of locality in local tasks.

Additionally, we develop information theoretic tools for analyzing randomize d algorithms, which may be of interest for statistical research.
****
Xinyue Liu,Hualin Zhang,Bin Gu,Hong Chen
General Stability Analysis for Zeroth-Order Optimization Algorithms
Zeroth-order optimization algorithms are widely used for black-box optimization problems, such as those in machine learning and prompt engineering, where the gr adients are approximated using function evaluations. Recently, a generalization result was provided for zeroth-order stochastic gradient descent (SGD) algorithm s through stability analysis. However, this result was limited to the vanilla 2-point zeroth-order estimate of Gaussian distribution used in SGD algorithms. To address these limitations, we propose a general proof framework for stability an alysis that applies to convex, strongly convex, and non-convex conditions, and y ields results for popular zeroth-order optimization algorithms, including SGD, G D, and SVRG, as well as various zeroth-order estimates, such as 1-point and 2-po int with different distributions and coordinate estimates. Our general analysis shows that coordinate estimation can lead to tighter generalization bounds for S GD, GD, and SVRG versions of zeroth-order optimization algorithms, due to the sm aller expansion brought by coordinate estimates to stability analysis.
****
Zhen Liu,Yao Feng,Yuliang Xiu,Weiyang Liu,Liam Paull,Michael J. Black,Bernhard S chölkopf
Ghost on the Shell: An Expressive Representation of General 3D Shapes
The creation of photorealistic virtual worlds requires the accurate modeling of 3D surface geometry for a wide range of objects. For this, meshes are appealing since they enable 1) fast physics-based rendering with realistic material and li ghting, 2) physical simulation, and 3) are memory-efficient for modern graphics pipelines. Recent work on reconstructing and statistically modeling 3D shape, ho

wever, has critiqued meshes as being topologically inflexible. To capture a wide range of object shapes, any 3D representation must be able to model solid, watertight, shapes as well as thin, open, surfaces. Recent work has focused on the former, and methods for reconstructing open surfaces do not support fast reconstruction with material and lighting or unconditional generative modelling. Inspired by the observation that open surfaces can be seen as islands floating on watertight surfaces, we parametrize open surfaces by defining a manifold signed distance field on watertight templates. With this parametrization, we further develop a grid-based and differentiable representation that parametrizes both watertight and non-watertight meshes of arbitrary topology. Our new representation, called Ghost-on-the-Shell (G-Shell), enables two important applications: differentiable rasterization-based reconstruction from multiview images and generative modelling of non-watertight meshes. We empirically demonstrate that G-Shell achieves state-of-the-art performance on non-watertight mesh reconstruction and generation tasks, while also performing effectively for watertight meshes.
****
Guodong Wang,Yunhong Wang,Xiuguo Bao,Di Huang
Rotation has two sides: Evaluating Data Augmentation for Deep One-class Classification
One-class classification (OCC) involves predicting whether a new data is normal or anomalous based solely on the data from a single class during training. Various attempts have been made to learn suitable representations for OCC within a self-supervised framework. Notably, discriminative methods that use geometric visual transformations, such as rotation, to generate pseudo-anomaly samples have exhibited impressive detection performance. Although rotation is commonly viewed as a distribution-shifting transformation and is widely used in the literature, its effectiveness remains a mystery. In this study, we make a surprising observation: there exists a strong linear relationship (Pearson's Correlation, $r > 0.9$) between the accuracy of rotation prediction and the performance of OCC. This suggests that a classifier that effectively distinguishes different rotations is more likely to excel in OCC, and vice versa. The root cause of this phenomenon can be attributed to the transformation bias in the dataset, where representations learned from transformations already present in the dataset tend to be less effective, making it essential to accurately estimate the transformation distribution before utilizing pretext tasks involving these transformations for reliable self-supervised representation learning. To the end, we propose a novel two-stage method to estimate the transformation distribution within the dataset. In the first stage, we learn general representations through standard contrastive pre-training. In the second stage, we select potentially semantics-preserving samples from the entire augmented dataset, which includes all rotations, by employing density matching with the provided reference distribution. By sorting samples based on semantics-preserving versus shifting transformations, we achieve improved performance on OCC benchmarks.
****
Xuheng Li,Yihe Deng,Jingfeng Wu,Dongruo Zhou,Quanquan Gu
Risk Bounds of Accelerated SGD for Overparameterized Linear Regression
Accelerated stochastic gradient descent (ASGD) is a workhorse in deep learning and often achieves better generalization performance than SGD. However, existing optimization theory can only explain the faster convergence of ASGD, but cannot explain its better generalization. In this paper, we study the generalization of ASGD for overparameterized linear regression, which is possibly the simplest setting of learning with overparameterization. We establish an instance-dependent excess risk bound for ASGD within each eigen-subspace of the data covariance matrix. Our analysis shows that (i) ASGD outperforms SGD in the subspace of small eigenvalues, exhibiting a faster rate of exponential decay for bias error, while in the subspace of large eigenvalues, its bias error decays slower than SGD; and (ii) the variance error of ASGD is always larger than that of SGD. Our result suggests that ASGD can outperform SGD when the difference between the initialization and the true weight vector is mostly confined to the subspace of small eigenvalues. Additionally, when our analysis is specialized to linear regression in t

he strongly convex setting, it yields a tighter bound for bias error than the be
st-known result.
****
Eran Rosenbluth,Jan Tönshoff,Martin Ritzert,Berke Kisin,Martin Grohe
Distinguished In Uniform: Self-Attention Vs. Virtual Nodes
Graph Transformers (GTs) such as SAN and GPS are graph processing models that co
mbine Message-Passing GNNs (MPGNNs) with global Self-Attention. They were shown
to be universal function approximators, with two reservations: 1. The initial no
de features must be augmented with certain positional encodings. 2. The approxim
ation is non-uniform: Graphs of different sizes may require a different approxim
ating network.

We first clarify that this form of universality is not unique to GTs: Using the
same positional encodings, also pure MPGNNs and even 2-layer MLPs are non-unifor
m universal approximators. We then consider uniform expressivity: The target fun
ction is to be approximated by a single network for graphs of all sizes. There,
we compare GTs to the more efficient MPGNN + Virtual Node architecture. The esse
ntial difference between the two model definitions is in their global computatio
n method: Self-Attention Vs Virtual Node. We prove that none of the models is a
uniform-universal approximator, before proving our main result: Neither model’s
uniform expressivity subsumes the other’s. We demonstrate the theory with experi
ments on synthetic data. We further augment our study with real-world datasets,
observing mixed results which indicate no clear ranking in practice as well.
****
Fran Jeleni■,Josip Juki■,Martin Tutek,Mate Puljiz,Jan Snajder
Out-of-Distribution Detection by Leveraging Between-Layer Transformation Smoothn
ess
Effective out-of-distribution (OOD) detection is crucial for reliable machine le
arning models, yet most current methods are limited in practical use due to requ
irements like access to training data or intervention in training. We present a
novel method for detecting OOD data in Transformers based on transformation smoo
thness between intermediate layers of a network (BLOOD), which is applicable to
pre-trained models without access to training data. BLOOD utilizes the tendency
of between-layer representation transformations of in-distribution (ID) data to
be smoother than the corresponding transformations of OOD data, a property that
we also demonstrate empirically. We evaluate BLOOD on several text classificatio
n tasks with Transformer networks and demonstrate that it outperforms methods wi
th comparable resource requirements. Our analysis also suggests that when learni
ng simpler tasks, OOD data transformations maintain their original sharpness, wh
ereas sharpness increases with more complex tasks.
****
Huanran Chen,Yichi Zhang,Yinpeng Dong,Xiao Yang,Hang Su,Jun Zhu
Rethinking Model Ensemble in Transfer-based Adversarial Attacks
It is widely recognized that deep learning models lack robustness to adversarial
 examples. An intriguing property of adversarial examples is that they can trans
fer across different models, which enables black-box attacks without any knowled
ge of the victim model. An effective strategy to improve the transferability is
attacking an ensemble of models. However, previous works simply average the outp
uts of different models, lacking an in-depth analysis on how and why model ensem
ble methods can strongly improve the transferability. In this paper, we rethink
the ensemble in adversarial attacks and define the common weakness of model ense
mble with two properties: 1) the flatness of loss landscape; and 2) the closenes
s to the local optimum of each model. We empirically and theoretically show that
 both properties are strongly correlated with the transferability and propose a
Common Weakness Attack (CWA) to generate more transferable adversarial examples
by promoting these two properties. Experimental results on both image classifica
tion and object detection tasks validate the effectiveness of our approach to im
proving the adversarial transferability, especially when attacking adversarially
 trained models. We also successfully apply our method to attack a black-box lar
ge vision-language model -- Google's Bard, showing the practical effectiveness.

Code is available at \url{https://github.com/huanranchen/AdversarialAttacks}.
****
Namyong Park,Xing Wang,Antoine Simoulin,Shuai Yang,Grey Yang,Ryan A. Rossi,Puja Trivedi,Nesreen K. Ahmed
Forward Learning of Graph Neural Networks
Graph neural networks (GNNs) have achieved remarkable success across a wide range of applications, such as recommendation, drug discovery, and question answering. Behind the success of GNNs lies the backpropagation (BP) algorithm, which is the de facto standard for training deep neural networks (NNs). However, despite its effectiveness, BP imposes several constraints, which are not only biologically implausible, but also limit the scalability, parallelism, and flexibility in learning NNs. Examples of such constraints include storage of neural activities computed in the forward pass for use in the subsequent backward pass, and the dependence of parameter updates on non-local signals. To address these limitations, the forward-forward algorithm (FF) was recently proposed as an alternative to BP in the image classification domain, which trains NNs by performing two forward passes over positive and negative data. Inspired by this advance, we propose ForwardGNN in this work, a new forward learning procedure for GNNs, which avoids the constraints imposed by BP via an effective layer-wise local forward training. ForwardGNN extends the original FF to deal with graph data and GNNs, and makes it possible to operate without generating negative inputs (hence no longer forward-forward). Further, ForwardGNN enables each layer to learn from both the bottom-up and top-down signals without relying on the backpropagation of errors. Extensive experiments on real-world datasets show the effectiveness and generality of the proposed forward graph learning framework. We release our code at https://github.com/facebookresearch/forwardgnn.
****
Pierre Marion,Yu-Han Wu,Michael Eli Sander,Gérard Biau
Implicit regularization of deep residual networks towards neural ODEs
Residual neural networks are state-of-the-art deep learning models. Their continuous-depth analog, neural ordinary differential equations (ODEs), are also widely used. Despite their success, the link between the discrete and continuous models still lacks a solid mathematical foundation. In this article, we take a step in this direction by establishing an implicit regularization of deep residual networks towards neural ODEs, for nonlinear networks trained with gradient flow. We prove that if the network is initialized as a discretization of a neural ODE, then such a discretization holds throughout training. Our results are valid for a finite training time, and also as the training time tends to infinity provided that the network satisfies a Polyak-■ojasiewicz condition. Importantly, this condition holds for a family of residual networks where the residuals are two-layer perceptrons with an overparameterization in width that is only linear, and implies the convergence of gradient flow to a global minimum. Numerical experiments illustrate our results.
****
Renyu Zhang,Aly A Khan,Yuxin Chen,Robert L. Grossman
Enhancing Instance-Level Image Classification with Set-Level Labels
Instance-level image classification tasks have traditionally relied on single-instance labels to train models, e.g., few-shot learning and transfer learning. However, set-level coarse-grained labels that capture relationships among instances can provide richer information in real-world scenarios. In this paper, we present a novel approach to enhance instance-level image classification by leveraging set-level labels. We provide a theoretical analysis of the proposed method, including recognition conditions for fast excess risk rate, shedding light on the theoretical foundations of our approach. We conducted experiments on two distinct categories of datasets: natural image datasets and histopathology image datasets. Our experimental results demonstrate the effectiveness of our approach, showcasing improved classification performance compared to traditional single-instance label-based methods. Notably, our algorithm achieves 13\% improvement in classification accuracy compared to the strongest baseline on the histopathology image classification benchmarks. Importantly, our experimental findings align with

the theoretical analysis, reinforcing the robustness and reliability of our proposed method. This work bridges the gap between instance-level and set-level image classification, offering a promising avenue for advancing the capabilities of image classification models with set-level coarse-grained labels.
****
Yeongwoo Song,Hawoong Jeong
Towards Cross Domain Generalization of Hamiltonian Representation via Meta Learning
Recent advances in deep learning for physics have focused on discovering shared representations of target systems by incorporating physics priors or inductive biases into neural networks. While effective, these methods are limited to the system domain, where the type of system remains consistent and thus cannot ensure the adaptation to new, or unseen physical systems governed by different laws. For instance, a neural network trained on a mass-spring system cannot guarantee accurate predictions for the behavior of a two-body system or any other system with different physical laws.
In this work, we take a significant leap forward by targeting cross domain generalization within the field of Hamiltonian dynamics.
We model our system with a graph neural network (GNN) and employ a meta learning algorithm to enable the model to gain experience over a distribution of systems and make it adapt to new physics. Our approach aims to learn a unified Hamiltonian representation that is generalizable across multiple system domains, thereby overcoming the limitations of system-specific models.
We demonstrate that the meta-trained model captures the generalized Hamiltonian representation that is consistent across different physical domains.
Overall, through the use of meta learning, we offer a framework that achieves cross domain generalization, providing a step towards a unified model for understanding a wide array of dynamical systems via deep learning.
****
Zhou Lu,Qiuyi Zhang,Xinyi Chen,Fred Zhang,David Woodruff,Elad Hazan
Adaptive Regret for Bandits Made Possible: Two Queries Suffice
Fast changing states or volatile environments pose a significant challenge to online optimization, which needs to perform rapid adaptation under limited observation. In this paper, we give query and regret optimal bandit algorithms under the strict notion of strongly adaptive regret, which measures the maximum regret over any contiguous interval I. Due to its worst-case nature, there is an almost-linear $\Omega(|I|^{1-\epsilon})$ regret lower bound, when only one query per round is allowed [Daniely el al, ICML 2015]. Surprisingly, with just two queries per round, we give Strongly Adaptive Bandit Learner (StABL) that achieves $\widetilde{O}(\sqrt{n|I|})$ adaptive regret for multi-armed bandits with n arms. The bound is tight and cannot be improved in general. Our algorithm leverages a multiplicative update scheme of varying stepsizes and a carefully chosen observation distribution to control the variance. Furthermore, we extend our results and provide optimal algorithms in the bandit convex optimization setting. Finally, we empirically demonstrate the superior performance of our algorithms under volatile environments and for downstream tasks, such as algorithm selection for hyperparameter optimization.
****
Ruizhe Shi,Yuyao Liu,Yanjie Ze,Simon Shaolei Du,Huazhe Xu
Unleashing the Power of Pre-trained Language Models for Offline Reinforcement Learning
Offline reinforcement learning (RL) aims to find a near-optimal policy using pre-collected datasets. Given recent advances in Large Language Models (LLMs) and their few-shot learning prowess, this paper introduces $\textbf{La}$nguage Models for $\textbf{Mo}$tion Control ($\textbf{LaMo}$), a general framework based on Decision Transformers to effectively use pre-trained Language Models (LMs) for offline RL. Our framework highlights four crucial components: (1) Initializing Decision Transformers with sequentially pre-trained LMs, (2) employing the LoRA fine-tuning method, in contrast to full-weight fine-tuning, to combine the pre-trained knowledge from LMs and in-domain knowledge effectively, (3) using the non-l

inear MLP transformation instead of linear projections, to generate embeddings, and (4) integrating an auxiliary language prediction loss during fine-tuning to stabilize the LMs and retain their original abilities on languages. Empirical results indicate $\textbf{LaMo}$ achieves state-of-the-art performance in sparse-reward tasks and closes the gap between value-based offline RL methods and decision transformers in dense-reward tasks. In particular, our method demonstrates superior performance in scenarios with limited data samples.
****
Shikun Feng,Minghao Li,Yinjun Jia,Wei-Ying Ma,Yanyan Lan
Protein-ligand binding representation learning from fine-grained interactions
The binding between proteins and ligands plays a crucial role in the realm of drug discovery. Previous deep learning approaches have shown promising results over traditional computationally intensive methods, but resulting in poor generalization due to limited supervised data. In this paper, we propose to learn protein-ligand binding representation in a self-supervised learning manner. Different from existing pre-training approaches which treat proteins and ligands individually, we emphasize to discern the intricate binding patterns from fine-grained interactions. Specifically, this self-supervised learning problem is formulated as a prediction of the conclusive binding complex structure given a pocket and ligand with a Transformer based interaction module, which naturally emulates the binding process. To ensure the representation of rich binding information, we introduce two pre-training tasks, i.e. atomic pairwise distance map prediction and mask ligand reconstruction, which comprehensively model the fine-grained interactions from both structure and feature space. Extensive experiments have demonstrated the superiority of our method across various binding tasks, including protein-ligand affinity prediction, virtual screening and protein-ligand docking.
****
Jiaxin Lu,Zetian Jiang,Tianzhe Wang,Junchi Yan
M3C: A Framework towards Convergent, Flexible, and Unsupervised Learning of Mixture Graph Matching and Clustering
Existing graph matching methods typically assume that there are similar structures between graphs and they are matchable. This work addresses a more realistic scenario where graphs exhibit diverse modes, requiring graph grouping before or along with matching, a task termed mixture graph matching and clustering. Specifically, we introduce Minorize-Maximization Matching and Clustering (M3C), a learning-free algorithm that guarantees theoretical convergence through the Minorize-Maximization framework and offers enhanced flexibility via relaxed clustering. Building on M3C, we further develop UM3C, an unsupervised model that incorporates novel edge-wise affinity learning and pseudo label selection. Extensive experimental results on public benchmarks demonstrate that our method outperforms state-of-the-art graph matching and mixture graph matching and clustering approaches in both accuracy and efficiency.
****
Chaoming Wang,Tianqiu Zhang,Sichao He,Hongyaoxing Gu,Shangyang Li,Si Wu
A differentiable brain simulator bridging brain simulation and brain-inspired computing
Brain simulation builds dynamical models to mimic the structure and functions of the brain, while brain-inspired computing (BIC) develops intelligent systems by learning from the structure and functions of the brain. The two fields are intertwined and should share a common programming framework to facilitate each other's development. However, none of the existing software in the fields can achieve this goal, because traditional brain simulators lack differentiability for training, while existing deep learning (DL) frameworks fail to capture the biophysical realism and complexity of brain dynamics. In this paper, we introduce BrainPy, a differentiable brain simulator developed using JAX and XLA, with the aim of bridging the gap between brain simulation and BIC. BrainPy expands upon the functionalities of JAX, a powerful AI framework, by introducing complete capabilities for flexible, efficient, and scalable brain simulation. It offers a range of sparse and event-driven operators for efficient and scalable brain simulation, an abstraction for managing the intricacies of synaptic computations, a modular an

d flexible interface for constructing multi-scale brain models, and an object-or iented just-in-time compilation approach to handle the memory-intensive nature o f brain dynamics. We showcase the efficiency and scalability of BrainPy on bench mark tasks, and highlight its differentiable simulation for biologically plausib le spiking models.
****
Tianyu He,Junliang Guo,Runyi Yu,Yuchi Wang,jialiang zhu,Kaikai An,Leyi Li,Xu Tan ,Chunyu Wang,Han Hu,HsiangTao Wu,sheng zhao,Jiang Bian
GAIA: Zero-shot Talking Avatar Generation
Zero-shot talking avatar generation aims at synthesizing natural talking videos from speech and a single portrait image. Previous methods have relied on domain- specific heuristics such as warping-based motion representation and 3D Morphable Models, which limit the naturalness and diversity of the generated avatars. In this work, we introduce GAIA (Generative AI for Avatar), which eliminates the do main priors in talking avatar generation. In light of the observation that the s peech only drives the motion of the avatar while the appearance of the avatar an d the background typically remain the same throughout the entire video, we divid e our approach into two stages: 1) disentangling each frame into motion and appe arance representations; 2) generating motion sequences conditioned on the speech and reference portrait image. We collect a large-scale high-quality talking ava tar dataset and train the model on it with different scales (up to 2B parameters). Experimental results verify the superiority, scalability, and flexibility of GAIA as 1) the resulting model beats previous baseline models in terms of natura lness, diversity, lip-sync quality, and visual quality; 2) the framework is scal able since larger models yield better results; 3) it is general and enables diff erent applications like controllable talking avatar generation and text-instruct ed avatar generation.
****
Bobak Kiani,Thien Le,Hannah Lawrence,Stefanie Jegelka,Melanie Weber
On the hardness of learning under symmetries
We study the problem of learning equivariant neural networks via gradient descen t. The incorporation of known symmetries ("equivariance") into neural nets has empirically improved the performance of learning pipelines, in domains ranging f rom biology to computer vision. However, a rich yet separate line of learning th eoretic research has demonstrated that actually learning shallow, fully-connecte d (i.e. non-symmetric) networks has exponential complexity in the correlational statistical query (CSQ) model, a framework encompassing gradient descent. In thi s work, we ask: are known problem symmetries sufficient to alleviate the fundame ntal hardness of learning neural nets with gradient descent? We answer this ques tion in the negative. In particular, we give lower bounds for shallow graph neur al networks, convolutional networks, invariant polynomials, and frame-averaged n etworks for permutation subgroups, which all scale either superpolynomially or e xponentially in the relevant input dimension. Therefore, in spite of the signifi cant inductive bias imparted via symmetry, actually learning the complete classe s of functions represented by equivariant neural networks via gradient descent r emains hard.
****
Guan Wang,Sijie Cheng,Xianyuan Zhan,Xiangang Li,Sen Song,Yang Liu
OpenChat: Advancing Open-source Language Models with Mixed-Quality Data
Nowadays, open-source large language models like LLaMA have emerged. Recent deve lopments have incorporated supervised fine-tuning (SFT) and reinforcement learni ng fine-tuning (RLFT) to align these models with human goals. However, SFT metho ds treat all training data with mixed quality equally, while RLFT methods requir e high-quality pairwise or ranking-based preference data. In this study, we pres ent a novel framework, named OpenChat, to advance open-source language models wi th mixed-quality data. Specifically, we consider the general SFT training data, consisting of a small amount of expert data mixed with a large proportion of sub -optimal data, without any preference labels. We propose the C(onditioned)-RLFT, which regards different data sources as coarse-grained reward labels and learns a class-conditioned policy to leverage complementary data quality information.

Interestingly, the optimal policy in C-RLFT can be easily solved through single-stage, RL-free supervised learning, which is lightweight and avoids costly human preference labeling.
Through extensive experiments on three standard benchmarks, our openchat-13b fine-tuned with C-RLFT achieves the highest average performance among all 13b open-source language models. Moreover, we use AGIEval to validate the model generalization performance, in which only openchat-13b surpasses the base model. Finally, we conduct a series of analyses to shed light on the effectiveness and robustness of OpenChat. Our code, data, and models are publicly available at https://github.com/imoneoi/openchat and https://huggingface.co/openchat.
****
Zahra Kadkhodaie,Florentin Guth,Eero P Simoncelli,Stéphane Mallat
Generalization in diffusion models arises from geometry-adaptive harmonic representations
Deep neural networks (DNNs) trained for image denoising are able to generate high-quality samples with score-based reverse diffusion algorithms. These impressive capabilities seem to imply an escape from the curse of dimensionality, but recent reports of memorization of the training set raise the question of whether these networks are learning the "true" continuous density of the data. Here, we show that two DNNs trained on non-overlapping subsets of a dataset learn nearly the same score function, and thus the same density, when the number of training images is large enough. In this regime of strong generalization, diffusion-generated images are distinct from the training set, and are of high visual quality, suggesting that the inductive biases of the DNNs are well-aligned with the data density. We analyze the learned denoising functions and show that the inductive biases give rise to a shrinkage operation in a basis adapted to the underlying image. Examination of these bases reveals oscillating harmonic structures along contours and in homogeneous regions. We demonstrate that trained denoisers are inductively biased towards these geometry-adaptive harmonic bases since they arise not only when the network is trained on photographic images, but also when it is trained on image classes supported on low-dimensional manifolds for which the harmonic basis is suboptimal. Finally, we show that when trained on regular image classes for which the optimal basis is known to be geometry-adaptive and harmonic, the denoising performance of the networks is near-optimal.
****
Jiayi Wei,Greg Durrett,Isil Dillig
Coeditor: Leveraging Repo-level Diffs for Code Auto-editing
Developers often dedicate significant time to maintaining and refactoring existing code. However, most prior work on generative models for code focuses solely on creating new code, overlooking the distinctive needs of editing existing code. In this work, we explore a multi-round code auto-editing setting, aiming to predict edits to a code region based on recent changes within the same codebase. Our model, Coeditor, is a fine-tuned language model specifically designed for code editing tasks. We represent code changes using a line diff format and employ static analysis to form large customized model contexts, ensuring the availability of appropriate information for prediction. We collect a code editing dataset from the commit histories of 1650 open-source Python projects for training and evaluation. In a simplified single-round, single-edit task, Coeditor significantly outperforms GPT-3.5 and SOTA open-source code completion models (bringing exact-match accuracy from 34.7 up to 60.4), demonstrating the benefits of incorporating editing history for code completion. In a multi-round, multi-edit setting, we observe substantial gains by iteratively conditioning on additional user edits. We have open-sourced our code, data, and model weights to encourage future research and have released a VSCode extension powered by our model for interactive IDE usage.
****
Jinyang Jiang,Zeliang Zhang,Chenliang Xu,Zhaofei Yu,Yijie Peng
One Forward is Enough for Neural Network Training via Likelihood Ratio Method
While backpropagation (BP) is the mainstream approach for gradient computation in neural network training, its heavy reliance on the chain rule of differentiati

on constrains the designing flexibility of network architecture and training pip elines. We avoid the recursive computation in BP and develop a unified likelihoo d ratio (ULR) method for gradient estimation with only one forward propagation. Not only can ULR be extended to train a wide variety of neural network architect ures, but the computation flow in BP can also be rearranged by ULR for better de vice adaptation. Moreover, we propose several variance reduction techniques to f urther accelerate the training process. Our experiments offer numerical results across diverse aspects, including various neural network training scenarios, com putation flow rearrangement, and fine-tuning of pre-trained models. All findings demonstrate that ULR effectively enhances the flexibility of neural network tra ining by permitting localized module training without compromising the global ob jective and significantly boosts the network robustness.
****

Kesen Zhao,Liang Zhang
Causality-Inspired Spatial-Temporal Explanations for Dynamic Graph Neural Networ ks
Dynamic Graph Neural Networks (DyGNNs) have gained significant popularity in the research of dynamic graphs, but are limited by the low transparency, such that human-understandable insights can hardly be drawn from their predictions. Althou gh a number of existing research have been devoted to investigating the interpre tability of graph neural networks (GNNs), achieving the interpretability of DyGN Ns is pivotally challenging due to the complex spatial-temporal correlations in dynamic graphs. To this end, we propose an innovative causality-inspired generat ive model based on structural causal model (SCM), which explores the underlying philosophies of DyGNN predictions by identifying the trivial, static, and dynami c causal relationships. To reach this goal, two critical tasks need to be accomp lished including (1) disentangling the complex causal relationships, and (2) fit ting the spatial-temporal explanations of DyGNNs in the SCM architecture. To tac kle these challenges, the proposed method incorporates a contrastive learning mo dule to disentangle trivial and causal relationships, and a dynamic correlating module to disentangle dynamic and static causal relationships, respectively. A d ynamic VGAE-based framework is further developed, which generates causal-and-dyn amic masks for spatial interpretability, and recognizes dynamic relationships al ong the time horizon through causal invention for temporal interpretability. Com prehensive experiments have been conducted on both synthetic and real-world data sets, where our approach yields substantial improvements, thereby demonstrating significant superiority.
****

Haomin Zhuang,Mingxian Yu,Hao Wang,Yang Hua,Jian Li,Xu Yuan
Backdoor Federated Learning by Poisoning Backdoor-Critical Layers
Federated learning (FL) has been widely deployed to enable machine learning trai ning on sensitive data across distributed devices. However, the decentralized le arning paradigm and heterogeneity of FL further extend the attack surface for ba ckdoor attacks. Existing FL attack and defense methodologies typically focus on the whole model. None of them recognizes the existence of backdoor-critical (BC) layers-a small subset of layers that dominate the model vulnerabilities. Attack ing the BC layers achieves equivalent effects as attacking the whole model but a t a far smaller chance of being detected by state-of-the-art (SOTA) defenses. Th is paper proposes a general in-situ approach that identifies and verifies BC lay ers from the perspective of attackers. Based on the identified BC layers, we car efully craft a new backdoor attack methodology that adaptively seeks a fundament al balance between attacking effects and stealthiness under various defense stra tegies. Extensive experiments show that our BC layer-aware backdoor attacks can successfully backdoor FL under seven SOTA defenses with only 10% malicious clien ts and outperform the latest backdoor attack methods.
****

Xurui Li,Ziming Huang,Feng Xue,Yu Zhou
MuSc: Zero-Shot Industrial Anomaly Classification and Segmentation with Mutual S coring of the Unlabeled Images
This paper studies zero-shot anomaly classification (AC) and segmentation (AS) i

n industrial vision.
We reveal that the abundant normal and abnormal cues implicit in unlabeled test images can be exploited for anomaly determination, which is ignored by prior methods.
Our key observation is that for the industrial product images, the normal image patches could find a relatively large number of similar patches in other unlabeled images,
while the abnormal ones only have a few similar patches.
We leverage such a discriminative characteristic to design a novel zero-shot AC/AS method by Mutual Scoring (MuSc) of the unlabeled images,
which does not need any training or prompts.
Specifically, we perform Local Neighborhood Aggregation with Multiple Degrees (LNAMD) to obtain the patch features that are capable of representing anomalies in varying sizes.
Then we propose the Mutual Scoring Mechanism (MSM) to leverage the unlabeled test images to assign the anomaly score to each other.
Furthermore, we present an optimization approach named Re-scoring with Constrained Image-level Neighborhood (RsCIN) for image-level anomaly classification to suppress the false positives caused by noises in normal images.
The superior performance on the challenging MVTec AD and VisA datasets demonstrates the effectiveness of our approach.
Compared with the state-of-the-art zero-shot approaches,
MuSc achieves a $\textbf{21.1}$% PRO absolute gain (from 72.7\% to 93.8\%) on MVTec AD, a $\textbf{19.4}$% pixel-AP gain and a $\textbf{14.7}$% pixel-AUROC gain on VisA.
In addition, our zero-shot approach outperforms most of the few-shot approaches and is comparable to some one-class methods.
Code is available at https://github.com/xrli-U/MuSc.
****
Xin Zhang,Dong Zhang,Shimin Li,Yaqian Zhou,Xipeng Qiu
SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models
Current speech large language models build upon discrete speech representations, which can be categorized into semantic tokens and acoustic tokens. However, existing speech tokens are not specifically designed for speech language modeling. To assess the suitability of speech tokens for building speech language models, we established the first benchmark, SLMTokBench. Our results indicate that neither semantic nor acoustic tokens are ideal for this purpose. Therefore, we
propose SpeechTokenizer, a unified speech tokenizer for speech large language models. SpeechTokenizer adopts the Encoder-Decoder architecture with residual vector quantization (RVQ). Unifying semantic and acoustic tokens, SpeechTokenizer disentangles different aspects of speech information hierarchically across different RVQ layers. Furthermore, We construct a Unified Speech Language Model (USLM) leveraging SpeechTokenizer. Experiments show that SpeechTokenizer performs comparably to EnCodec in speech reconstruction and demonstrates strong performance on the SLMTokBench benchmark. Also, USLM outperforms VALL-E in zero-shot Text-to-Speech tasks. Code and models are available at https://github.com/ZhangXInFD/SpeechTokenizer/.
****
Jae-Woo Choi,Youngwoo Yoon,Hyobin Ong,Jaehong Kim,Minsu Jang
LoTa-Bench: Benchmarking Language-oriented Task Planners for Embodied Agents
Large language models (LLMs) have recently received considerable attention as alternative solutions for task planning. However, comparing the performance of language-oriented task planners becomes difficult, and there exists a dearth of detailed exploration regarding the effects of various factors such as pre-trained model selection and prompt construction. To address this, we propose a benchmark system for automatically quantifying performance of task planning for home-service embodied agents. Task planners are tested on two pairs of datasets and simulators: 1) ALFRED and AI2-THOR, 2) an extension of Watch-And-Help and VirtualHome. Using the proposed benchmark system, we perform extensive experiments with LLMs

and prompts, and explore several enhancements of the baseline planner. We expect that the proposed benchmark tool would accelerate the development of language-oriented task planners.
****
Jacob S. Prince,Gabriel Fajardo,George A. Alvarez,Talia Konkle
Manipulating dropout reveals an optimal balance of efficiency and robustness in biological and machine visual systems
According to the efficient coding hypothesis, neural populations encode information optimally when representations are high-dimensional and uncorrelated. However, such codes may carry a cost in terms of generalization and robustness. Past empirical studies of early visual cortex (V1) in rodents have suggested that this tradeoff indeed constrains sensory representations. However, it remains unclear whether these insights generalize across the hierarchy of the human visual system, and particularly to object representations in high-level occipitotemporal cortex (OTC). To gain new empirical clarity, here we develop a family of object recognition models with parametrically varying dropout proportion p, which induces systematically varying dimensionality of internal responses (while controlling all other inductive biases). We find that increasing dropout produces an increasingly smooth, low-dimensional representational space. Optimal robustness to lesioning is observed at around 70% dropout, after which both accuracy and robustness decline. Representational comparison to large-scale 7T fMRI data from occipitotemporal cortex in the Natural Scenes Dataset reveals that this optimal degree of dropout is also associated with maximal emergent neural predictivity. Finally, using new techniques for achieving denoised estimates of the eigenspectrum of human fMRI responses, we compare the rate of eigenspectrum decay between model and brain feature spaces. We observe that the match between model and brain representations is associated with a common balance between efficiency and robustness in the representational space. These results suggest that varying dropout may reveal an optimal point of balance between the efficiency of high-dimensional codes and the robustness of low dimensional codes in hierarchical vision systems.
****
Changbin Li,Kangshuo Li,Yuzhe Ou,Lance M. Kaplan,Audun Jøsang,Jin-Hee Cho,DONG HYUN JEONG,Feng Chen
Hyper Evidential Deep Learning to Quantify Composite Classification Uncertainty
Deep neural networks (DNNs) have been shown to perform well on exclusive, multi-class classification tasks. However, when different classes have similar visual features, it becomes challenging for human annotators to differentiate them. When an image is ambiguous, such as a blurry one where an annotator can't distinguish between a husky and a wolf, it may be labeled with both classes: {husky, wolf}. This scenario necessitates the use of composite set labels.
In this paper, we propose a novel framework called Hyper-Evidential Neural Network (HENN) that explicitly models predictive uncertainty caused by composite set labels in training data in the context of the belief theory called Subjective Logic (SL).
By placing a Grouped Dirichlet distribution on the class probabilities, we treat predictions of a neural network as parameters of hyper-subjective opinions and learn the network that collects both single and composite evidence leading to these hyper-opinions by a deterministic DNN from data.
We introduce a new uncertainty type called vagueness originally designed for hyper-opinions in SL to quantify composite classification uncertainty for DNNs.
Our experiments prove that HENN outperforms its state-of-the-art counterparts based on four image datasets.
The code and datasets are available at: https://shorturl.at/dhoqx.
****
HeeSun Bae,Seungjae Shin,Byeonghu Na,Il-chul Moon
Dirichlet-based Per-Sample Weighting by Transition Matrix for Noisy Label Learning
For learning with noisy labels, the transition matrix, which explicitly models the relation between noisy label distribution and clean label distribution, has been utilized to achieve the statistical consistency of either the classifier or

the risk. Previous researches have focused more on how to estimate this transiti on matrix well, rather than how to utilize it. We propose good utilization of th e transition matrix is crucial and suggest a new utilization method based on res ampling, coined RENT. Specifically, we first demonstrate current utilizations ca n have potential limitations for implementation. As an extension to Reweighting, we suggest the Dirichlet distribution-based per-sample Weight Sampling (DWS) fr amework, and compare reweighting and resampling under DWS framework. With the an alyses from DWS, we propose RENT, a REsampling method with Noise Transition matr ix. Empirically, RENT consistently outperforms existing transition matrix utiliz ation methods, which includes reweighting, on various benchmark datasets. Our co de is available at https://github.com/BaeHeeSun/RENT.
****
Fengrui Tian,Yueqi Duan,Angtian Wang,Jianfei Guo,Shaoyi Du
Semantic Flow: Learning Semantic Fields of Dynamic Scenes from Monocular Videos
In this work, we pioneer Semantic Flow, a neural semantic representation of dyna mic scenes from monocular videos. In contrast to previous NeRF methods that reco nstruct dynamic scenes from the colors and volume densities of individual points , Semantic Flow learns semantics from continuous flows that contain rich 3D moti on information. As there is 2D-to-3D ambiguity problem in the viewing direction when extracting 3D flow features from 2D video frames, we consider the volume de nsities as opacity priors that describe the contributions of flow features to th e semantics on the frames. More specifically, we first learn a flow network to p redict flows in the dynamic scene, and propose a flow feature aggregation module to extract flow features from video frames. Then, we propose a flow attention m odule to extract motion information from flow features, which is followed by a s emantic network to output semantic logits of flows. We integrate the logits with volume densities in the viewing direction to supervise the flow features with se mantic labels on video frames. Experimental results show that our model is able to learn from multiple dynamic scenes and supports a series of new tasks such as instance-level scene editing, semantic completions, dynamic scene tracking and semantic adaption on novel scenes.
****
Mahan Fathi,Clement Gehring,Jonathan Pilault,David Kanaa,Pierre-Luc Bacon,Ross G oroshin
Course Correcting Koopman Representations
Koopman representations aim to learn features of nonlinear dynamical systems (NL DS) which lead to linear dynamics in the latent space. Theoretically, such featu res can be used to simplify many problems in modeling and control of NLDS. In th is work we study autoencoder formulations of this problem, and different ways th ey can be used to model dynamics, specifically for future state prediction over long horizons. We discover several limitations of predicting future states in th e latent space and propose an inference-time mechanism, which we refer to as Per iodic Reencoding, for faithfully capturing long term dynamics. We justify this m ethod both analytically and empirically via experiments in low and high dimensio nal NLDS.
****
Samyak Jain,Robert Kirk,Ekdeep Singh Lubana,Robert P. Dick,Hidenori Tanaka,Tim R ocktäschel,Edward Grefenstette,David Krueger
Mechanistically analyzing the effects of fine-tuning on procedurally defined tas ks
Fine-tuning large pre-trained models has become the de facto strategy for develo ping both task-specific and general-purpose machine learning systems, including developing models that are safe to deploy. Despite its clear importance, there h as been little work that explains how fine-tuning alters the underlying capabili ties learnt by a model during pretraining: does fine-tuning yield entirely novel capabilities or does it just inhibit existing ones? An answer to this question would improve our ability to trust fine-tuning protocols meant to improve the sa fety of pre-trained models and delete unsafe capabilities.
We aim to make progress on this question by answering it in controlled settings where we can use mechanistic interpretability tools (e.g.~ network pruning and p

robing) to understand how the model's underlying capabilities are changing.
We perform an exhaustive analysis of the effects of fine-tuning in these setting s, and show: (i) the ubiquitous protocol of fine-tuning with a small learning ra te rarely alters the underlying model capabilities; (ii) often a minimal transfo rmation, which we call a wrapper, is learned on top of the underlying model capa bility, yielding the impression that a new capability has been learned or a prio r capability has been deleted; and (iii) continuing the fine-tuning process on a task where the pretraining capabilities are relevant leads to sample-efficient ``revival'' of the capability, i.e., the model starts to accurately reuse that c apability in just a few gradient steps. \textit{This potentially indicates a pra ctitioner could unintentionally render a safe model to be unsafe by merely fine-tuning on a downstream task.} We additionally perform analysis on language model s trained on the TinyStories dataset to support our claims in a realistic settin g.
****
Ayesha Vermani,Il Memming Park,Josue Nassar
Leveraging Generative Models for Unsupervised Alignment of Neural Time Series Da ta
Large scale inference models are widely used in neuroscience to extract latent r epresentations from high-dimensional neural recordings. Due to the statistical h eterogeneities between sessions and animals, a new model is trained from scratch to infer the underlying dynamics for each new dataset. This is computationally expensive and does not fully leverage all the available data. Moreover, as these models get more complex, they can be challenging to train. In parallel, it is b ecoming common to use pre-trained models in the machine learning community for f ew shot and transfer learning. One major hurdle that prevents the re-use of gene rative models in neuroscience is the complex spatio-temporal structure of neural dynamics within and across animals. Interestingly, the underlying dynamics iden tified from different datasets on the same task are qualitatively similar. In th is work, we exploit this observation and propose a source-free and unsupervised alignment approach that utilizes the learnt dynamics and enables the re-use of t rained generative models. We validate our approach on simulations and show the e fficacy of the alignment on neural recordings from the motor cortex obtained dur ing a reaching task.
****
Yulai Zhao,Wenhao Zhan,Xiaoyan Hu,Ho-fung Leung,Farzan Farnia,Wen Sun,Jason D. L ee
Provably Efficient CVaR RL in Low-rank MDPs
We study risk-sensitive Reinforcement Learning (RL), where we aim to maximize the Conditional Value at Risk (CVaR) with a fixed risk tolerance τ.
Prior theoretical work studying risk-sensitive RL focuses on the tabular Markov Decision Processes (MDPs) setting.
To extend CVaR RL to settings where state space is large, function approximation must be deployed.
We study CVaR RL in low-rank MDPs with nonlinear function approximation. Low-ra nk MDPs assume the underlying transition kernel admits a low-rank decomposition, but unlike prior linear models, low-rank MDPs do not assume the feature or stat e-action representation is known.
We propose a novel Upper Confidence Bound (UCB) bonus-driven algorithm to carefu lly balance the interplay between exploration, exploitation, and representation learning in CVaR RL.
We prove that our algorithm achieves a sample complexity of $\tilde{O}\left(\frac{H^7 A^2 d^4}{\tau^2 \epsilon^2}\right)$ to yield an ϵ-optimal CVaR, w here H is the length of each episode, A is the capacity of action space, and d is the dimension of representations.
Computational-wise, we design a novel discretized Least-Squares Value Iteration (LSVI) algorithm for the CVaR objective as the planning oracle and show that we can find the near-optimal policy in a polynomial running time with a Maximum Lik elihood Estimation oracle.
To our knowledge, this is the first provably efficient CVaR RL algorithm in low-

rank MDPs.
****
Jonghyun Lee,Dahuin Jung,Saehyung Lee,Junsung Park,Juhyeon Shin,Uiwon Hwang,Sungroh Yoon
Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors
Test-time adaptation (TTA) fine-tunes pre-trained deep neural networks for unseen test data. The primary challenge of TTA is limited access to the entire test dataset during online updates, causing error accumulation. To mitigate it, TTA methods have utilized the model output's entropy as a confidence metric that aims to determine which samples have a lower likelihood of causing error. Through experimental studies, however, we observed the unreliability of entropy as a confidence metric for TTA under biased scenarios and theoretically revealed that it stems from the neglect of the influence of latent disentangled factors of data on predictions. Building upon these findings, we introduce a novel TTA method named Destroy Your Object (DeYO), which leverages a newly proposed confidence metric named Pseudo-Label Probability Difference (PLPD). PLPD quantifies the influence of the shape of an object on prediction by measuring the difference between predictions before and after applying an object-destructive transformation. DeYO consists of sample selection and sample weighting, which employ entropy and PLPD concurrently. For robust adaptation, DeYO prioritizes samples that dominantly incorporate shape information when making predictions. Our extensive experiments demonstrate the consistent superiority of DeYO over baseline methods across various scenarios, including biased and wild. Project page is publicly available at https://whitesnowdrop.github.io/DeYO/.
****
Yin Fang,Ningyu Zhang,Zhuo Chen,Lingbing Guo,Xiaohui Fan,Huajun Chen
Domain-Agnostic Molecular Generation with Chemical Feedback
The generation of molecules with desired properties has become increasingly popular, revolutionizing the way scientists design molecular structures and providing valuable support for chemical and drug design. However, despite the potential of language models in molecule generation, they face challenges such as generating syntactically or chemically flawed molecules, having narrow domain focus, and struggling to create diverse and feasible molecules due to limited annotated data or external molecular databases.
To tackle these challenges, we introduce MolGen, a pre-trained molecular language model tailored specifically for molecule generation. Through the reconstruction of over 100 million molecular SELFIES, MolGen internalizes structural and grammatical insights. This is further enhanced by domain-agnostic molecular prefix tuning, fostering robust knowledge transfer across diverse domains. Importantly, our chemical feedback paradigm steers the model away from "molecular hallucinations", ensuring alignment between the model's estimated probabilities and real-world chemical preferences. Extensive experiments on well-known benchmarks underscore MolGen's optimization capabilities in properties such as penalized logP, QED, and molecular docking. Additional analyses confirm its proficiency in accurately capturing molecule distributions, discerning intricate structural patterns, and efficiently exploring the chemical space (https://github.com/zjunlp/MolGen).
****
Yilun Du,Sherry Yang,Pete Florence,Fei Xia,Ayzaan Wahid,brian ichter,Pierre Sermanet,Tianhe Yu,Pieter Abbeel,Joshua B. Tenenbaum,Leslie Pack Kaelbling,Andy Zeng,Jonathan Tompson
Video Language Planning
We are interested in enabling visual planning for complex long-horizon tasks in the space of generated videos and language, leveraging recent advances in large generative models pretrained on Internet-scale data. To this end, we present video language planning (VLP), an algorithm that consists of a tree search procedure, where we train (i) vision-language models to serve as both policies and value functions, and (ii) text-to-video models as dynamics models. VLP takes as input a long-horizon task instruction and current image observation, and outputs a long video plan that provides detailed multimodal (video and language) specificat

ions that describe how to complete the final task. VLP scales with increasing co
mputation budget where more computation time results in improved video plans, an
d is able to synthesize long-horizon video plans across different robotics domai
ns -- from multi-object rearrangement, to multi-camera bi-arm dexterous manipula
tion. Generated video plans can be translated into real robot actions via goal-c
onditioned policies, conditioned on each intermediate frame of the generated vid
eo. Experiments show that VLP substantially improves long-horizon task success r
ates compared to prior methods on both simulated and real robots (across 3 hardw
are platforms).
****
Yixuan Weng,Minjun Zhu,Fei Xia,Bin Li,Shizhu He,Kang Liu,Jun Zhao
Mastering Symbolic Operations: Augmenting Language Models with Compiled Neural N
etworks
Language models' (LMs) proficiency in handling deterministic symbolic reasoning
and rule-based tasks remains limited due to their dependency implicit learning o
n textual data. To endow LMs with genuine rule comprehension abilities, we propo
se "Neural Comprehension" - a framework that synergistically integrates compiled
 neural networks (CoNNs) into the standard transformer architecture. CoNNs are n
eural modules designed to explicitly encode rules through artificially generated
 attention weights. By incorporating CoNN modules, the Neural Comprehension fram
ework enables LMs to accurately and robustly execute rule-intensive symbolic tas
ks. Extensive experiments demonstrate the superiority of our approach over exist
ing techniques in terms of length generalization, efficiency, and interpretabili
ty for symbolic operations. Furthermore, it can be applied to LMs across differe
nt model scales, outperforming tool-calling methods in arithmetic reasoning task
s while maintaining superior inference efficiency. Our work highlights the poten
tial of seamlessly unifying explicit rule learning via CoNNs and implicit patter
n learning in LMs, paving the way for true symbolic comprehension capabilities.
The code is released at: \url{https://github.com/wengsyx/Neural-Comprehension}.
****
Yongchan Kwon,Eric Wu,Kevin Wu,James Zou
DataInf: Efficiently Estimating Data Influence in LoRA-tuned LLMs and Diffusion
Models
Quantifying the impact of training data points is crucial for understanding the
outputs of machine learning models and for improving the transparency of the AI
pipeline. The influence function is a principled and popular data attribution me
thod, but its computational cost often makes it challenging to use. This issue b
ecomes more pronounced in the setting of large language models and text-to-image
 models. In this work, we propose DataInf, an efficient influence approximation
method that is practical for large-scale generative AI models. Leveraging an eas
y-to-compute closed-form expression, DataInf outperforms existing influence comp
utation algorithms in terms of computational and memory efficiency. Our theoreti
cal analysis shows that DataInf is particularly well-suited for parameter-effici
ent fine-tuning techniques such as LoRA. Through systematic empirical evaluation
s, we show that DataInf accurately approximates influence scores and is orders o
f magnitude faster than existing methods. In applications to RoBERTa-large, Llam
a-2-13B-chat, and stable-diffusion-v1.5 models, DataInf effectively identifies t
he most influential fine-tuning examples better than other approximate influence
 scores. Moreover, it can help to identify which data points are mislabeled.
****
Adam Lechowicz,Rik Sengupta,Bo Sun,Shahin Kamali,Mohammad Hajiesmaili
Time Fairness in Online Knapsack Problems
The online knapsack problem is a classic problem in the field of online algorith
ms. Its canonical version asks how to pack items of different values and weights
 arriving online into a capacity-limited knapsack so as to maximize the total va
lue of the admitted items. Although optimal competitive algorithms are known for
 this problem, they may be fundamentally unfair, i.e., individual items may be t
reated inequitably in different ways. We formalize a practically-relevant notion
 of time fairness which effectively models a trade off between static and dynami
c pricing in a motivating application such as cloud resource allocation, and sho

w that existing algorithms perform poorly under this metric. We propose a param eterized deterministic algorithm where the parameter precisely captures the Pare to-optimal trade-off between fairness (static pricing) and competitiveness (dyna mic pricing). We show that randomization is theoretically powerful enough to be simultaneously competitive and fair; however, it does not work well in experimen ts. To further improve the trade-off between fairness and competitiveness, we de velop a nearly-optimal learning-augmented algorithm which is fair, consistent, a nd robust (competitive), showing substantial performance improvements in numeric al experiments.
****
Chenchen Gu,Xiang Lisa Li,Percy Liang,Tatsunori Hashimoto
On the Learnability of Watermarks for Language Models
Language model watermarking enables reliable detection of model-generated text, which has many applications in the responsible deployment of language models. Ex isting watermarking strategies operate by altering the decoder of an existing la nguage model, and the ability for a language model to directly learn to generate the watermark would have significant implications for the real-world deployment of watermarks. First, learned watermarks could be used to build open models tha t naturally generate watermarked text, allowing for open models to benefit from watermarking. Second, if watermarking is used to determine the provenance of gen erated text, an adversary can damage the reputation of a victim model by spoofin g its watermark and generating harmful watermarked text. To investigate the lear nability of watermarks, we propose watermark distillation, which trains a studen t model to behave like a teacher model that uses decoding-based watermarking. We test our approach on three distinct decoding-based watermarking strategies, fin ding that models can learn to generate watermarked text with high detectability. We also find limitations to learnability, including the loss of watermarking ca pabilities under fine-tuning on normal text and high sample complexity when lear ning low-distortion watermarks.
****
Yongsheng Mei,Mahdi Imani,Tian Lan
Bayesian Optimization through Gaussian Cox Process Models for Spatio-temporal Da ta
Bayesian optimization (BO) has established itself as a leading strategy for effi ciently optimizing expensive-to-evaluate functions. Existing BO methods mostly r ely on Gaussian process (GP) surrogate models and are not applicable to (doubly-stochastic) Gaussian Cox processes, where the observation process is modulated b y a latent intensity function modeled as a GP. In this paper, we propose a novel maximum *a posteriori* inference of Gaussian Cox processes. It leverages the La place approximation and change of kernel technique to transform the problem into a new reproducing kernel Hilbert space, where it becomes more tractable computa tionally. It enables us to obtain both a functional posterior of the latent inte nsity function and the covariance of the posterior, thus extending existing work s that often focus on specific link functions or estimating the posterior mean.
Using the result, we propose a BO framework based on the Gaussian Cox process mo del and further develop a Nyström approximation for efficient computation. Exten sive evaluations on various synthetic and real-world datasets demonstrate signif icant improvement over state-of-the-art inference solutions for Gaussian Cox pro cesses, as well as effective BO with a wide range of acquisition functions desig ned through the underlying Gaussian Cox process model.
****
Guihong Li,Hsiang Hsu,Chun-Fu Chen,Radu Marculescu
Machine Unlearning for Image-to-Image Generative Models
Machine unlearning has emerged as a new paradigm to deliberately forget data sam ples from a given model in order to adhere to stringent regulations.
However, existing machine unlearning methods have been primarily focused on clas sification models, leaving the landscape of unlearning for generative models rel atively unexplored.
This paper serves as a bridge, addressing the gap by providing a unifying framew ork of machine unlearning for image-to-image generative models.

Within this framework, we propose a computationally-efficient algorithm, underpinned by rigorous theoretical analysis, that demonstrates negligible performance degradation on the retain samples, while effectively removing the information from the forget samples.
Empirical studies on two large-scale datasets, ImageNet-1K and Places-365, further show that our algorithm does not rely on the availability of the retain samples, which further complies with data retention policy.
To our best knowledge, this work is the first that represents systemic, theoretical, empirical explorations of machine unlearning specifically tailored for image-to-image generative models.
****
Yuanwen Yue,Sabarinath Mahadevan,Jonas Schult,Francis Engelmann,Bastian Leibe,Konrad Schindler,Theodora Kontogianni
AGILE3D: Attention Guided Interactive Multi-object 3D Segmentation
During interactive segmentation, a model and a user work together to delineate objects of interest in a 3D point cloud. In an iterative process, the model assigns each data point to an object (or the background), while the user corrects errors in the resulting segmentation and feeds them back into the model. The current best practice formulates the problem as binary classification and segments objects one at a time. The model expects the user to provide positive clicks to indicate regions wrongly assigned to the background and negative clicks on regions wrongly assigned to the object. Sequentially visiting objects is wasteful since it disregards synergies between objects: a positive click for a given object can, by definition, serve as a negative click for nearby objects. Moreover, a direct competition between adjacent objects can speed up the identification of their common boundary. We introduce AGILE3D, an efficient, attention-based model that (1) supports simultaneous segmentation of multiple 3D objects, (2) yields more accurate segmentation masks with fewer user clicks, and (3) offers faster inference. Our core idea is to encode user clicks as spatial-temporal queries and enable explicit interactions between click queries as well as between them and the 3D scene through a click attention module. Every time new clicks are added, we only need to run a lightweight decoder that produces updated segmentation masks. In experiments with four different 3D point cloud datasets, AGILE3D sets a new state-of-the-art. Moreover, we also verify its practicality in real-world setups with real user studies. Project page: https://ywyue.github.io/AGILE3D.
****
Aditya Chattopadhyay,Kwan Ho Ryan Chan,Rene Vidal
Bootstrapping Variational Information Pursuit with Large Language and Vision Models for Interpretable Image Classification
Variational Information Pursuit (V-IP) is an interpretable-by-design framework that makes predictions by sequentially selecting a short chain of task-relevant, user-defined interpretable queries about the data that are most informative for the task. The selected query-answer chain serves as an explanation for the prediction. Applying the framework to any task requires (i) specification of a query set, and (ii) densely annotated data with query answers to train classifiers to answer queries at test time. This limits V-IP’s application to small-scale tasks where manual data annotation is feasible. In this work, we focus on image classification tasks and propose to relieve this bottleneck by leveraging pretrained language and vision models. Specifically, following recent work, we propose to use GPT, a Large Language Model, to propose semantic concepts as queries for a given classification task. To answer these queries, we propose a Concept Question-Answering network (Concept-QA) which learns to answer binary queries about semantic concepts in images. We design pseudo-labels to train our Concept-QA model using GPT and CLIP (a Vision-Language Model). Empirically, we find our Concept-QA model to be competitive with state-of-the-art VQA models in terms of answer- ing accuracy but with an order of magnitude fewer parameters. This allows for seamless integration of Concept-QA into the V-IP framework as a fast-answering mechanism. We name this method Concept-QA+V-IP. Finally, we show on several datasets that Concept-QA+V-IP produces shorter, interpretable query chains which are more accurate than V-IP trained with CLIP-based answering systems.

****
Yukai Shi,Jianan Wang,He CAO,Boshi Tang,Xianbiao Qi,Tianyu Yang,Yukun Huang,Shil
ong Liu,Lei Zhang,Heung-Yeung Shum
TOSS: High-quality Text-guided Novel View Synthesis from a Single Image
In this paper, we present TOSS, which introduces text to the task of novel view
synthesis (NVS) from just a single RGB image.
While Zero123 has demonstrated impressive zero-shot open-set NVS capabilities, i
t treats NVS as a pure image-to-image translation problem. This approach suffers
 from the challengingly under-constrained nature of single-view NVS: the process
 lacks means of explicit user control and often result in implausible NVS genera
tions.
To address this limitation, TOSS uses text as high-level semantic information to
 constrain the NVS solution space.
TOSS fine-tunes text-to-image Stable Diffusion pre-trained on large-scale text-i
mage pairs and introduces modules specifically tailored to image and camera pose
 conditioning, as well as dedicated training for pose correctness and preservati
on of fine details.
Comprehensive experiments are conducted with results showing that our proposed T
OSS outperforms Zero123 with higher-quality NVS results and faster convergence.
We further support these results with comprehensive ablations that underscore th
e effectiveness and potential of
the introduced semantic guidance and architecture design.
****
Zengwei Yao,Liyong Guo,Xiaoyu Yang,Wei Kang,Fangjun Kuang,Yifan Yang,Zengrui Jin
,Long Lin,Daniel Povey
Zipformer: A faster and better encoder for automatic speech recognition
The Conformer has become the most popular encoder model for automatic speech rec
ognition (ASR). It adds convolution modules to a transformer to learn both loca
l and global dependencies. In this work we describe a faster, more memory-effici
ent, and better-performing transformer, called Zipformer. Modeling changes incl
ude: 1) a U-Net-like encoder structure where middle stacks operate at lower fram
e rates; 2) reorganized block structure with more modules, within which we re-us
e attention weights for efficiency; 3) a modified form of LayerNorm called BiasN
orm allows us to retain some length information; 4) new activation functions Sw
ooshR and SwooshL work better than Swish. We also propose a new optimizer, call
ed ScaledAdam, which scales the update by each tensor's current scale to keep th
e relative change about the same, and also explictly learns the parameter scale.
 It achieves faster converge and better performance than Adam. Extensive experim
ents on LibriSpeech, Aishell-1, and WenetSpeech datasets demonstrate the effecti
veness of our proposed Zipformer over other state-of-the-art ASR models. Our cod
e is publicly available at https://github.com/k2-fsa/icefall.
****
Max F Burg,Thomas Zenkel,Michaela Vystr■ilová,Jonathan Oesterle,Larissa Höfling,
Konstantin Friedrich Willeke,Jan Lause,Sarah Müller,Paul G. Fahey,Zhiwei Ding,Ke
lli Restivo,Shashwat Sridhar,Tim Gollisch,Philipp Berens,Andreas S. Tolias,Thoma
s Euler,Matthias Bethge,Alexander S Ecker
Most discriminative stimuli for functional cell type clustering
Identifying cell types and understanding their functional properties is crucial
for unraveling the mechanisms underlying perception and cognition. In the retina
, functional types can be identified by carefully selected stimuli, but this req
uires expert domain knowledge and biases the procedure towards previously known
cell types. In the visual cortex, it is still unknown what functional types exis
t and how to identify them. Thus, for unbiased identification of the functional
cell types in retina and visual cortex, new approaches are needed. Here we propo
se an optimization-based clustering approach using deep predictive models to obt
ain functional clusters of neurons using Most Discriminative Stimuli (MDS). Our
approach alternates between stimulus optimization with cluster reassignment akin
 to an expectation-maximization algorithm. The algorithm recovers functional clu
sters in mouse retina, marmoset retina and macaque visual area V4. This demonstr
ates that our approach can successfully find discriminative stimuli across speci

es, stages of the visual system and recording techniques. The resulting most discriminative stimuli can be used to assign functional cell types fast and on the fly, without the need to train complex predictive models or show a large natural scene dataset, paving the way for experiments that were previously limited by experimental time. Crucially, MDS are interpretable: they visualize the distinctive stimulus patterns that most unambiguously identify a specific type of neuron.
****
Weian Mao,Muzhi Zhu,Zheng Sun,Shuaike Shen,Lin Yuanbo Wu,Hao Chen,Chunhua Shen
De novo Protein Design Using Geometric Vector Field Networks
Advances like protein diffusion have marked revolutionary progress in $\textit{de novo}$ protein design, a central topic in life science. These methods typically depend on protein structure encoders to model residue backbone frames, where atoms do not exist. Most prior encoders rely on atom-wise features, such as angles and distances between atoms, which are not available in this context. Only a few basic encoders, like IPA, have been proposed for this scenario, exposing the frame modeling as a bottleneck. In this work, we introduce the Vector Field Network (VFN), that enables network layers to perform learnable vector computations between coordinates of frame-anchored virtual atoms, thus achieving a higher capability for modeling frames. The vector computation operates in a manner similar to a linear layer, with each input channel receiving 3D virtual atom coordinates instead of scalar values. The multiple feature vectors output by the vector computation are then used to update the residue representations and virtual atom coordinates via attention aggregation. Remarkably, VFN also excels in modeling both frames and atoms, as the real atoms can be treated as the virtual atoms for modeling, positioning VFN as a potential $\textit{universal encoder}$. In protein diffusion (frame modeling), VFN exhibits a impressive performance advantage over IPA, excelling in terms of both designability ($\textbf{67.04}$\% vs. 53.58\%) and diversity ($\textbf{66.54}$\% vs. 51.98\%). In inverse folding(frame and atom modeling), VFN outperforms the previous SoTA model, PiFold ($\textbf{54.7}$\% vs. 51.66\%), on sequence recovery rate; we also propose a method of equipping VFN with the ESM model, which significantly surpasses the previous ESM-based SoTA ($\textbf{62.67}$\% vs. 55.65\%), LM-Design, by a substantial margin. Code is available at https://github.com/aim-uofa/VFN
****
Mouïn Ben Ammar,Nacim Belkhir,Sebastian Popescu,Antoine Manzanera,Gianni Franchi
NECO: NEural Collapse Based Out-of-distribution detection
Detecting out-of-distribution (OOD) data is a critical challenge in machine learning due to model overconfidence, often without awareness of their epistemological limits. We hypothesize that "neural collapse", a phenomenon affecting in-distribution data for models trained beyond loss convergence, also influences OOD data. To benefit from this interplay, we introduce NECO, a novel post-hoc method for OOD detection, which leverages the geometric properties of “neural collapse” and of principal component spaces to identify OOD data. Our extensive experiments demonstrate that NECO achieves state-of-the-art results on both small and large-scale OOD detection tasks while exhibiting strong generalization capabilities across different network architectures. Furthermore, we provide a theoretical explanation for the effectiveness of our method in OOD detection. We plan to release the code after the anonymity period.
****
Jiayuan Ye,Anastasia Borovykh,Soufiane Hayou,Reza Shokri
Leave-one-out Distinguishability in Machine Learning
We introduce a new analytical framework to quantify the changes in a machine learning algorithm's output distribution following the inclusion of a few data points in its training set, a notion we define as leave-one-out distinguishability (LOOD). This problem is key to measuring data **memorization** and information **leakage** in machine learning, and the **influence** of training data points on model predictions. We illustrate how our method broadens and refines existing empirical measures of memorization and privacy risks associated with training data. We use Gaussian processes to model the randomness of machine learning algorithms, and validate LOOD with extensive empirical analysis of information leakage

using membership inference attacks. Our theoretical framework enables us to investigate the causes of information leakage and where the leakage is high. For example, we analyze the influence of activation functions, on data memorization. Additionally, our method allows us to optimize queries that disclose the most significant information about the training data in the leave-one-out setting. We illustrate how optimal queries can be used for accurate **reconstruction** of training data.
****
David Brellmann,Eloïse Berthier,David Filliat,Goran Frehse
On Double Descent in Reinforcement Learning with LSTD and Random Features
Temporal Difference (TD) algorithms are widely used in Deep Reinforcement Learning (RL). Their performance is heavily influenced by the size of the neural network. While in supervised learning, the regime of over-parameterization and its benefits are well understood, the situation in RL is much less clear. In this paper, we present a theoretical analysis of the influence of network size and l_2-regularization on performance. We identify the ratio between the number of parameters and the number of visited states as a crucial factor and define over-parameterization as the regime when it is larger than one. Furthermore, we observe a double descent phenomenon, i.e., a sudden drop in performance around the parameter/state ratio of one. Leveraging random features and the lazy training regime, we study the regularized Least-Square Temporal Difference (LSTD) algorithm in an asymptotic regime, as both the number of parameters and states go to infinity, maintaining a constant ratio. We derive deterministic limits of both the empirical and the true Mean-Squared Bellman Error (MSBE) that feature correction terms responsible for the double descent. Correction terms vanish when the l_2-regularization is increased or the number of unvisited states goes to zero. Numerical experiments with synthetic and small real-world environments closely match the theoretical predictions.
****
Xuming Hu,Junzhe Chen,Xiaochuan Li,Yufei Guo,Lijie Wen,Philip S. Yu,Zhijiang Guo
Towards Understanding Factual Knowledge of Large Language Models
Large language models (LLMs) have recently driven striking performance improvements across a range of natural language processing tasks. The factual knowledge acquired during pretraining and instruction tuning can be useful in various downstream tasks, such as question answering, and language generation. Unlike conventional Knowledge Bases (KBs) that explicitly store factual knowledge, LLMs implicitly store facts in their parameters. Content generated by the LLMs can often exhibit inaccuracies or deviations from the truth, due to facts that can be incorrectly induced or become obsolete over time. To this end, we aim to explore the extent and scope of factual knowledge within LLMs by designing the benchmark Pinocchio. Pinocchio contains 20K diverse factual questions that span different sources, timelines, domains, regions, and languages. Furthermore, we investigate whether LLMs can compose multiple facts, update factual knowledge temporally, reason over multiple pieces of facts, identify subtle factual differences, and resist adversarial examples. Extensive experiments on different sizes and types of LLMs show that existing LLMs still lack factual knowledge and suffer from various spurious correlations. We believe this is a critical bottleneck for realizing trustworthy artificial intelligence. The dataset Pinocchio and our codes are publicly available at: https://github.com/THU-BPM/Pinocchio.
****
Xindi Yang,Zeke Xie,Xiong Zhou,Boyu Liu,Buhua Liu,Yi Liu,Haoran Wang,YUNFENG CAI,Mingming Sun
Neural Field Classifiers via Target Encoding and Classification Loss
Neural field methods have seen great progress in various long-standing tasks in computer vision and computer graphics, including novel view synthesis and geometry reconstruction. As existing neural field methods try to predict some coordinate-based continuous target values, such as RGB for Neural Radiance Field (NeRF), all of these methods are regression models and are optimized by some regression loss. However, are regression models really better than classification models for neural field methods? In this work, we try to visit this very fundamental but

overlooked question for neural fields from a machine learning perspective. We s uccessfully propose a novel Neural Field Classifier (NFC) framework which formul ates existing neural field methods as classification tasks rather than regressio n tasks. The proposed NFC can easily transform arbitrary Neural Field Regressor (NFR) into its classification variant via employing a novel Target Encoding modu le and optimizing a classification loss. By encoding a continuous regression tar get into a high-dimensional discrete encoding, we naturally formulate a multi-la bel classification task. Extensive experiments demonstrate the impressive effect iveness of NFC at the nearly free extra computational costs. Moreover, NFC also shows robustness to sparse inputs, corrupted images, and dynamic scenes.
****
Izzeddin Gur,Hiroki Furuta,Austin V Huang,Mustafa Safdari,Yutaka Matsuo,Douglas Eck,Aleksandra Faust
A Real-World WebAgent with Planning, Long Context Understanding, and Program Syn thesis
Pre-trained large language models (LLMs) have recently achieved better generaliz ation and sample efficiency in autonomous web automation.
However, the performance on real-world websites has still suffered from (1) open domainness, (2) limited context length, and (3) lack of inductive bias on HTML.
We introduce WebAgent, an LLM-driven agent that learns from self-experience to c omplete tasks on real websites following natural language instructions.
WebAgent plans ahead by decomposing instructions into canonical sub-instructions , summarizes long HTML documents into task-relevant snippets, and acts on websit es via Python programs generated from those.
We design WebAgent with Flan-U-PaLM, for grounded code generation, and HTML-T5, new pre-trained LLMs for long HTML documents using local and global attention me chanisms and a mixture of long-span denoising objectives, for planning and summa rization.
We empirically demonstrate that our modular recipe improves the success on real websites by over 50%, and that HTML-T5 is the best model to solve various HTML u nderstanding tasks; achieving 18.7% higher success rate than the prior method on MiniWoB web automation benchmark, and SoTA performance on Mind2Web, an offline task planning evaluation.
****
JangHo Park,Gihyun Kwon,Jong Chul Ye
ED-NeRF: Efficient Text-Guided Editing of 3D Scene With Latent Space NeRF
Recently, there has been a significant advancement in text-to-image diffusion mo dels, leading to groundbreaking performance in 2D image generation. These advanc ements have been extended to 3D models, enabling the generation of novel 3D obje cts from textual descriptions. This has evolved into NeRF editing methods, whic h allow the manipulation of existing 3D objects through textual conditioning. Ho wever, existing NeRF editing techniques have faced limitations in their performa nce due to slow training speeds and the use of loss functions that do not adequa tely consider editing. To address this, here we present a novel 3D NeRF editing approach dubbed ED-NeRF by successfully embedding real-world scenes into the la tent space of the latent diffusion model (LDM) through a unique refinement layer . This approach enables us to obtain a NeRF backbone that is not only faster but also more amenable to editing compared to traditional image space NeRF editing. Furthermore, we propose an improved loss function tailored for editing by migra ting the delta denoising score (DDS) distillation loss, originally used in 2D im age editing to the three-dimensional domain. This novel loss function surpasses the well-known score distillation sampling (SDS) loss in terms of suitability fo r editing purposes. Our experimental results demonstrate that ED-NeRF achieves f aster editing speed while producing improved output quality compared to state-of -the-art 3D editing models.
****
Jiajun Ma,Tianyang Hu,Wenjia Wang,Jiacheng Sun
Elucidating the design space of classifier-guided diffusion generation
Guidance in conditional diffusion generation is of great importance for sample q uality and controllability.

However, existing guidance schemes are to be desired.
On one hand, mainstream methods such as classifier guidance and classifier-free guidance both require extra training with labeled data, which is time-consuming and unable to adapt to new conditions.
On the other hand, training-free methods such as universal guidance, though more flexible, have yet to demonstrate comparable performance.
In this work, through a comprehensive investigation into the design space, we show that it is possible to achieve significant performance improvements over existing guidance schemes by leveraging off-the-shelf classifiers in a training-free fashion, enjoying the best of both worlds.
Employing calibration as a general guideline, we propose several pre-conditioning techniques to better exploit pretrained off-the-shelf classifiers for guiding diffusion generation.
Extensive experiments on ImageNet validate our proposed method, showing that state-of-the-art (SOTA) diffusion models (DDPM, EDM, DiT) can be further improved (up to 20\%) using off-the-shelf classifiers with barely any extra computational cost.
With the proliferation of publicly available pretrained classifiers, our proposed approach has great potential and can be readily scaled up to text-to-image generation tasks.
****
Yijie Lin,Jie Zhang,Zhenyu Huang,Jia Liu,zujie wen,Xi Peng
Multi-granularity Correspondence Learning from Long-term Noisy Videos
Existing video-language studies mainly focus on learning short video clips, leaving long-term temporal dependencies rarely explored due to over-high computational cost of modeling long videos. To address this issue, one feasible solution is learning the correspondence between video clips and captions, which however inevitably encounters the multi-granularity noisy correspondence (MNC) problem. To be specific, MNC refers to the clip-caption misalignment (coarse-grained) and frame-word misalignment (fine-grained), hindering temporal learning and video understanding. In this paper, we propose NOise Robust Temporal Optimal traNsport (Norton) that addresses MNC in a unified optimal transport (OT) framework. In brief, Norton employs video-paragraph and clip-caption contrastive losses to capture long-term dependencies based on OT. To address coarse-grained misalignment in video-paragraph contrast, Norton filters out the irrelevant clips and captions through an alignable prompt bucket and realigns asynchronous clip-caption pairs based on transport distance. To address the fine-grained misalignment, Norton incorporates a soft-maximum operator to identify crucial words and key frames. Additionally, Norton exploits the potential faulty negative samples in clip-caption contrast by rectifying the alignment target with OT assignment to ensure precise temporal modeling. Extensive experiments on video retrieval, videoQA, and action segmentation verify the effectiveness of our method.
Code is available at https://lin-yijie.github.io/projects/Norton.
****
Kai Cui,Sascha H. Hauck,Christian Fabian,Heinz Koeppl
Learning Decentralized Partially Observable Mean Field Control for Artificial Collective Behavior
Recent reinforcement learning (RL) methods have achieved success in various domains. However, multi-agent RL (MARL) remains a challenge in terms of decentralization, partial observability and scalability to many agents. Meanwhile, collective behavior requires resolution of the aforementioned challenges, and remains of importance to many state-of-the-art applications such as active matter physics, self-organizing systems, opinion dynamics, and biological or robotic swarms. Here, MARL via mean field control (MFC) offers a potential solution to scalability, but fails to consider decentralized and partially observable systems. In this paper, we enable decentralized behavior of agents under partial information by proposing novel models for decentralized partially observable MFC (Dec-POMFC), a broad class of problems with permutation-invariant agents allowing for reduction to tractable single-agent Markov decision processes (MDP) with single-agent RL solution. We provide rigorous theoretical results, including a dynamic programmin

g principle, together with optimality guarantees for Dec-POMFC solutions applied to finite swarms of interest. Algorithmically, we propose Dec-POMFC-based policy gradient methods for MARL via centralized training and decentralized execution, together with policy gradient approximation guarantees. In addition, we improve upon state-of-the-art histogram-based MFC by kernel methods, which is of separate interest also for fully observable MFC. We evaluate numerically on representative collective behavior tasks such as adapted Kuramoto and Vicsek swarming models, being on par with state-of-the-art MARL. Overall, our framework takes a step towards RL-based engineering of artificial collective behavior via MFC.
****
Nikita Srivatsan,Sofia Samaniego,Omar Florez,Taylor Berg-Kirkpatrick
Alt-Text with Context: Improving Accessibility for Images on Twitter
In this work we present an approach for generating alternative text (or alt-text) descriptions for images shared on social media, specifically Twitter. More than just a special case of image captioning, alt-text is both more literally descriptive and context-specific. Also critically, images posted to Twitter are often accompanied by user-written text that despite not necessarily describing the image may provide useful context that if properly leveraged can be informative. We address this task with a multimodal model that conditions on both textual information from the associated social media post as well as visual signal from the image, and demonstrate that the utility of these two information sources stacks. We put forward a new dataset of 371k images paired with alt-text and tweets scraped from Twitter and evaluate on it across a variety of automated metrics as well as human evaluation. We show that our approach of conditioning on both tweet text and visual information significantly outperforms prior work, by more than 2x on BLEU@4.
****
Junfeng Long,ZiRui Wang,Quanyi Li,Liu Cao,Jiawei Gao,Jiangmiao Pang
Hybrid Internal Model: Learning Agile Legged Locomotion with Simulated Robot Response
Robust locomotion control depends on accurate state estimations. However, the sensors of most legged robots can only provide partial and noisy observations, making the estimation particularly challenging, especially for external states like terrain frictions and elevation maps. Inspired by the classical Internal Model Control principle, we consider these external states as disturbances and introduce Hybrid Internal Model (HIM) to estimate them according to the response of the robot. The response, which we refer to as the hybrid internal embedding, contains the robot’s explicit velocity and implicit stability representation, corresponding to two primary goals for locomotion tasks: explicitly tracking velocity and implicitly maintaining stability. We use contrastive learning to optimize the embedding to be close to the robot’s successor state, in which the response is naturally embedded. HIM has several appealing benefits: It only needs the robot’s proprioceptions, i.e., those from joint encoders and IMU as observations. It innovatively maintains consistent observations between simulation reference and reality that avoids information loss in mimicking learning. It exploits batch-level information that is more robust to noises and keeps better sample efficiency. It only requires 1 hour of training on an RTX 4090 to enable a quadruped robot to traverse any terrain under any disturbances. A wealth of real-world experiments demonstrates its agility, even in high-difficulty tasks and cases never occurred during the training process, revealing remarkable open-world generalizability.
****
Kathan Shah,Chawin Sitawarin
SPDER: Semiperiodic Damping-Enabled Object Representation
We present a neural network architecture designed to naturally learn a positional embedding and overcome the spectral bias towards lower frequencies faced by conventional implicit neural representation networks. Our proposed architecture, SPDER, is a simple MLP that uses an activation function composed of a sinusoidal multiplied by a sublinear function, called the damping function. The sinusoidal enables the network to automatically learn the positional embedding of an input

coordinate while the damping passes on the actual coordinate value by preventing it from being projected down to within a finite range of values. Our results indicate that SPDERs speed up training by 10 times and converge to losses 1,500 to 50,000 times lower than that of the state-of-the-art for image representation. SPDER is also state-of-the-art in audio representation. The superior representation capability allows SPDER to also excel on multiple downstream tasks such as image super-resolution and video frame interpolation. We provide intuition as to why SPDER significantly improves fitting compared to that of other INR methods while requiring no hyperparameter tuning or preprocessing. See code at https://github.com/katop1234/SPDER.
****
Christian Horvat,Jean-Pascal Pfister
On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models
Diffusion models are generative models that have recently demonstrated impressive performances in terms of sampling quality and density estimation in high dimensions. They rely on a forward continuous diffusion process and a backward continuous denoising process, which can be described by a time-dependent vector field and is used as a generative model. In the original formulation of the diffusion model, this vector field is assumed to be the score function (i.e. it is the gradient of the log-probability at a given time in the diffusion process). Curiously, on the practical side, most studies on diffusion models implement this vector field as a neural network function and do not constrain it be the gradient of some energy function (that is, most studies do not constrain the vector field to be conservative). Even though some studies investigated empirically whether such a constraint will lead to a performance gain, they lead to contradicting results and failed to provide analytical results. Here, we provide three analytical results regarding the extent of the modeling freedom of this vector field. {Firstly, we propose a novel decomposition of vector fields into a conservative component and an orthogonal component which satisfies a given (gauge) freedom. Secondly, from this orthogonal decomposition, we show that exact density estimation and exact sampling is achieved when the conservative component is exactly equals to the true score and therefore conservativity is neither necessary nor sufficient to obtain exact density estimation and exact sampling. Finally, we show that when it comes to inferring local information of the data manifold, constraining the vector field to be conservative is desirable.
****
Youn-Yeol Yu,Jeongwhan Choi,Woojin Cho,Kookjin Lee,Nayong Kim,Kiseok Chang,ChangSeung Woo,ILHO KIM,SeokWoo Lee,Joon Young Yang,SOOYOUNG YOON,Noseong Park
Learning Flexible Body Collision Dynamics with Hierarchical Contact Mesh Transformer
Recently, many mesh-based graph neural network (GNN) models have been proposed for modeling complex high-dimensional physical systems. Remarkable achievements have been made in significantly reducing the solving time compared to traditional numerical solvers. These methods are typically designed to i) reduce the computational cost in solving physical dynamics and/or ii) propose techniques to enhance the solution accuracy in fluid and rigid body dynamics. However, it remains under-explored whether they are effective in addressing the challenges of flexible body dynamics, where instantaneous collisions occur within a very short timeframe. In this paper, we present Hierarchical Contact Mesh Transformer (HCMT), which uses hierarchical mesh structures and can learn long-range dependencies (occurred by collisions) among spatially distant positions of a body --- two close positions in a higher-level mesh corresponds to two distant positions in a lower-level mesh. HCMT enables long-range interactions, and the hierarchical mesh structure quickly propagates collision effects to faraway positions. To this end, it consists of a contact mesh Transformer and a hierarchical mesh Transformer (CMT and HMT, respectively). Lastly, we propose a flexible body dynamics dataset, consisting of trajectories that reflect experimental settings frequently used in the display industry for product designs. We also compare the performance of several baselines using well-known benchmark datasets. Our results show that HCMT pr

ovides significant performance improvements over existing methods. Our code is available at https://github.com/yuyudeep/hcmt.
****
Yinya Huang,Xiaohan Lin,Zhengying Liu,Qingxing Cao,Huajian Xin,Haiming Wang,Zhenguo Li,Linqi Song,Xiaodan Liang
MUSTARD: Mastering Uniform Synthesis of Theorem and Proof Data
Recent large language models (LLMs) have witnessed significant advancement in various tasks, including mathematical reasoning and theorem proving. As these two tasks require strict and formal multi-step inference, they are appealing domains for exploring the reasoning ability of LLMs but still face important challenges. Previous studies such as Chain-of-Thought (CoT) have revealed the effectiveness of intermediate steps guidance. However, such step-wise annotation requires heavy labor, leading to insufficient training steps for current benchmarks. To fill this gap, this work introduces MUSTARD, a data generation framework that masters uniform synthesis of theorem and proof data of high quality and diversity. MUSTARD synthesizes data in three stages: (1) It samples a few mathematical concept seeds as the problem category. (2) Then, it prompts a generative language model with the sampled concepts to obtain both the problems and their step-wise formal solutions. (3) Lastly, the framework utilizes a proof assistant (e.g., Lean Prover) to filter the valid proofs. With the proposed MUSTARD, we present a theorem-and-proof benchmark MUSTARDSAUCE with 5,866 valid data points. Each data point contains an informal statement, an informal proof, and a translated formal proof that passes the prover validation. We perform extensive analysis and demonstrate that MUSTARD generates validated high-quality step-by-step data. We further apply the MUSTARDSAUCE for fine-tuning smaller language models. The fine-tuned Llama 2-7B achieves a 15.41% average relative performance gain in automated theorem proving, and 8.18% in math word problems. Codes and data are available at https://github.com/Eleanor-H/MUSTARD.
****
Sejun Park,Sanghyuk Chun,Wonyeol Lee
What does automatic differentiation compute for neural networks?
Forward- or reverse-mode automatic differentiation (AD) is a popular algorithm for computing the derivative of a function expressed by a program. AD always outputs the correct derivative if a program does not use any non-differentiable functions and control flows; however, it may return an arbitrary value otherwise. In this work, we investigate what AD computes for neural networks that may contain non-differentiable functions such as ReLU and maxpools. We first prove that AD always returns a generalized derivative called a Clarke subderivative for networks with pointwise activation functions, if the minibatch size is one and all non-differentiable neurons have distinct bias parameters. We show that the same conclusion does not hold otherwise, but does hold under some mild sufficient conditions. We also prove similar results for more general networks that can use maxpools and bias parameters shared across different neurons. We empirically check our sufficient conditions over popular network architectures and observe that AD almost always computes a Clarke subderivative in practical learning setups.
****
Nikolas Patris,Ioannis Panageas
Learning Nash Equilibria in Rank-1 Games
Learning Nash equilibria (NE) in games has garnered significant attention, particularly in the context of training Generative Adversarial Networks (GANs) and multi-agent Reinforcement Learning. The current state-of-the-art in efficiently learning games focuses on landscapes that meet the (weak) Minty property or games characterized by a unique function, often referred to as potential games. A significant challenge in this domain is that computing Nash equilibria is a computationally intractable task [Daskalakis et al. 2009].

In this paper we focus on bimatrix games (A,B) called rank-1. These are games in which the sum of the payoff matrices A+B is a rank 1 matrix; note that standard zero-sum games are rank 0. We show that optimistic gradient descent/ascent converges to an \epsilon-approximate NE after 1/\epsilon^2 log(1/\epsilon) iterates

in rank-1 games. We achieve this by leveraging structural results about the NE l
andscape of rank-1 games Adsul et al. 2021. Notably, our approach bypasses the f
act that these games do not satisfy the MVI property.
****
Nikhil Prakash,Tamar Rott Shaham,Tal Haklay,Yonatan Belinkov,David Bau
Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking
Fine-tuning on generalized tasks such as instruction following, code generation,
 and mathematics has been shown to enhance language models' performance on a ran
ge of tasks. Nevertheless, explanations of how such fine-tuning influences the i
nternal computations in these models remain elusive. We study how fine-tuning af
fects the internal mechanisms implemented in language models. As a case study, w
e explore the property of entity tracking, a crucial facet of language comprehen
sion, where models fine-tuned on mathematics have substantial performance gains.
 We identify a mechanism that enables entity tracking and show that (i) both the
 original model and its fine-tuned version implement entity tracking with the sa
me circuit. In fact, the entity tracking circuit of the fine-tuned version perfo
rms better than the full original model. (ii) The circuits of all the models imp
lement roughly the same functionality, that is entity tracking is performed by t
racking the position of the correct entity in both the original model and its fi
ne-tuned version. (iii) Performance boost in the fine-tuned model is primarily a
ttributed to its improved ability to handle positional information. To uncover t
hese findings, we employ two methods: DCM, which automatically detects model com
ponents responsible for specific semantics, and CMAP, a new approach for patchin
g activations across models to reveal improved mechanisms. Our findings suggest
that fine-tuning enhances, rather than fundamentally alters, the mechanistic ope
ration of the model.
****
Arvind V. Mahankali,Tatsunori Hashimoto,Tengyu Ma
One Step of Gradient Descent is Provably the Optimal In-Context Learner with One
 Layer of Linear Self-Attention
Recent works have empirically analyzed in-context learning and shown that transf
ormers trained on synthetic linear regression tasks can learn to implement ridge
 regression, which is the Bayes-optimal predictor, given sufficient capacity (Ak
yurek et al., 2023), while one-layer transformers with linear self-attention and
 no MLP layer will learn to implement one step of gradient descent (GD) on a lea
st-squares linear regression objective (von Oswald et al., 2022). However, the t
heory behind these observations remains poorly understood. We theoretically stud
y transformers with a single layer of linear self-attention, trained on syntheti
c noisy linear regression data. First, we mathematically show that when the cova
riates are drawn from a standard Gaussian distribution, the one-layer transforme
r which minimizes the pre-training loss will implement a single step of GD on th
e least-squares linear regression objective. Then, we find that changing the dis
tribution of the covariates and weight vector to a non-isotropic Gaussian distri
bution has a strong impact on the learned algorithm: the global minimizer of the
 pre-training loss now implements a single step of $\textit{pre-conditioned}$ GD
. However, if only the distribution of the responses is changed, then this does
not have a large effect on the learned algorithm: even when the response comes f
rom a more general family of $\textit{nonlinear}$ functions, the global minimize
r of the pre-training loss still implements a single step of GD on a least-squar
es linear regression objective.
****
Beatrix Miranda Ginn Nielsen,Anders Christensen,Andrea Dittadi,Ole Winther
DiffEnc: Variational Diffusion with a Learned Encoder
Diffusion models may be viewed as hierarchical variational autoencoders (VAEs) w
ith two improvements: parameter sharing for the conditionals in the generative p
rocess and efficient computation of the loss as independent terms over the hiera
rchy. We consider two changes to the diffusion model that retain these advantage
s while adding flexibility to the model. Firstly, we introduce a data and depth-
dependent mean function in the diffusion process, which leads to a modified diff
usion loss. Our proposed framework, DiffEnc, achieves a statistically significan

t improvement in likelihood on CIFAR-10. Secondly, we let the ratio of the noise variance of the reverse encoder process and the generative process be a free weight parameter rather than being fixed to one. This leads to theoretical insights: For a finite depth hierarchy, the evidence lower bound (ELBO) can be used as an objective for a weighted diffusion loss approach and for optimizing the noise schedule specifically for inference. For the infinite-depth hierarchy, on the other hand, the weight parameter has to be one to have a well-defined ELBO.
****
Sai Surya Duvvuri,Fnu Devvrit,Rohan Anil,Cho-Jui Hsieh,Inderjit S Dhillon
Combining Axes Preconditioners through Kronecker Approximation for Deep Learning
Adaptive regularization based optimization methods such as full-matrix Adagrad which use gradient second-moment information hold significant potential for fast convergence in deep neural network (DNN) training, but are memory intensive and computationally demanding for large neural nets. We develop a technique called Combining AxeS PReconditioners (CASPR), which optimizes matrix-shaped DNN parameters by finding different preconditioners for each mode/axis of the parameter and combining them using a Kronecker-sum based approximation. We show tighter convergence guarantees in stochastic optimization compared to a Kronecker product based preconditioner, Shampoo, which arises as a special case of CASPR. Furthermore, our experiments demonstrates that CASPR approximates the gradient second-moment matrix in full-matrix Adagrad more accurately, and shows significant improvement in training and generalization performance compared to existing practical adaptive regularization based methods such as Shampoo and Adam in a variety of tasks including graph neural network on OGBG-molpcba, Transformer on a universal dependencies dataset and auto-regressive large language modeling on C4 dataset.
****
Fabian Mentzer,David Minnen,Eirikur Agustsson,Michael Tschannen
Finite Scalar Quantization: VQ-VAE Made Simple
We propose to replace vector quantization (VQ) in the latent representation of VQ-VAEs
with a simple scheme termed finite scalar quantization (FSQ), where we project the VAE representation down to a few dimensions (typically less than 10).
Each dimension is quantized to a small set of fixed values, leading to an (implicit) codebook given by the product of these sets.
By appropriately choosing the number of dimensions and values each dimension can take, we obtain the same codebook size as in VQ.
On top of such discrete representations,
we can train the same models that have been trained on VQ-VAE representations. For example, autoregressive and masked transformer models for image generation, multimodal generation, and dense prediction computer vision tasks.
Concretely, we employ FSQ with MaskGIT for image generation, and with UViM for depth estimation, colorization, and panoptic segmentation.
Despite the much simpler design of FSQ, we obtain competitive performance in all these tasks.
We emphasize that FSQ does not suffer from codebook collapse and does not need the complex machinery employed in VQ (commitment losses, codebook reseeding, code splitting, entropy penalties, etc.) to learn expressive discrete representations.
****
Jiawei Liang,Siyuan Liang,Aishan Liu,Xiaojun Jia,Junhao Kuang,Xiaochun Cao
Poisoned Forgery Face: Towards Backdoor Attacks on Face Forgery Detection
The proliferation of face forgery techniques has raised significant concerns within society, thereby motivating the development of face forgery detection methods. These methods aim to distinguish forged faces from genuine ones and have proven effective in practical applications. However, this paper introduces a novel and previously unrecognized threat in face forgery detection scenarios caused by backdoor attack. By embedding backdoors into models and incorporating specific trigger patterns into the input, attackers can deceive detectors into producing erroneous predictions for forged faces. To achieve this goal, this paper proposes \emph{Poisoned Forgery Face} framework, which enables clean-label backdoor atta

cks on face forgery detectors. Our approach involves constructing a scalable tri
gger generator and utilizing a novel convolving process to generate translation-
sensitive trigger patterns. Moreover, we employ a relative embedding method base
d on landmark-based regions to enhance the stealthiness of the poisoned samples.
 Consequently, detectors trained on our poisoned samples are embedded with backd
oors. Notably, our approach surpasses SoTA backdoor baselines with a significant
 improvement in attack success rate (+16.39\% BD-AUC) and reduction in visibilit
y (-12.65\% L_∞). Furthermore, our attack exhibits promising performance
against backdoor defenses. We anticipate that this paper will draw greater atten
tion to the potential threats posed by backdoor attacks in face forgery detectio
n scenarios. Our codes will be made available at \url{https://github.com/JWLiang
007/PFF}.
****
Cong Lei,Yuxuan Du,Peng Mi,Jun Yu,Tongliang Liu
Neural Auto-designer for Enhanced Quantum Kernels
Quantum kernels hold great promise for offering computational advantages over cl
assical learners, with the effectiveness of these kernels closely tied to the de
sign of the feature map. However, the challenge of designing effective quantum f
eature maps for real-world datasets, particularly in the absence of sufficient p
rior information, remains a significant obstacle. In this study, we present a da
ta-driven approach that automates the design of problem-specific quantum feature
 maps. Our approach leverages feature-selection techniques to handle high-dimens
ional data on near-term quantum machines with limited qubits, and incorporates a
 deep neural predictor to efficiently evaluate the performance of various candid
ate quantum kernels. Through extensive numerical simulations on different datase
ts, we demonstrate the superiority of our proposal over prior methods, especiall
y for the capability of eliminating the kernel concentration issue and identifyi
ng the feature map with prediction advantages. Our work not only unlocks the po
tential of quantum kernels for enhancing real-world tasks, but also highlights t
he substantial role of deep learning in advancing quantum machine learning.
****
Zitao Song,Wendi Ren,Shuang Li
Amortized Network Intervention to Steer the Excitatory Point Processes
We tackle the challenge of large-scale network intervention for guiding excitato
ry point processes, such as infectious disease spread or traffic congestion cont
rol. Our model-based reinforcement learning method utilizes neural ODEs to captu
re how the networked excitatory point processes will evolve subject to the time-
varying changes in network topology. Our approach incorporates Gradient-Descent
based Model Predictive Control (GD-MPC), offering policy flexibility to accommod
ate prior knowledge and constraints. To address the intricacies of planning and
overcome the high dimensionality inherent to such decision-making problems, we d
esign an Amortize Network Interventions (ANI) framework, allowing for the poolin
g of optimal policies from history and other contexts, while ensuring a permutat
ion equivalent property. This property enables efficient knowledge transfer and
sharing across diverse contexts. Our approach has broad applications, from curbi
ng infectious disease spread to reducing carbon emissions through traffic light
optimization, and thus has the potential to address critical societal and enviro
nmental challenges.
****
Seungone Kim,Jamin Shin,Yejin Cho,Joel Jang,Shayne Longpre,Hwaran Lee,Sangdoo Yu
n,Seongjin Shin,Sungdong Kim,James Thorne,Minjoon Seo
Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models
Recently, GPT-4 has become the de facto evaluator for long-form text generated b
y large language models (LLMs). However, for practitioners and researchers with
large and custom evaluation tasks, GPT-4 is unreliable due to its closed-source
nature, uncontrolled versioning, and prohibitive costs. In this work, we propose
 PROMETHEUS a fully open-source LLM that is on par with GPT-4’s evaluation capab
ilities when the appropriate reference materials (reference answer, score rubric
) are accompanied. For this purpose, we construct a new dataset – FEEDBACK COLLE
CTION – that consists of 1K fine-grained score rubrics, 20K instructions, and 10

0K natural language feedback generated by GPT-4. Using the FEEDBACK COLLECTION, we train PROMETHEUS, a 13B evaluation-specific LLM that can assess any given response based on novel and unseen score rubrics and reference materials provided by the user. Our dataset's versatility and diversity make our model generalize to challenging real-world criteria, such as prioritizing conciseness, child-readability, or varying levels of formality. We show that PROMETHEUS shows a stronger correlation with GPT-4 evaluation compared to ChatGPT on seven evaluation benchmarks (Two Feedback Collection testsets, MT Bench, Vicuna Bench, Flask Eval, MT Bench Human Judgment, and HHH Alignment), showing the efficacy of our model and dataset design. During human evaluation with hand-crafted score rubrics, PROMETHEUS shows a Pearson correlation of 0.897 with human evaluators, which is on par with GPT-4-0613 (0.882), and greatly outperforms ChatGPT (0.392). Remarkably, when assessing the quality of the generated feedback, PROMETHEUS demonstrates a win rate of 58.62% when compared to GPT-4 evaluation and a win rate of 79.57% when compared to ChatGPT evaluation. Our findings suggests that by adding reference materials and training on GPT-4 feedback, we can obtain effective open-source evaluator LMs.
****
Linlin Yu,Yifei Lou,Feng Chen
Uncertainty-aware Graph-based Hyperspectral Image Classification
Hyperspectral imaging (HSI) technology captures spectral information across a broad wavelength range, providing richer pixel features compared to traditional color images with only three channels. Although pixel classification in HSI has been extensively studied, especially using graph convolution neural networks (GCNs), quantifying epistemic and aleatoric uncertainties associated with the HSI classification (HSIC) results remains an unexplored area. These two uncertainties are effective for out-of-distribution (OOD) and misclassification detection, respectively. In this paper, we adapt two advanced uncertainty quantification models, evidential GCNs (EGCN) and graph posterior networks (GPN), designed for node classifications in graphs, into the realm of HSIC. We first reveal theoretically that a popular uncertainty cross-entropy (UCE) loss function is insufficient to produce good epistemic uncertainty when learning EGCNs. To mitigate the limitations, we propose two regularization terms. One leverages the inherent property of HSI data where each feature vector is a linear combination of the spectra signatures of the confounding materials, while the other is the total variation (TV) regularization to enforce the spatial smoothness of the evidence with edge-preserving. We demonstrate the effectiveness of the proposed regularization terms on both EGCN and GPN on three real-world HSIC datasets for OOD and misclassification detection tasks. The code is available at GitHub.
****
Wenqi Shao,Mengzhao Chen,Zhaoyang Zhang,Peng Xu,Lirui Zhao,Zhiqian Li,Kaipeng Zhang,Peng Gao,Yu Qiao,Ping Luo
OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models
Large language models (LLMs) have revolutionized natural language processing tasks. However, their practical deployment is hindered by their immense memory and computation requirements. Although recent post-training quantization (PTQ) methods are effective in reducing memory footprint and improving the computational efficiency of LLM, they hand-craft quantization parameters, leading to low performance, especially in extremely low-bit quantization. To tackle this issue, we introduce an Omnidirectionally calibrated Quantization ($\textbf{OmniQuant}$) technique for LLMs, which achieves good performance in diverse quantization settings while maintaining the computational efficiency of PTQ by efficiently optimizing various quantization parameters. OmniQuant comprises two innovative components including Learnable Weight Clipping (LWC) and Learnable Equivalent Transformation (LET). LWC modulates the extreme values of weights by optimizing the clipping threshold. Meanwhile, LET tackles activation outliers by shifting the challenge of quantization from activations to weights. Operating within a differentiable framework using block-wise error minimization, OmniQuant can optimize the quantization process efficiently for both weight-only and weight-activation quantization. For instance, the LLaMA-2 model family size 7-70B can be processed with OmniQu

ant on a single A100-40G GPU within 1-16 hours using 128 samples. Extensive experiments validate OmniQuant's superior performance across diverse quantization configurations such as W4A4 (4-bit weight, 4-bit activation), W6A6, W4A16, W3A16, and W2A16. Additionally, OmniQuant demonstrates effectiveness in instruction-tuned models and delivers notable improvements in inference speed and memory reduction on real devices. Codes are available at
\url{https://github.com/OpenGVLab/OmniQuant}.
****
Sharut Gupta,Stefanie Jegelka,David Lopez-Paz,Kartik Ahuja
Context is Environment
Two lines of work are taking the central stage in AI research. On the one hand, the community is making increasing efforts to build models that discard spurious correlations and generalize better in novel test environments. Unfortunately, the hard lesson so far is that no proposal convincingly outperforms a simple empirical risk minimization baseline. On the other hand, large language models (LLMs) have erupted as algorithms able to learn in-context, generalizing on-the-fly to eclectic contextual circumstances that users enforce by means of prompting. In this paper, we argue that context is environment, and posit that in-context learning holds the key to better domain generalization. Via extensive theory and experiments, we show that paying attention to context$\unicode{x2013}\unicode{x2013}$unlabeled examples as they arrive$\unicode{x2013}\unicode{x2013}$allows our proposed In-Context Risk Minimization (ICRM) algorithm to zoom-in on the test environment risk minimizer, leading to significant out-of-distribution performance improvements. Furthermore, training with context helps the model learn a better featurizer. From all of this, two messages are worth taking home. Researchers in domain generalization should consider environment as context, and harness the adaptive power of in-context learning. Researchers in LLMs should consider context as environment, to better structure data towards generalization. Code is available at https://github.com/facebookresearch/ICRM.
****
Haitao Mao,Juanhui Li,Harry Shomer,Bingheng Li,Wenqi Fan,Yao Ma,Tong Zhao,Neil Shah,Jiliang Tang
Revisiting Link Prediction: a data perspective
Link prediction, a fundamental task on graphs, has proven indispensable in various applications, e.g., friend recommendation, protein analysis, and drug interaction prediction. However, since datasets span a multitude of domains, they could have distinct underlying mechanisms of link formation. Evidence in existing literature underscores the absence of a universally best algorithm suitable for all datasets. In this paper, we endeavor to explore principles of link prediction across diverse datasets from a data-centric perspective. We recognize three fundamental factors critical to link prediction: local structural proximity, global structural proximity, and feature proximity. We then unearth relationships among those factors where (i) global structural proximity only shows effectiveness when local structural proximity is deficient. (ii) The incompatibility can be found between feature and structural proximity. Such incompatibility leads to GNNs for Link Prediction (GNN4LP) consistently underperforming on edges where the feature proximity factor dominates. Inspired by these new insights from a data perspective, we offer practical instruction for GNN4LP model design and guidelines for selecting appropriate benchmark datasets for more comprehensive evaluations.
****
Haowei Lin,Yijia Shao,Weinan Qian,Ningxin Pan,Yiduo Guo,Bing Liu
Class Incremental Learning via Likelihood Ratio Based Task Prediction
Class incremental learning (CIL) is a challenging setting of continual learning, which learns a series of tasks sequentially. Each task consists of a set of unique classes. The key feature of CIL is that no task identifier (or task-id) is provided at test time. Predicting the task-id for each test sample is a challenging problem. An emerging theory-guided approach (called TIL+OOD) is to train a task-specific model for each task in a shared network for all tasks based on a task-incremental learning (TIL) method to deal with catastrophic forgetting. The model for each task is an out-of-distribution (OOD) detector rather than a convent

ional classifier. The OOD detector can perform both within-task (in-distribution
(IND)) class prediction and OOD detection. The OOD detection capability is the
key to task-id prediction during inference. However, this paper argues that usin
g a traditional OOD detector for task-id prediction is sub-optimal because addit
ional information (e.g., the replay data and the learned tasks) available in CIL
can be exploited to design a better and principled method for task-id predictio
n. We call the new method TPL (Task-id Prediction based on Likelihood Ratio). TP
L markedly outperforms strong CIL baselines and has negligible catastrophic forg
etting. The code of TPL is publicly available at https://github.com/linhaowei1/T
PL.
****
Jiayu Xiao,Henglei Lv,Liang Li,Shuhui Wang,Qingming Huang
R&B: Region and Boundary Aware Zero-shot Grounded Text-to-image Generation
Recent text-to-image (T2I) diffusion models have achieved remarkable progress in
generating high-quality images given text-prompts as input. However, these mode
ls fail to convey appropriate spatial composition specified by a layout instruct
ion. In this work, we probe into zero-shot grounded T2I generation with diffusio
n models, that is, generating images corresponding to the input layout informati
on without training auxiliary modules or finetuning diffusion models. We propose
a **R**egion and **B**oundary (R&B) aware cross-attention guidance approach tha
t gradually modulates the attention maps of diffusion model during generative pr
ocess, and assists the model to synthesize images (1) with high fidelity, (2) hi
ghly compatible with textual input, and (3) interpreting layout instructions acc
urately. Specifically, we leverage the discrete sampling to bridge the gap betwe
en consecutive attention maps and discrete layout constraints, and design a regi
on-aware loss to refine the generative layout during diffusion process. We furth
er propose a boundary-aware loss to strengthen object discriminability within th
e corresponding regions. Experimental results show that our method outperforms e
xisting state-of-the-art zero-shot grounded T2I generation methods by a large ma
rgin both qualitatively and quantitatively on several benchmarks.
Project page: https://sagileo.github.io/Region-and-Boundary.
****
Juncai He,Xinliang Liu,Jinchao Xu
MgNO: Efficient Parameterization of Linear Operators via Multigrid
In this work, we propose a concise neural operator architecture for operator lea
rning. Drawing an analogy with a conventional fully connected neural network, we
define the neural operator as follows: the output of the i-th neuron in a non
linear operator layer is defined by $\mathcal O_i(u) = \sigma\left(\sum_j \mat
hcal W_{ij} u + \mathcal B_{ij}\right)$. Here, $\mathcal W_{ij}$ denotes the bou
nded linear operator connecting j-th input neuron to i-th output neuron, and
the bias $\mathcal B_{ij}$ takes the form of a function rather than a scalar. G
iven its new universal approximation property, the efficient parameterization of
the bounded linear operators between two neurons (Banach spaces) plays a critic
al role. As a result, we introduce MgNO, utilizing multigrid structures to param
eterize these linear operators between neurons. This approach offers both mathe
matical rigor and practical expressivity. Additionally, MgNO obviates the need f
or conventional lifting and projecting operators typically required in previous
neural operators. Moreover, it seamlessly accommodates diverse boundary conditio
ns. Our empirical observations reveal that MgNO exhibits superior ease of traini
ng compared to CNN-based models, while also displaying a reduced susceptibility
to overfitting when contrasted with spectral-type neural operators. We demonstra
te the efficiency and accuracy of our method with consistently state-of-the-art
performance on different types of partial differential equations (PDEs).
****
Chengyu Dong,Liyuan Liu,Hao Cheng,Jingbo Shang,Jianfeng Gao,Xiaodong Liu
Fast-ELECTRA for Efficient Pre-training
ELECTRA pre-trains language models by detecting tokens in a sequence that have b
een replaced by an auxiliary model. Although ELECTRA offers a significant boost
in efficiency, its potential is constrained by the training cost brought by the
auxiliary model. Notably, this model, which is jointly trained with the main mod

el, only serves to assist the training of the main model and is discarded post-training. This results in a substantial amount of training cost being expended in vain. To mitigate this issue, we propose Fast-ELECTRA, which leverages an existing language model as the auxiliary model. To construct a learning curriculum for the main model, we smooth its output distribution via temperature scaling following a descending schedule. Our approach rivals the performance of state-of-the-art ELECTRA-style pre-training methods, while significantly eliminating the computation and memory cost brought by the joint training of the auxiliary model. Our method also reduces the sensitivity to hyper-parameters and enhances the pre-training stability.
****
Shangmin Guo,Yi Ren,Stefano V Albrecht,Kenny Smith
lpNTK: Better Generalisation but Less Data via Sample Interaction during Learning
Although much research has been done on proposing new models or loss functions to improve the generalisation of artificial neural networks (ANNs), less attention has been directed to the impact of the training data on generalisation. In this work, we start from approximating the interaction between samples, i.e. how learning one sample would modify the model's prediction on other samples. Through analysing the terms involved in weight updates in supervised learning, we find that labels influence the interaction between samples. Therefore, we propose the labelled pseudo Neural Tangent Kernel (lpNTK) which takes label information into consideration when measuring the interactions between samples. We first prove that lpNTK asymptotically converges to the empirical neural tangent kernel in terms of the Frobenius norm under certain assumptions. Secondly, we illustrate how lpNTK helps to understand learning phenomena identified in previous work, specifically the learning difficulty of samples and forgetting events during learning. Moreover, we also show that using lpNTK to identify and remove poisoning training samples does not hurt the generalisation performance of ANNs.
****
Haoxuan You,Mandy Guo,Zhecan Wang,Kai-Wei Chang,Jason Michael Baldridge,Jiahui Yu
CoBIT: A Contrastive Bi-directional Image-Text Generation Model
The field of Vision-and-Language (VL) has witnessed a proliferation of pretrained foundation models. Current techniques typically employ only one type of training objective, whether it's (1) contrastive objectives (like CLIP), (2) image-to-text generative objectives (like PaLI), or (3) text-to-image generative objectives (like Parti). However, all these three objectives are mutually relevant and are all based on image-text pairs. Intuitively, the first two objectives can be considered as complementary projections between two modalities, and contrastive learning can preserve global alignment and generations facilitate fine-grained understanding. Inspired by this, we present a Contrastive Bi-directional Image-Text generation model (CoBIT) to first time unify the three pre-training objectives in one framework. Specifically, CoBIT employs a novel unicoder-decoder structure consisting of an image unicoder, a text unicoder, and a cross-modal decoder. The image/text unicoders can switch between encoding and decoding in different tasks, enabling flexibility and shared knowledge that benefits both image-to-text and text-to-image generations. CoBIT achieves superior performance in image understanding, image-text understanding (Retrieval, Captioning, VQA, SNLI-VE), and text-based content creation, particularly in zero-shot scenarios.
****
Guillaume Bono,Leonid Antsfeld,Assem Sadek,Gianluca Monaci,Christian Wolf
Learning with a Mole: Transferable latent spatial representations for navigation without reconstruction
Agents navigating in 3D environments require some form of memory, which should hold a compact and actionable representation of the history of observations useful for decision taking and planning. In most end-to-end learning approaches the representation is latent and usually does not have a clearly defined interpretation, whereas classical robotics addresses this with scene reconstruction resulting in some form of map, usually estimated with geometry and sensor models and/or

learning. In this work we propose to learn an actionable representation of the scene independently of the targeted downstream task and without explicitly optimizing reconstruction. The learned representation is optimized by a blind auxiliary agent trained to navigate with it on multiple short sub episodes branching out from a waypoint and, most importantly, without any direct visual observation. We argue and show that the blindness property is important and forces the (trained) latent representation to be the only means for planning. With probing experiments we show that the learned representation optimizes navigability and not reconstruction. On downstream tasks we show that it is robust to changes in distribution, in particular the sim2real gap, which we evaluate with a real physical robot in a real office building, significantly improving performance.
****
Xueyang Tang,Song Guo,Jie ZHANG,Jingcai Guo
Learning Personalized Causally Invariant Representations for Heterogeneous Federated Clients
Personalized federated learning (PFL) has gained great success in tackling the scenarios where target datasets are heterogeneous across the local clients. However, the application of the existing PFL methods to real-world setting is hindered by the common assumption that the test data on each client is in-distribution (IND) with respect to its training data. Due to the bias of training dataset, the modern machine learning model prefers to rely on shortcut which can perform well on the training data but fail to generalize to the unseen test data that is out-of-distribution (OOD). This pervasive phenomenon is called shortcut learning and has attracted plentiful efforts in centralized situations. In PFL, the limited data diversity on federated clients makes mitigating shortcut and meanwhile preserving personalization knowledge rather difficult. In this paper, we analyse this challenging problem by formulating the structural causal models (SCMs) for heterogeneous federated clients. From the proposed SCMs, we derive two significant causal signatures which inspire a provable shortcut discovery and removal method under federated learning, namely FedSDR. Specifically, FedSDR is divided into two steps: 1) utilizing the available training data distributed among local clients to discover all the shortcut features in a collaborative manner. 2) developing the optimal personalized causally invariant predictor for each client by eliminating the discovered shortcut features. We provide theoretical analysis to prove that our method can draw complete shortcut features and produce the optimal personalized invariant predictor that can generalize to unseen OOD data on each client. The experimental results on diverse datasets validate the superiority of FedSDR over the state-of-the-art PFL methods on OOD generalization performance.
****
June Yong Yang,Geondo Park,Joowon Kim,Hyeongwon Jang,Eunho Yang
Language-Interfaced Tabular Oversampling via Progressive Imputation and Self-Authentication
Tabular data in the wild are frequently afflicted with class-imbalance, biasing machine learning model predictions towards major classes. A data-centric solution to this problem is oversampling - where the classes are balanced by adding synthetic minority samples via generative methods. However, although tabular generative models are capable of generating synthetic samples under a balanced distribution, their integrity suffers when the number of minority samples is low. To this end, pre-trained generative language models with rich prior knowledge are a fitting candidate for the task at hand. Nevertheless, an oversampling strategy tailored for tabular data that utilizes the extensive capabilities of such language models is yet to emerge. In this paper, we propose a novel oversampling framework for tabular data to channel the abilities of generative language models. By leveraging its conditional sampling capabilities, we synthesize minority samples by progressively masking the important features of the majority class samples and imputing them towards the minority distribution. To reduce the inclusion of imperfectly converted samples, we utilize the power of the language model itself to self-authenticate the labels of the samples generated by itself, sifting out ill-converted samples. Extensive experiments on a variety of datasets and imbala

nce ratios reveal that the proposed method successfully generates reliable minority samples to boost the performance of machine learning classifiers, even under heavy imbalance ratios.
****
Sen Cui,Abudukelimu Wuerkaixi,Weishen Pan,Jian Liang,Lei Fang,Changshui Zhang,Fei Wang
CLAP: Collaborative Adaptation for Patchwork Learning
In this paper, we investigate a new practical learning scenario, where the data distributed in different sources/clients are typically generated with various modalities. Existing research on learning from multi-source data mostly assume that each client owns the data of all modalities, which may largely limit its practicability. In light of the expensiveness and sparsity of multimodal data, we propose patchwork learning to jointly learn from fragmented multimodal data in distributed clients. Considering the concerns on data privacy, patchwork learning aims to impute incomplete multimodal data for diverse downstream tasks without accessing the raw data directly. Local clients could miss different modality combinations. Due to the statistical heterogeneity induced by non-i.i.d. data, the imputation is more challenging since the learned dependencies fail to adapt to the imputation of other clients. In this paper, we provide a novel imputation framework to tackle modality combination heterogeneity and statistical heterogeneity simultaneously, called ``collaborative adaptation''. In particular, for two observed modality combinations from two clients, we learn the transformations between their maximal intersection and other modalities by proposing a novel ELBO. We improve the worst-performing required transformations through a Pareto min-max optimization framework. In extensive experiments, we demonstrate the superiority of the proposed method compared to existing related methods on benchmark data sets and a real-world clinical data set.
****
Mao Hong,Zhengling Qi,Yanxun Xu
A Policy Gradient Method for Confounded POMDPs
In this paper, we propose a policy gradient method for confounded partially observable Markov decision processes (POMDPs) with continuous state and observation spaces in the offline setting. We first establish a novel identification result to non-parametrically estimate any history-dependent policy gradient under POMDPs using the offline data. The identification enables us to solve a sequence of conditional moment restrictions and adopt the min-max learning procedure with general function approximation for estimating the policy gradient. We then provide a finite-sample non-asymptotic bound for estimating the gradient uniformly over a pre-specified policy class in terms of the sample size, length of horizon, concentratability coefficient and the measure of ill-posedness in solving the conditional moment restrictions. Lastly, by deploying the proposed gradient estimation in the gradient ascent algorithm, we show the global convergence of the proposed algorithm in finding the history-dependent optimal policy under some technical conditions. To the best of our knowledge, this is the first work studying the policy gradient method for POMDPs under the offline setting.
****
Artur Back de Luca,Kimon Fountoulakis,Shenghao Yang
Local Graph Clustering with Noisy Labels
The growing interest in machine learning problems over graphs with additional node information such as texts, images, or labels has popularized methods that require the costly operation of processing the entire graph. Yet, little effort has been made to the development of fast local methods (i.e. without accessing the entire graph) that extract useful information from such data. To that end, we propose a study of local graph clustering using noisy node labels as a proxy for additional node information. In this setting, nodes receive initial binary labels based on cluster affiliation: 1 if they belong to the target cluster and 0 otherwise. Subsequently, a fraction of these labels is flipped. We investigate the benefits of incorporating noisy labels for local graph clustering. By constructing a weighted graph with such labels, we study the performance of graph diffusion-based local clustering method on both the original and the weighted graphs. Fro

m a theoretical perspective, we consider recovering an unknown target cluster wi
th a single seed node in a random graph with independent noisy node labels. We p
rovide sufficient conditions on the label noise under which, with high probabili
ty, using diffusion in the weighted graph yields a more accurate recovery of the
 target cluster. This approach proves more effective than using the given labels
 alone or using diffusion in the label-free original graph. Empirically, we show
 that reliable node labels can be obtained with just a few samples from an attri
buted graph. Moreover, utilizing these labels via diffusion in the weighted grap
h leads to significantly better local clustering performance across several real
-world datasets, improving F1 scores by up to 13\%.
****
Han Zhang,Yu Lei,Lin Gui,Min Yang,Yulan He,Hui Wang,Ruifeng Xu
CPPO: Continual Learning for Reinforcement Learning with Human Feedback
The approach of Reinforcement Learning from Human Feedback (RLHF) is widely used
 for enhancing pre-trained Language Models (LM), enabling them to better align w
ith human preferences. Existing RLHF-based LMs however require complete retraini
ng whenever new queries or feedback are introduced, as human preferences may dif
fer across different domains or topics. LM retraining is of■ten impracticable in
 most real-world scenarios, due to the substantial time and computational costs
involved, as well as data privacy concerns. To address this limitation, we propo
se Continual Proximal Policy Optimization (CPPO), a novel method that is able to
 continually align LM with dynamic human preferences. Specifically, CPPO adopts
a weighting strategy to decide which samples should be utilized for enhancing po
licy learning and which should be used for solidifying past experiences. This se
eks a good trade-off between policy learning and knowledge retention. Our experi
mental results show that CPPO outperforms strong Contin■uous learning (CL) basel
ines when it comes to consistently aligning with human preferences. Furthermore,
 compared to PPO, CPPO offers more efficient and stable learning in non-continua
l scenarios.
****
Sreyan Ghosh,Ashish Seth,Sonal Kumar,Utkarsh Tyagi,Chandra Kiran Reddy Evuru,Ram
aneswaran S,S Sakshi,Oriol Nieto,Ramani Duraiswami,Dinesh Manocha
CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models
A fundamental characteristic of audio is its compositional nature. Audio-languag
e models (ALMs) trained using a contrastive approach (e.g., CLAP) that learns a
shared representation between audio and language modalities have improved perfor
mance in many downstream applications, including zero-shot audio classification,
 audio retrieval, etc. However, the ability of these models to effectively perfo
rm compositional reasoning remains largely unexplored and necessitates additiona
l research. In this paper, we propose CompA, a collection of two expert-annotate
d benchmarks with a majority of real-world audio samples, to evaluate compositio
nal reasoning in ALMs. Our proposed CompA-order evaluates how well an ALM unders
tands the order or occurrence of acoustic events in audio, and CompA-attribute e
valuates attribute-binding of acoustic events. An instance from either benchmark
 consists of two audio-caption pairs, where both audios have the same acoustic e
vents but with different compositions. An ALM is evaluated on how well it matche
s the right audio to the right caption. Using this benchmark, we first show that
 current ALMs perform only marginally better than random chance, thereby struggl
ing with compositional reasoning. Next, we propose CompA-CLAP, where we fine-tun
e CLAP using a novel learning method to improve its compositional reasoning abil
ities. To train CompA-CLAP, we first propose improvements to contrastive trainin
g with composition-aware hard negatives, allowing for more focused training. Nex
t, we propose a novel modular contrastive loss that helps the model learn fine-g
rained compositional understanding and overcomes the acute scarcity of openly av
ailable compositional audios. CompA-CLAP significantly improves over all our bas
eline models on the CompA benchmark, indicating its superior compositional reaso
ning capabilities.
****
Yuxin Wen,Yuchen Liu,Chen Chen,Lingjuan Lyu
Detecting, Explaining, and Mitigating Memorization in Diffusion Models

Recent breakthroughs in diffusion models have exhibited exceptional image-generation capabilities. However, studies show that some outputs are merely replications of training data. Such replications present potential legal challenges for model owners, especially when the generated content contains proprietary information. In this work, we introduce a straightforward yet effective method for detecting memorized prompts by inspecting the magnitude of text-conditional predictions. Our proposed method seamlessly integrates without disrupting sampling algorithms, and delivers high accuracy even at the first generation step, with a single generation per prompt. Building on our detection strategy, we unveil an explainable approach that shows the contribution of individual words or tokens to memorization. This offers an interactive medium for users to adjust their prompts. Moreover, we propose two strategies i.e., to mitigate memorization by leveraging the magnitude of text-conditional predictions, either through minimization during inference or filtering during training. These proposed strategies effectively counteract memorization while maintaining high-generation quality. Code is available at https://github.com/YuxinWenRick/diffusion_memorization.
****

Yuxing Tian,Yiyan Qi,Fan Guo
FreeDyG: Frequency Enhanced Continuous-Time Dynamic Graph Model for Link Prediction
Link prediction is a crucial task in dynamic graph learning. Recent advancements in continuous-time dynamic graph models, primarily by leveraging richer temporal details, have significantly improved link prediction performance. However, due to their complex modules, they still face several challenges, such as overfitting and optimization difficulties. More importantly, it is challenging for these methods to capture the 'shift' phenomenon, where node interaction patterns change over time. To address these issues, we propose a simple yet novel method called \textbf{Fre}quency \textbf{E}nhanced Continuous-Time \textbf{Dy}namic \textbf{G}raph ({\bf FreeDyG}) model for link prediction. Specifically, we propose a node interaction frequency encoding module that both explicitly captures the proportion of common neighbors and the frequency of the interaction of the node pair. Unlike previous works that primarily focus on the time domain, we delve into the frequency domain, allowing a deeper and more nuanced extraction of interaction patterns, revealing periodic and "shift" behaviors. Extensive experiments conducted on seven real-world continuous-time dynamic graph datasets validate the effectiveness of FreeDyG. The results consistently demonstrate that FreeDyG outperforms existing methods in both transductive and inductive settings. Our code is available at this repository: \href{https://github.com/Tianxzzz/FreeDyG}{https://github.com/Tianxzzz/FreeDyG}
****

Fabian Akkerman,Julius Luy,Wouter van Heeswijk,Maximilian Schiffer
Dynamic Neighborhood Construction for Structured Large Discrete Action Spaces
Large discrete action spaces (LDAS) remain a central challenge in reinforcement learning. Existing solution approaches can handle unstructured LDAS with up to a few million actions. However, many real-world applications in logistics, production, and transportation systems have combinatorial action spaces, whose size grows well beyond millions of actions, even on small instances. Fortunately, such action spaces exhibit structure, e.g., equally spaced discrete resource units. With this work, we focus on handling structured LDAS (SLDAS) with sizes that cannot be handled by current benchmarks: we propose Dynamic Neighborhood Construction (DNC), a novel exploitation paradigm for SLDAS. We present a scalable neighborhood exploration heuristic that utilizes this paradigm and efficiently explores the discrete neighborhood around the continuous proxy action in structured action spaces with up to 10^{73} actions. We demonstrate the performance of our method by benchmarking it against three state-of-the-art approaches designed for large discrete action spaces across three distinct environments. Our results show that DNC matches or outperforms state-of-the-art approaches while being computationally more efficient. Furthermore, our method scales to action spaces that so far remained computationally intractable for existing methodologies.
****

Zhaoyi Zhou,Chuning Zhu,Runlong Zhou,Qiwen Cui,Abhishek Gupta,Simon Shaolei Du
Free from Bellman Completeness: Trajectory Stitching via Model-based Return-conditioned Supervised Learning
Off-policy dynamic programming (DP) techniques such as Q-learning have proven to be important in sequential decision-making problems. In the presence of function approximation, however, these techniques often diverge due to the absence of Bellman completeness in the function classes considered, a crucial condition for the success of DP-based methods. In this paper, we show how off-policy learning techniques based on return-conditioned supervised learning (RCSL) are able to circumvent these challenges of Bellman completeness, converging under significantly more relaxed assumptions inherited from supervised learning. We prove there exists a natural environment in which if one uses two-layer multilayer perceptron as the function approximator, the layer width needs to grow *linearly* with the state space size to satisfy Bellman completeness while a constant layer width is enough for RCSL. These findings take a step towards explaining the superior empirical performance of RCSL methods compared to DP-based methods in environments with near-optimal datasets. Furthermore, in order to learn from sub-optimal datasets, we propose a simple framework called MBRCSL, granting RCSL methods the ability of dynamic programming to stitch together segments from distinct trajectories. MBRCSL leverages learned dynamics models and forward sampling to accomplish trajectory stitching while avoiding the need for Bellman completeness that plagues all dynamic programming algorithms. We propose both theoretical analysis and experimental evaluation to back these claims, outperforming state-of-the-art model-free and model-based offline RL algorithms across several simulated robotics problems.
****
Shujian Yu,Xi Yu,Sigurd Løkse,Robert Jenssen,Jose C Principe
Cauchy-Schwarz Divergence Information Bottleneck for Regression
The information bottleneck (IB) approach is popular to improve the generalization, robustness and explainability of deep neural networks. Essentially, it aims to find a minimum sufficient representation $\mathbf{t}$ by striking a trade-off between a compression term $I(\mathbf{x};\mathbf{t})$ and a prediction term $I(y;\mathbf{t})$, where $I(\cdot;\cdot)$ refers to the mutual information (MI). MI is for the IB for the most part expressed in terms of the Kullback-Leibler (KL) divergence, which in the regression case corresponds to prediction based on mean squared error (MSE) loss with Gaussian assumption and compression approximated by variational inference.
In this paper, we study the IB principle for the regression problem and develop a new way to parameterize the IB with deep neural networks by exploiting favorable properties of the Cauchy-Schwarz (CS) divergence. By doing so, we move away from MSE-based regression and ease estimation by avoiding variational approximations or distributional assumptions. We investigate the improved generalization ability of our proposed CS-IB and demonstrate strong adversarial robustness guarantees. We demonstrate its superior performance on six real-world regression tasks over other popular deep IB approaches. We additionally observe that the solutions discovered by CS-IB always achieve the best trade-off between prediction accuracy and compression ratio in the information plane. The code is available at \url{https://github.com/SJYuCNEL/Cauchy-Schwarz-Information-Bottleneck}.
****
Tuo Xu,Lei Zou
Rethinking and Extending the Probabilistic Inference Capacity of GNNs
Designing expressive Graph Neural Networks (GNNs) is an important topic in graph machine learning fields. Despite the existence of numerous approaches proposed to enhance GNNs based on Weisfeiler-Lehman (WL) tests, what GNNs can and cannot learn still lacks a deeper understanding. This paper adopts a fundamentally different approach to examine the expressive power of GNNs from a probabilistic perspective. By establishing connections between GNNs' predictions and the central inference problems of probabilistic graphical models (PGMs), we can analyze previous GNN variants with a novel hierarchical framework and gain new insights into their node-level and link-level behaviors. Additionally, we introduce novel meth

ods that can provably enhance GNNs' ability to capture complex dependencies and make complex predictions. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of our approaches.
****
Meirui Jiang,Anjie Le,Xiaoxiao Li,Qi Dou
Heterogeneous Personalized Federated Learning by Local-Global Updates Mixing via Convergence Rate
Personalized federated learning (PFL) has emerged as a promising technique for addressing the challenge of data heterogeneity. While recent studies have made notable progress in mitigating heterogeneity associated with label distributions, the issue of effectively handling feature heterogeneity remains an open question. In this paper, we propose a personalization approach by Local-global updates Mixing (LG-Mix) via Neural Tangent Kernel (NTK)-based convergence. The core idea is to leverage the convergence rate induced by NTK to quantify the importance of local and global updates, and subsequently mix these updates based on their importance. Specifically, we find the trace of the NTK matrix can manifest the convergence rate, and propose an efficient and effective approximation to calculate the trace of a feature matrix instead of the NTK matrix. Such approximation significantly reduces the cost of computing NTK, and the feature matrix explicitly considers the heterogeneous features among samples. We have theoretically analyzed the convergence of our method in the over-parameterize regime, and experimentally evaluated our method on five datasets. These datasets present heterogeneous data features in natural and medical images. With comprehensive comparison to existing state-of-the-art approaches, our LG-Mix has consistently outperformed them across all datasets (largest accuracy improvement of 5.01\%), demonstrating the outstanding efficacy of our method for model personalization. Code is available at \url{https://github.com/med-air/HeteroPFL}.
****
Shiyu Wang,Haixu Wu,Xiaoming Shi,Tengge Hu,Huakun Luo,Lintao Ma,James Y. Zhang,JUN ZHOU
TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting
Time series forecasting is widely used in extensive applications, such as traffic planning and weather forecasting. However, real-world time series usually present intricate temporal variations, making forecasting extremely challenging. Going beyond the mainstream paradigms of plain decomposition and multiperiodicity analysis, we analyze temporal variations in a novel view of multiscale-mixing, where time series present distinct patterns in different sampling scales. Specifically, the microscopic and the macroscopic information are reflected in fine and coarse scales, respectively, and thereby complex variations are inherently disentangled. Based on this observation, we propose TimeMixer as a fully MLP-based architecture with Past-Decomposable-Mixing (PDM) and Future-Multipredictor-Mixing (FMM) blocks to take full advantage of disentangled multiscale series in both past extraction and future prediction phases. Concretely, PDM applies the decomposition to multiscale series and further mixes the decomposed seasonal and trend components in fine-to-coarse and coarse-to-fine directions separately, which successively aggregates the microscopic seasonal and macroscopic trend information. FMM further ensembles multiple predictors to utilize complementary forecasting capabilities in multiscale observations. Consequently, our proposed TimeMixer is able to achieve consistent state-of-the-art performances in both long-term and short-term forecasting tasks with favorable run-time efficiency.
****
Shaofeng Zhang,Jinfa Huang,Qiang Zhou,zhibin wang,Fan Wang,Jiebo Luo,Junchi Yan
Continuous-Multiple Image Outpainting in One-Step via Positional Query and A Diffusion-based Approach
Image outpainting aims to generate the content of an input sub-image beyond its original boundaries. It is an important task in content generation yet remains an open problem for generative models. This paper pushes the technical frontier of image outpainting in two directions that have not been resolved in literature: 1) outpainting with arbitrary and continuous multiples (without restriction), and 2) outpainting in a single step (even for large expansion multiples). Moreove

r, we develop a method that does not depend on a pre-trained backbone network, which is in contrast commonly required by the previous SOTA outpainting methods. The arbitrary multiple outpainting is achieved by utilizing randomly cropped views from the same image during training to capture arbitrary relative positional information. Specifically, by feeding one view and positional embeddings as queries, we can reconstruct another view. At inference, we generate images with arbitrary expansion multiples by inputting an anchor image and its corresponding positional embeddings. The one-step outpainting ability here is particularly noteworthy in contrast to previous methods that need to be performed for N times to obtain a final multiple which is N times of its basic and fixed multiple. We evaluate the proposed approach (called PQDiff as we adopt a diffusion-based generator as our embodiment, under our proposed \textbf{P}ositional \textbf{Q}uery scheme) on public benchmarks, demonstrating its superior performance over state-of-the-art approaches. Specifically, PQDiff achieves state-of-the-art FID scores on the Scenery (\textbf{21.512}), Building Facades (\textbf{25.310}), and WikiArts (\textbf{36.212}) datasets. Furthermore, under the 2.25x, 5x and 11.7x outpainting settings, PQDiff only takes \textbf{40.6\%}, \textbf{20.3\%} and \textbf{10.2\%} of the time of the benchmark state-of-the-art (SOTA) method.
****
Md Mofijul Islam,Alexi Gladstone,Riashat Islam,Tariq Iqbal
EQA-MX: Embodied Question Answering using Multimodal Expression
Humans predominantly use verbal utterances and nonverbal gestures (e.g., eye gaze and pointing gestures) in their natural interactions. For instance, pointing gestures and verbal information is often required to comprehend questions such as "what object is that?" Thus, this question-answering (QA) task involves complex reasoning of multimodal expressions (verbal utterances and nonverbal gestures). However, prior works have explored QA tasks in non-embodied settings, where questions solely contain verbal utterances from a single verbal and visual perspective. In this paper, we have introduced 8 novel embodied question answering (EQA) tasks to develop learning models to comprehend embodied questions with multimodal expressions. We have developed a novel large-scale dataset, EQA-MX, with over 8 million diverse embodied QA data samples involving multimodal expressions from multiple visual and verbal perspectives. To learn salient multimodal representations from discrete verbal embeddings and continuous wrapping of multiview visual representations, we propose a vector-quantization (VQ) based multimodal representation learning model, VQ-Fusion, for the EQA tasks. Our extensive experimental results suggest that VQ-Fusion can improve the performance of existing state-of-the-art visual-language models up to 13% across EQA tasks.
****
Diego Gomez,Michael Bowling,Marlos C. Machado
Proper Laplacian Representation Learning
The ability to learn good representations of states is essential for solving large reinforcement learning problems, where exploration, generalization, and transfer are particularly challenging. The _Laplacian representation_ is a promising approach to address these problems by inducing informative state encoding and intrinsic rewards for temporally-extended action discovery and reward shaping. To obtain the Laplacian representation one needs to compute the eigensystem of the graph Laplacian, which is often approximated through optimization objectives compatible with deep learning approaches. These approximations, however, depend on hyperparameters that are impossible to tune efficiently, converge to arbitrary rotations of the desired eigenvectors, and are unable to accurately recover the corresponding eigenvalues. In this paper we introduce a theoretically sound objective and corresponding optimization algorithm for approximating the Laplacian representation. Our approach naturally recovers both the true eigenvectors and eigenvalues while eliminating the hyperparameter dependence of previous approximations. We provide theoretical guarantees for our method and we show that those results translate empirically into robust learning across multiple environments.
****
Xiaolin Sun,Zizhan Zheng
Belief-Enriched Pessimistic Q-Learning against Adversarial State Perturbations

Reinforcement learning (RL) has achieved phenomenal success in various domains.
However, its data-driven nature also introduces new vulnerabilities that can be
exploited by malicious opponents. Recent work shows that a well-trained RL agent
 can be easily manipulated by strategically perturbing its state observations at
 the test stage. Existing solutions either introduce a regularization term to im
prove the smoothness of the trained policy against perturbations or alternativel
y train the agent's policy and the attacker's policy. However, the former does n
ot provide sufficient protection against strong attacks, while the latter is com
putationally prohibitive for large environments. In this work, we propose a new
robust RL algorithm for deriving a pessimistic policy to safeguard against an ag
ent's uncertainty about true states. This approach is further enhanced with beli
ef state inference and diffusion-based state purification to reduce uncertainty.
 Empirical results show that our approach obtains superb performance under stron
g attacks and has a comparable training overhead with regularization-based metho
ds. Our code is available at https://github.com/SliencerX/Belief-enriched-robust
-Q-learning.
****
Hongwei Ren,Yue Zhou,Xiaopeng LIN,Yulong Huang,Haotian FU,Jie Song,Bojun Cheng
SpikePoint: An Efficient Point-based Spiking Neural Network for Event Cameras Ac
tion Recognition
Event cameras are bio-inspired sensors that respond to local changes in light in
tensity and feature low latency, high energy efficiency, and high dynamic range.
 Meanwhile, Spiking Neural Networks (SNNs) have gained significant attention due
 to their remarkable efficiency and fault tolerance. By synergistically harnessi
ng the energy efficiency inherent in event cameras and the spike-based processin
g capabilities of SNNs, their integration could enable ultra-low-power applicati
on scenarios, such as action recognition tasks. However, existing approaches oft
en entail converting asynchronous events into conventional frames, leading to ad
ditional data mapping efforts and a loss of sparsity, contradicting the design c
oncept of SNNs and event cameras. To address this challenge, we propose SpikePoi
nt, a novel end-to-end point-based SNN architecture. SpikePoint excels at proces
sing sparse event cloud data, effectively extracting both global and local featu
res through a singular-stage structure. Leveraging the surrogate training method
, SpikePoint achieves high accuracy with few parameters and maintains low power
consumption, specifically employing the identity mapping feature extractor on di
verse datasets. SpikePoint achieves state-of-the-art (SOTA) performance on four
event-based action recognition datasets using only 16 timesteps, surpassing othe
r SNN methods. Moreover, it also achieves SOTA performance across all methods on
 three datasets, utilizing approximately 0.3 % of the parameters and 0.5 % of po
wer consumption employed by artificial neural networks (ANNs). These results emp
hasize the significance of Point Cloud and pave the way for many ultra-low-power
 event-based data processing applications.
****
Vaidehi Patil,Peter Hase,Mohit Bansal
Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against
 Extraction Attacks
Pretrained language models sometimes possess knowledge that we do not wish them
to, including memorized personal information and knowledge that could be used to
 harm people. They can also output toxic or harmful text. To mitigate these safe
ty and informational issues, we propose an attack-and-defense framework for stud
ying the task of deleting sensitive information directly from model weights. We
study direct edits to model weights because (1) this approach should guarantee t
hat particular deleted information is never extracted by future prompt attacks,
and (2) it should protect against whitebox attacks, which is necessary for makin
g claims about safety/privacy in a setting where publicly available model weight
s could be used to elicit sensitive information. Our threat model assumes that a
n attack succeeds if the answer to a sensitive question is located among a set o
f B generated candidates, based on scenarios where the information would be inse
cure if the answer is among B candidates. Experimentally, we show that even stat
e-of-the-art model editing methods such as ROME struggle to truly delete factual

 information from models like GPT-J, as our whitebox and blackbox attacks can re
cover “deleted” information from an edited model 38% of the time. These attacks
leverage two key observations: (1) that traces of deleted information can be fou
nd in intermediate model hidden states, and (2) that applying an editing method
for one question may not delete information across rephrased versions of the que
stion. Finally, we provide new defense methods that protect against some extract
ion attacks, but we do not find a single universally effective defense method. O
ur results suggest that truly deleting sensitive information is a tractable but
difficult problem, since even relatively low attack success rates have potential
ly severe implications for the deployment of language models in a world where in
dividuals enjoy ownership of their personal data, a right to privacy, and safety
 from harmful model outputs.
****
Hamidreza Almasi,Harsh Mishra,Balajee Vamanan,Sathya N. Ravi
Flag Aggregator: Scalable Distributed Training under Failures and Augmented Loss
es using Convex Optimization
Modern ML applications increasingly rely on complex deep learning models and lar
ge datasets. There has been an exponential growth in the amount of computation n
eeded to train the largest models. Therefore, to scale computation and data, the
se models are inevitably trained in a distributed manner in clusters of nodes, a
nd their updates are aggregated before being applied to the model. However, a di
stributed setup is prone to Byzantine failures of individual nodes, components,
and software. With data augmentation added to these settings, there is a critica
l need for robust and efficient aggregation systems. We define the quality of wo
rkers as reconstruction ratios $\in (0,1]$, and formulate aggregation as a Maxim
um Likelihood Estimation procedure using Beta densities. We show that the Regula
rized form of log-likelihood wrt subspace can be approximately solved using iter
ative least squares solver, and provide convergence guarantees using recent Conv
ex Optimization landscape results. Our empirical findings demonstrate that our a
pproach significantly enhances the robustness of state-of-the-art Byzantine resi
lient aggregators. We evaluate our method in a distributed setup with a paramete
r server, and show simultaneous improvements in communication efficiency and acc
uracy across various tasks.
****
YUTONG WU,Han Qiu,Shangwei Guo,Jiwei Li,Tianwei Zhang
You Only Query Once: An Efficient Label-Only Membership Inference Attack
As one of the privacy threats to machine learning models, the membership inferen
ce attack (MIA) tries to infer whether a given sample is in the original trainin
g set of a victim model by analyzing its outputs. Recent studies only use the pr
edicted hard labels to achieve impressive membership inference accuracy. However
, such label-only MIA approach requires very high query budgets to evaluate the
distance of the target sample from the victim model's decision boundary.
   We propose YOQO, a novel label-only attack to overcome the above limitation.Y
OQO aims at identifying a special area (called improvement area) around the targ
et sample and crafting a query sample, whose hard label from the victim model ca
n reliably reflect the target sample's membership. YOQO can successfully reduce
the query budget from more than 1,000 times to only ONCE. Experiments demonstrat
e that YOQO is not only as effective as SOTA attack methods, but also performs c
omparably or even more robustly against many sophisticated defenses.
****
Tom Hosking,Phil Blunsom,Max Bartolo
Human Feedback is not Gold Standard
Human feedback has become the de facto standard for evaluating the performance o
f Large Language Models, and is increasingly being used as a training objective.
 However, it is not clear which properties of a generated output this single `pr
eference' score captures. We hypothesise that preference scores are subjective a
nd open to undesirable biases. We critically analyse the use of human feedback f
or both training and evaluation, to verify whether it fully captures a range of
crucial error criteria. We find that while preference scores have fairly good co
verage, they under-represent important aspects like factuality. We further hypot

hesise that both preference scores and error annotation may be affected by confounders, and leverage instruction-tuned models to generate outputs that vary along two possible confounding dimensions: assertiveness and complexity. We find that the assertiveness of an output skews the perceived rate of factuality errors, indicating that human annotations are not a fully reliable evaluation metric or training objective. Finally, we offer preliminary evidence that using human feedback as a training objective disproportionately increases the assertiveness of model outputs. We encourage future work to carefully consider whether preference scores are well aligned with the desired objective.
****
Thaddäus Wiedemer,Jack Brady,Alexander Panfilov,Attila Juhos,Matthias Bethge,Wieland Brendel
Provable Compositional Generalization for Object-Centric Learning
Learning representations that generalize to novel compositions of known concepts is crucial for bridging the gap between human and machine perception. One prominent effort is learning object-centric representations, which are widely conjectured to enable compositional generalization. Yet, it remains unclear when this conjecture will be true, as a principled theoretical or empirical understanding of compositional generalization is lacking. In this work, we investigate when compositional generalization is guaranteed for object-centric representations through the lens of identifiability theory. We show that autoencoders that satisfy structural assumptions on the decoder and enforce encoder-decoder consistency will learn object-centric representations that provably generalize compositionally. We validate our theoretical result and highlight the practical relevance of our assumptions through experiments on synthetic image data.
****
Jeff Guo,Philippe Schwaller
Beam Enumeration: Probabilistic Explainability For Sample Efficient Self-conditioned Molecular Design
Generative molecular design has moved from proof-of-concept to real-world applicability, as marked by the surge in very recent papers reporting experimental validation. Key challenges in explainability and sample efficiency present opportunities to enhance generative design to directly optimize expensive high-fidelity oracles and provide actionable insights to domain experts. Here, we propose Beam Enumeration to exhaustively enumerate the most probable sub-sequences from language-based molecular generative models and show that molecular substructures can be extracted. When coupled with reinforcement learning, extracted substructures become meaningful, providing a source of explainability and improving sample efficiency through self-conditioned generation. Beam Enumeration is generally applicable to any language-based molecular generative model and notably further improves the performance of the recently reported Augmented Memory algorithm, which achieved the new state-of-the-art on the Practical Molecular Optimization benchmark for sample efficiency. The combined algorithm generates more high reward molecules and faster, given a fixed oracle budget. Beam Enumeration shows that improvements to explainability and sample efficiency for molecular design can be made synergistic.
****
Yapei Chang,Kyle Lo,Tanya Goyal,Mohit Iyyer
BooookScore: A systematic exploration of book-length summarization in the era of LLMs
Summarizing book-length documents ($>$100K tokens) that exceed the context window size of large language models (LLMs) requires first breaking the input document into smaller chunks and then prompting an LLM to merge, update, and compress chunk-level summaries. Despite the complexity and importance of this task, it has yet to be meaningfully studied due to the challenges of evaluation: existing book-length summarization datasets (e.g., BookSum) are in the pretraining data of most public LLMs, and existing evaluation methods struggle to capture errors made by modern LLM summarizers. In this paper, we present the first study of the coherence of LLM-based book-length summarizers implemented via two prompting workflows: (1) hierarchically merging chunk-level summaries, and (2) incrementally u

pdating a running summary. We obtain 1193 fine-grained human annotations on GPT-4 generated summaries of 100 recently-published books and identify eight common types of coherence errors made by LLMs. Because human evaluation is expensive and time-consuming, we develop an automatic metric, BooookScore, that measures the proportion of sentences in a summary that do not contain any of the identified error types. BooookScore has high agreement with human annotations and allows us to systematically evaluate the impact of many other critical parameters (e.g., chunk size, base LLM) while saving \$15K USD and 500 hours in human evaluation costs. We find that closed-source LLMs such as GPT-4 and Claude 2 produce summaries with higher BooookScore than those generated by open-source models. While LLaMA 2 falls behind other models, Mixtral achieves performance on par with GPT-3.5-Turbo. Incremental updating yields lower BooookScore but higher level of detail than hierarchical merging, a trade-off sometimes preferred by annotators. We release code and annotations to spur more principled research on book-length summarization.
****
Chenyu Wang,Sharut Gupta,Caroline Uhler,Tommi S. Jaakkola
Removing Biases from Molecular Representations via Information Maximization
High-throughput drug screening -- using cell imaging or gene expression measurements as readouts of drug effect -- is a critical tool in biotechnology to assess and understand the relationship between the chemical structure and biological activity of a drug. Since large-scale screens have to be divided into multiple experiments, a key difficulty is dealing with batch effects, which can introduce systematic errors and non-biological associations in the data. We propose InfoCORE, an Information maximization approach for COnfounder REmoval, to effectively deal with batch effects and obtain refined molecular representations. InfoCORE establishes a variational lower bound on the conditional mutual information of the latent representations given a batch identifier. It adaptively reweights samples to equalize their implied batch distribution. Extensive experiments on drug screening data reveal InfoCORE's superior performance in a multitude of tasks including molecular property prediction and molecule-phenotype retrieval. Additionally, we show results for how InfoCORE offers a versatile framework and resolves general distribution shifts and issues of data fairness by minimizing correlation with spurious features or removing sensitive attributes.
****
Kejun Tang,Jiayu Zhai,Xiaoliang Wan,Chao Yang
Adversarial Adaptive Sampling: Unify PINN and Optimal Transport for the Approximation of PDEs
Solving partial differential equations (PDEs) is a central task in scientific computing. Recently, neural network approximation of PDEs has received increasing attention due to its flexible meshless discretization and its potential for high-dimensional problems. One fundamental numerical difficulty is that random samples in the training set introduce statistical errors into the discretization of the loss functional which may become the dominant error in the final approximation, and therefore overshadow the modeling capability of the neural network. In this work, we propose a new minmax formulation to optimize simultaneously the approximate solution, given by a neural network model, and the random samples in the training set, provided by a deep generative model. The key idea is to use a deep generative model to adjust the random samples in the training set such that the residual induced by the neural network model can maintain a smooth profile in the training process. Such an idea is achieved by implicitly embedding the Wasserstein distance between the residual-induced distribution and the uniform distribution into the loss, which is then minimized together with the residual. A nearly uniform residual profile means that its variance is small for any normalized weight function such that the Monte Carlo approximation error of the loss functional is reduced significantly for a certain sample size. The adversarial adaptive sampling (AAS) approach proposed in this work is the first attempt to formulate two essential components, minimizing the residual and seeking the optimal training set, into one minmax objective functional for the neural network approximation of PDEs.

****
Yang He,Joey Tianyi Zhou
Data-independent Module-aware Pruning for Hierarchical Vision Transformers
Hierarchical vision transformers (ViTs) have two advantages over conventional ViTs. First, hierarchical ViTs achieve linear computational complexity with respect to image size by local self-attention. Second, hierarchical ViTs create hierarchical feature maps by merging image patches in deeper layers for dense prediction. However, existing pruning methods ignore the unique properties of hierarchical ViTs and use the magnitude value as the weight importance. This approach leads to two main drawbacks. First, the "local" attention weights are compared at a "global" level, which may cause some "locally" important weights to be pruned due to their relatively small magnitude "globally". The second issue with magnitude pruning is that it fails to consider the distinct weight distributions of the network, which are essential for extracting coarse to fine-grained features at various hierarchical levels.

To solve the aforementioned issues, we have developed a Data-independent Module-Aware Pruning method (DIMAP) to compress hierarchical ViTs. To ensure that "local" attention weights at different hierarchical levels are compared fairly in terms of their contribution, we treat them as a **module** and examine their contribution by analyzing their information distortion. Furthermore, we introduce a novel weight metric that is solely based on weights and does not require input images, thereby eliminating the **dependence** on the patch merging process. Our method validates its usefulness and strengths on Swin Transformers of different sizes on ImageNet-1k classification. Notably, the top-5 accuracy drop is only 0.07% when we remove 52.5% FLOPs and 52.7% parameters of Swin-B. When we reduce 33.2% FLOPs and 33.2% parameters of Swin-S, we can even achieve a 0.8% higher relative top-5 accuracy than the original model. Code is available at: https://github.com/he-y/Data-independent-Module-Aware-Pruning.
****
Weiyang Liu,Zeju Qiu,Yao Feng,Yuliang Xiu,Yuxuan Xue,Longhui Yu,Haiwen Feng,Zhen Liu,Juyeon Heo,Songyou Peng,Yandong Wen,Michael J. Black,Adrian Weller,Bernhard Schölkopf
Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization
Large foundation models are becoming ubiquitous, but training them from scratch is prohibitively expensive. Thus, efficiently adapting these powerful models to downstream tasks is increasingly important. In this paper, we study a principled finetuning paradigm -- Orthogonal Finetuning (OFT) -- for downstream task adaptation. Despite demonstrating good generalizability, OFT still uses a fairly large number of trainable parameters due to the high dimensionality of orthogonal matrices. To address this, we start by examining OFT from an information transmission perspective, and then identify a few key desiderata that enable better parameter-efficiency. Inspired by how the Cooley-Tukey fast Fourier transform algorithm enables efficient information transmission, we propose an efficient orthogonal parameterization using butterfly structures. We apply this parameterization to OFT, creating a novel parameter-efficient finetuning method, called Orthogonal Butterfly (BOFT). By subsuming OFT as a special case, BOFT introduces a generalized orthogonal finetuning framework. Finally, we conduct an extensive empirical study of adapting large vision transformers, large language models, and text-to-image diffusion models to various downstream tasks in computer vision and natural language. The results validate the effectiveness of BOFT as a generic finetuning method.
****
Byeonghwi Kim,Minhyuk Seo,Jonghyun Choi
Online Continual Learning for Interactive Instruction Following Agents
In learning an embodied agent executing daily tasks via language directives, the literature largely assumes that the agent learns all training data at the beginning. We argue that such a learning scenario is less realistic, since a robotic agent is supposed to learn the world continuously as it explores and perceives i

t. To take a step towards a more realistic embodied agent learning scenario, we propose two continual learning setups for embodied agents; learning new behaviors (Behavior Incremental Learning, Behavior-IL) and new environments (Environment Incremental Learning, Environment-IL) For the tasks, previous 'data prior' based continual learning methods maintain logits for the past tasks. However, the stored information is often insufficiently learned information and requires task boundary information, which might not always be available. Here, we propose to update them based on confidence scores without task boundary information (i.e., task-free) in a moving average fashion, named Confidence-Aware Moving Average (CAMA). In the proposed challenging Behavior-IL and Environment-IL setups, our simple CAMA outperforms prior arts in our empirical validations by noticeable margins.
****
Yoni Choukroun,Lior Wolf
A Foundation Model for Error Correction Codes
In recent years, Artificial Intelligence has undergone a paradigm shift with the rise of foundation models, which are trained on large amounts of data, typically in a self-supervised way, and can then be adapted to a wide range of downstream tasks. In this work, we propose the first foundation model for Error Correction Codes. This model is trained on multiple codes and can then be applied to an unseen code. To enable this, we extend the Transformer architecture in multiple ways: (1) a code-invariant initial embedding, which is also position- and length-invariant, (2) a learned modulation of the attention maps that is conditioned on the Tanner graph, and (3) a length-invariant code-aware noise prediction module that is based on the parity-check matrix. The proposed architecture is trained on multiple short- and medium-length codes and is able to generalize to unseen codes. Its performance on these codes matches and even outperforms the state of the art, despite having a smaller capacity than the leading code-specific transformers. The suggested framework therefore demonstrates, for the first time, the benefits of learning a universal decoder rather than a neural decoder optimized for a given code.
****
Xiaogeng Liu,Nan Xu,Muhao Chen,Chaowei Xiao
AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models
The aligned Large Language Models (LLMs) are powerful language understanding and decision-making tools that are created through extensive alignment with human feedback. However, these large models remain susceptible to jailbreak attacks, where adversaries manipulate prompts to elicit malicious outputs that should not be given by aligned LLMs. Investigating jailbreak prompts can lead us to delve into the limitations of LLMs and further guide us to secure them. Unfortunately, existing jailbreak techniques suffer from either (1) scalability issues, where attacks heavily rely on manual crafting of prompts, or (2) stealthiness problems, as attacks depend on token-based algorithms to generate prompts that are often semantically meaningless, making them susceptible to detection through basic perplexity testing. In light of these challenges, we intend to answer this question: Can we develop an approach that can automatically generate stealthy jailbreak prompts? In this paper, we introduce AutoDAN, a novel jailbreak attack against aligned LLMs. AutoDAN can automatically generate stealthy jailbreak prompts by the carefully designed hierarchical genetic algorithm. Extensive evaluations demonstrate that AutoDAN not only automates the process while preserving semantic meaningfulness, but also demonstrates superior attack strength in cross-model transferability, and cross-sample universality compared with the baseline. Moreover, we also compare AutoDAN with perplexity-based defense methods and show that AutoDAN can bypass them effectively. Code is available at https://github.com/SheltonLiu-N/AutoDAN.
****
Pum Jun Kim,Seojun Kim,Jaejun Yoo
STREAM: Spatio-TempoRal Evaluation and Analysis Metric for Video Generative Models
Image generative models have made significant progress in generating realistic a

nd diverse images, supported by comprehensive guidance from various evaluation m
etrics. However, current video generative models struggle to generate even
short video clips, with limited tools that provide insights for improvements. Cu
rrent video evaluation metrics are simple adaptations of image metrics by switch
ing the embeddings with video embedding networks, which may underestimate the un
ique characteristics of video. Our analysis reveals that the widely used Frechet
 Video Distance (FVD) has a stronger emphasis on the spatial aspect than the tem
poral naturalness of video and is inherently constrained by the input size of th
e embedding networks used, limiting it to 16 frames. Additionally, it demonstrat
es considerable instability and diverges from human evaluations. To address the
limitations, we propose STREAM, a new video evaluation metric uniquely designed
to independently evaluate spatial and temporal aspects. This feature allows comp
rehensive analysis and evaluation of video generative models from various perspe
ctives, unconstrained by video length. We provide analytical and experimental ev
idence demonstrating that STREAM provides an effective evaluation tool for both
visual and temporal quality of videos, offering insights into area of improvemen
t for video generative models. To the best of our knowledge, STREAM is the first
 evaluation metric that can separately assess the temporal and spatial aspects o
f videos. Our code is available at https://github.com/pro2nit/STREAM.
****
Aliasghar Khani,Saeid Asgari,Aditya Sanghi,Ali Mahdavi Amiri,Ghassan Hamarneh
SLiMe: Segment Like Me
Significant strides have been made using large vision-language models, like Stab
le Diffusion (SD), for a variety of downstream tasks, including image generation
, image editing, and 3D shape generation. Inspired by these advancements, we exp
lore leveraging these vision-language models for segmenting images at any desire
d granularity using as few as one annotated sample. We propose SLiMe, which fram
es this problem as an optimization task. Specifically, given a single image and
its segmentation mask, we first extract our novel “weighted accumulated self-att
ention map” along with cross-attention map from the SD prior. Then, using these
extracted maps, the text embeddings of SD are optimized to highlight the segment
ed region in these attention maps, which in turn can be used to derive new segme
ntation results. Moreover, leveraging additional training data when available, i
.e. few-shot, improves the performance of SLiMe. We performed comprehensive expe
riments examining various design factors and showed that SLiMe outperforms other
 existing one-shot and few-shot segmentation methods.
****
Zhenhui Ye,Tianyun Zhong,Yi Ren,Jiaqi Yang,Weichuang Li,Jiawei Huang,Ziyue Jiang
,Jinzheng He,Rongjie Huang,Jinglin Liu,Chen Zhang,Xiang Yin,Zejun MA,Zhou Zhao
Real3D-Portrait: One-shot Realistic 3D Talking Portrait Synthesis
One-shot 3D talking portrait generation aims to reconstruct a 3D avatar from an
unseen image, and then animate it with a reference video or audio to generate a
talking portrait video. The existing methods fail to simultaneously achieve the
goals of accurate 3D avatar reconstruction and stable talking face animation. Be
sides, while the existing works mainly focus on synthesizing the head part, it i
s also vital to generate natural torso and background segments to obtain a reali
stic talking portrait video. To address these limitations, we present Real3D-Pot
rait, a framework that (1) improves the one-shot 3D reconstruction power with a
large image-to-plane model that distills 3D prior knowledge from a 3D face gener
ative model; (2) facilitates accurate motion-conditioned animation with an effic
ient motion adapter; (3) synthesizes realistic video with natural torso movement
 and switchable background using a head-torso-background super-resolution model;
 and (4) supports one-shot audio-driven talking face generation with a generaliz
able audio-to-motion model. Extensive experiments show that Real3D-Portrait gene
ralizes well to unseen identities and generates more realistic talking portrait
videos compared to previous methods. Video samples are available at https://real
3dportrait.github.io.
****
Simone Magistri,Tomaso Trinci,Albin Soutif,Joost van de Weijer,Andrew D. Bagdano
v

Elastic Feature Consolidation For Cold Start Exemplar-Free Incremental Learning
Exemplar-Free Class Incremental Learning (EFCIL) aims to learn from a sequence of tasks without having access to previous task data. In this paper, we consider the challenging Cold Start scenario in which insufficient data is available in the first task to learn a high-quality backbone. This is especially challenging for EFCIL since it requires high plasticity, which results in feature drift which is difficult to compensate for in the exemplar-free setting. To address this problem, we propose a simple and effective approach that consolidates feature representations by regularizing drift in directions highly relevant to previous tasks and employs prototypes to reduce task-recency bias. Our method, called Elastic Feature Consolidation (EFC), exploits a tractable second-order approximation of feature drift based on an Empirical Feature Matrix (EFM). The EFM induces a pseudo-metric in feature space which we use to regularize feature drift in important directions and to update Gaussian prototypes used in a novel asymmetric cross entropy loss which effectively balances prototype rehearsal with data from new tasks. Experimental results on CIFAR-100, Tiny-ImageNet, ImageNet-Subset and ImageNet-1K demonstrate that Elastic Feature Consolidation is better able to learn new tasks by maintaining model plasticity and significantly outperform the state-of-the-art.
****
Karim Hamade,Reid McIlroy-Young,Siddhartha Sen,Jon Kleinberg,Ashton Anderson
Designing Skill-Compatible AI: Methodologies and Frameworks in Chess
Powerful artificial intelligence systems are often used in settings where they must interact with agents that are computationally much weaker, for example when they work alongside humans or operate in complex environments where some tasks are handled by algorithms, heuristics, or other entities of varying computational power. For AI agents to successfully interact in these settings, however, achieving superhuman performance alone is not sufficient; they also need to account for suboptimal actions or idiosyncratic style from their less-skilled counterparts. We propose a formal evaluation framework for assessing the compatibility of near-optimal AI with interaction partners who may have much lower levels of skill; we use popular collaborative chess variants as model systems to study and develop AI agents that can successfully interact with lower-skill entities. Traditional chess engines designed to output near-optimal moves prove to be inadequate partners when paired with engines of various lower skill levels in this domain, as they are not designed to consider the presence of other agents. We contribute three methodologies to explicitly create skill-compatible AI agents in complex decision-making settings, and two chess game frameworks designed to foster collaboration between powerful AI agents and less-skilled partners. On these frameworks, our agents outperform state-of-the-art chess AI (based on AlphaZero) despite being weaker in conventional chess, demonstrating that skill-compatibility is a tangible trait that is qualitatively and measurably distinct from raw performance. Our evaluations further explore and clarify the mechanisms by which our agents achieve skill-compatibility.
****
Marco Pacini,Xiaowen Dong,Bruno Lepri,Gabriele Santin
A Characterization Theorem for Equivariant Networks with Point-wise Activations
Equivariant neural networks have shown improved performance, expressiveness and sample complexity on symmetrical domains.
But for some specific symmetries, representations, and choice of coordinates, the most common point-wise activations, such as ReLU, are not equivariant, hence they cannot be employed in the design of equivariant neural networks.
The theorem we present in this paper describes all possibile combinations of representations, choice of coordinates and point-wise activations to obtain an equivariant layer, generalizing and strengthening existing characterizations.
Notable cases of practical relevance are discussed as corollaries. Indeed, we prove that rotation-equivariant networks can only be invariant, as it happens for any network which is equivariant with respect to connected compact groups. Then, we discuss implications of our findings when applied to important instances of equivariant networks. First, we completely characterize permutation equivariant

networks such as Invariant Graph Networks with point-wise nonlinearities and the
ir geometric counterparts, highlighting a plethora of models whose expressive po
wer and performance are still unknown.
Second, we show that feature spaces of disentangled steerable convolutional neur
al networks are trivial representations.
****
Mridul Gupta,Sahil Manchanda,HARIPRASAD KODAMANA,Sayan Ranu
Mirage: Model-agnostic Graph Distillation for Graph Classification
GNNs, like other deep learning models, are data and computation hungry. There is
 a pressing need to scale training of GNNs on large datasets to enable their usa
ge on low-resource environments. Graph distillation is an effort in that directi
on with the aim to construct a smaller synthetic training set from the original
training data without significantly compromising model performance. While initia
l efforts are promising, this work is motivated by two key observations: (1) Exi
sting graph distillation algorithms themselves rely on training with the full da
taset, which undermines the very premise of graph distillation. (2) The distilla
tion process is specific to the target GNN architecture and hyper-parameters and
 thus not robust to changes in the modeling pipeline. We circumvent these limita
tions by designing a distillation algorithm called MIRAGE for graph classificati
on. MIRAGE is built on the insight that a message-passing GNN decomposes the inp
ut graph into a multiset of computation trees. Furthermore, the frequency distri
bution of computation trees is often skewed in nature, enabling us to condense t
his data into a concise distilled summary. By compressing the computation data i
tself, as opposed to emulating gradient flows on the original training set—a pre
valent approach to date—MIRAGE transforms into an unsupervised and architecture-
agnostic distillation algorithm. Extensive benchmarking on real-world datasets u
nderscores MIRAGE’s superiority, showcasing enhanced generalization accuracy, da
ta compression, and distillation efficiency when compared to state-of-the-art ba
selines.
****
Hanyu Zhou,Yi Chang,Haoyue Liu,YAN WENDING,Yuxing Duan,Zhiwei Shi,Luxin Yan
Exploring the Common Appearance-Boundary Adaptation for Nighttime Optical Flow
We investigate a challenging task of nighttime optical flow, which suffers from
weakened texture and amplified noise. These degradations weaken discriminative v
isual features, thus causing invalid motion feature matching. Typically, existin
g methods employ domain adaptation to transfer knowledge from auxiliary domain t
o nighttime domain in either input visual space or output motion space. However,
 this direct adaptation is ineffective, since there exists a large domain gap du
e to the intrinsic heterogeneous nature of the feature representations between a
uxiliary and nighttime domains. To overcome this issue, we explore a common-late
nt space as the intermediate bridge to reinforce the feature alignment between a
uxiliary and nighttime domains. In this work, we exploit two auxiliary daytime a
nd event domains, and propose a novel common appearance-boundary adaptation fram
ework for nighttime optical flow. In appearance adaptation, we employ the intrin
sic image decomposition to embed the auxiliary daytime image and the nighttime i
mage into a reflectance-aligned common space. We discover that motion distributi
ons of the two reflectance maps are very similar, benefiting us to consistently
transfer motion appearance knowledge from daytime to nighttime domain. In bounda
ry adaptation, we theoretically derive the motion correlation formula between ni
ghttime image and accumulated events within a spatiotemporal gradient-aligned co
mmon space. We figure out that the correlation of the two spatiotemporal gradien
t maps shares significant discrepancy, benefitting us to contrastively transfer
boundary knowledge from event to nighttime domain. Moreover, appearance adaptati
on and boundary adaptation are complementary to each other, since they could joi
ntly transfer global motion and local boundary knowledge to the nighttime domain
. Extensive experiments have been performed to verify the superiority of the pro
posed method.
****
Ilia Igashov,Arne Schneuing,Marwin Segler,Michael M. Bronstein,Bruno Correia
RetroBridge: Modeling Retrosynthesis with Markov Bridges

Retrosynthesis planning is a fundamental challenge in chemistry which aims at designing multi-step reaction pathways from commercially available starting materials to a target molecule. Each step in multi-step retrosynthesis planning requires accurate prediction of possible precursor molecules given the target molecule and confidence estimates to guide heuristic search algorithms. We model single-step retrosynthesis as a distribution learning problem in a discrete state space. First, we introduce the Markov Bridge Model, a generative framework aimed to approximate the dependency between two intractable discrete distributions accessible via a finite sample of coupled data points. Our framework is based on the concept of a Markov bridge, a Markov process pinned at its endpoints. Unlike diffusion-based methods, our Markov Bridge Model does not need a tractable noise distribution as a sampling proxy and directly operates on the input product molecules as samples from the intractable prior distribution. We then address the retrosynthesis planning problem with our novel framework and introduce RetroBridge, a template-free retrosynthesis modeling approach that achieves state-of-the-art results on standard evaluation benchmarks.
****

LIN Yong,Fan Zhou,Lu Tan,Lintao Ma,Jianmeng Liu,Yansu HE,Yuan Yuan,Yu Liu,James Y. Zhang,Yujiu Yang,Hao Wang
Continuous Invariance Learning
Invariance learning methods aim to learn invariant features in the hope that they generalize under distributional shift. Although many tasks are naturally characterized by continuous domains, current invariance learning techniques generally assume categorically indexed domains. For example, auto-scaling in cloud computing often needs a CPU utilization prediction model that generalizes across different times (e.g., time of a day and date of a year), where `time' is a continuous domain index. In this paper, we start by theoretically showing that existing invariance learning methods can fail for continuous domain problems. Specifically, the naive solution of splitting continuous domains into discrete ones ignores the underlying relationship among domains, and therefore potentially leads to suboptimal performance. To address this challenge, we then propose Continuous Invariance Learning (CIL), which extracts invariant features across continuously indexed domains. CIL is a novel adversarial procedure which measures and controls the conditional independence between the labels and continuous domain indices given the extracted features. Our theoretical analysis demonstrates that CIL learns features that satisfy the invariant constraint with infinite samples. Empirical results on both synthetic and real-world datasets (including data collected from production systems) show that CIL consistently outperforms strong baselines among all the tasks.
****

Haolin Liu,Chen-Yu Wei,Julian Zimmert
Towards Optimal Regret in Adversarial Linear MDPs with Bandit Feedback
We study online reinforcement learning in linear Markov decision processes with adversarial losses and bandit feedback. We introduce two algorithms that achieve improved regret performance compared to existing approaches. The first algorithm, although computationally inefficient, achieves a regret of $\widetilde{O}(\sqrt{K})$ without relying on simulators, where K is the number of episodes. This is the first rate-optimal result in the considered setting. The second algorithm is computationally efficient and achieves a regret of $\widetilde{O}(K^{\frac{3}{4}})$. These results significantly improve over the prior state-of-the-art: a computationally inefficient algorithm by Kong et al. (2023) with $\widetilde{O}(K^{\frac{4}{5}}+1/\lambda_{\min})$ regret, and a computationally efficient algorithm by Sherman et al. (2023b) with $\widetilde{O}(K^{\frac{6}{7}})$ regret.
****

Hao Liu,Carmelo Sferrazza,Pieter Abbeel
Chain of Hindsight aligns Language Models with Feedback
Learning from human preferences is important for language models to match human needs and to align with human and social values.
Prior works have achieved remarkable successes by learning from human feedback to understand and follow instructions. Nonetheless, these methods are either foun

ded on hand-picked model generations that are favored by human annotators, rendering them inefficient in terms of data utilization and challenging to apply in general, or they depend on reinforcement learning, which often suffers from imperfect reward functions and relies on extremely challenging optimizations. In this work, we propose a novel technique, Chain of Hindsight, that is easy to optimize and can learn from any form of feedback, regardless of its polarity. Our idea is inspired by how humans learn from extensive feedback presented in the form of languages. We convert all types of feedback into sequences of sentences, which are then used to fine-tune the model, allowing us to take advantage of the language comprehension capabilities of language models.
We condition the model on a sequence of model generations paired with feedback. By doing so, the model is trained to generate outputs based on feedback, while learning to identify and correct negative attributes or errors. Applying our method to large language models, we observed that Chain of Hindsight significantly surpasses previous methods in aligning language models with human preferences. We report significant improvements on summarization and dialogue benchmarks, with our approach markedly preferred in human evaluations.
****
Ilya Shenbin,Sergey Nikolenko
ImplicitSLIM and How it Improves Embedding-based Collaborative Filtering
We present ImplicitSLIM, a novel unsupervised learning approach for sparse high-dimensional data, with applications to collaborative filtering. Sparse linear methods (SLIM) and their variations show outstanding performance, but they are memory-intensive and hard to scale. ImplicitSLIM improves embedding-based models by extracting embeddings from SLIM-like models in a computationally cheap and memory-efficient way, without explicit learning of heavy SLIM-like models. We show that ImplicitSLIM improves performance and speeds up convergence for both state of the art and classical collaborative filtering methods. The source code for ImplicitSLIM, related models, and applications is available at https://github.com/ilya-shenbin/ImplicitSLIM.
****
Sepanta Zeighami,Cyrus Shahabi
Towards Establishing Guaranteed Error for Learned Database Operations
Machine learning models have demonstrated substantial performance enhancements over non-learned alternatives in various fundamental data management operations, including indexing (locating items in an array), cardinality estimation (estimating the number of matching records in a database), and range-sum estimation (estimating aggregate attribute values for query-matched records). However, real-world systems frequently favor less efficient non-learned methods due to their ability to offer (worst-case) error guarantees — an aspect where learned approaches often fall short. The primary objective of these guarantees is to ensure system reliability, ensuring that the chosen approach consistently delivers the desired level of accuracy across all databases. In this paper, we embark on the first theoretical study of such guarantees for learned methods, presenting the necessary conditions for such guarantees to hold when using machine learning to perform indexing, cardinality estimation and range-sum estimation. Specifically, we present the first known lower bounds on the model size required to achieve the desired accuracy for these three key database operations. Our results bound the required model size for given average and worst-case errors in performing database operations, serving as the first theoretical guidelines governing how model size must change based on data size to be able to guarantee an accuracy level. More broadly, our established guarantees pave the way for the broader adoption and integration of learned models into real-world systems.
****
Xiaogang Jia,Denis Blessing,Xinkai Jiang,Moritz Reuss,Atalay Donat,Rudolf Lioutikov,Gerhard Neumann
Towards Diverse Behaviors: A Benchmark for Imitation Learning with Human Demonstrations
Imitation learning with human data has demonstrated remarkable success in teaching robots in a wide range of skills. However, the inherent diversity in human be

havior leads to the emergence of multi-modal data distributions, thereby presenting a formidable challenge for existing imitation learning algorithms. Quantifying a model's capacity to capture and replicate this diversity effectively is still an open problem. In this work, we introduce simulation benchmark environments and the corresponding *Datasets with Diverse human Demonstrations for Imitation Learning (D3IL)*, designed explicitly to evaluate a model's ability to learn multi-modal behavior. Our environments are designed to involve multiple sub-tasks that need to be solved, consider manipulation of multiple objects which increases the diversity of the behavior and can only be solved by policies that rely on closed loop sensory feedback. Other available datasets are missing at least one of these challenging properties.
To address the challenge of diversity quantification, we introduce tractable metrics that provide valuable insights into a model's ability to acquire and reproduce diverse behaviors. These metrics offer a practical means to assess the robustness and versatility of imitation learning algorithms. Furthermore, we conduct a thorough evaluation of state-of-the-art methods on the proposed task suite. This evaluation serves as a benchmark for assessing their capability to learn diverse behaviors. Our findings shed light on the effectiveness of these methods in tackling the intricate problem of capturing and generalizing multi-modal human behaviors, offering a valuable reference for the design of future imitation learning algorithms.
****
Aiwei Liu,Leyi Pan,Xuming Hu,Shiao Meng,Lijie Wen
A Semantic Invariant Robust Watermark for Large Language Models
Watermark algorithms for large language models (LLMs) have achieved extremely high accuracy in detecting text generated by LLMs. Such algorithms typically involve adding extra watermark logits to the LLM's logits at each generation step. However, prior algorithms face a trade-off between attack robustness and security robustness. This is because the watermark logits for a token are determined by a certain number of preceding tokens; a small number leads to low security robustness, while a large number results in insufficient attack robustness. In this work, we propose a semantic invariant watermarking method for LLMs that provides both attack robustness and security robustness. The watermark logits in our work are determined by the semantics of all preceding tokens. Specifically, we utilize another embedding LLM to generate semantic embeddings for all preceding tokens, and then these semantic embeddings are transformed into the watermark logits through our trained watermark model.
Subsequent analyses and experiments demonstrated the attack robustness of our method in semantically invariant settings: synonym substitution and text paraphrasing settings. Finally, we also show that our watermark possesses adequate security robustness.
****
Murong Yue,Jie Zhao,Min Zhang,Liang Du,Ziyu Yao
Large Language Model Cascades with Mixture of Thought Representations for Cost-Efficient Reasoning
Large language models (LLMs) such as GPT-4 have exhibited remarkable performance in a variety of tasks, but this strong performance often comes with the high expense of using paid API services. In this paper, we are motivated to study building an LLM "cascade" to save the cost of using LLMs, particularly for performing (e.g., mathematical, causal) reasoning tasks. Our cascade pipeline follows the intuition that simpler questions can be addressed by a weaker but more affordable LLM, whereas only the most challenging questions necessitate the stronger and more expensive LLM. To realize this decision-making, we consider the "answer consistency" of the weaker LLM as a signal of the question difficulty and propose several methods for answering sampling and consistency checking, including one leveraging a mixture of two thought representations (i.e., Chain-of-Thought and Program-of-Thought). Through experiments on six reasoning benchmark datasets, with GPT-3.5-turbo and GPT-4 being the weaker and stronger LLMs, respectively, our cascade pipeline demonstrates comparable performance but reduces about 60% of the cost compared with fully using the stronger LLM.

****
Sheng Shen,Le Hou,Yanqi Zhou,Nan Du,Shayne Longpre,Jason Wei,Hyung Won Chung,Bar
ret Zoph,William Fedus,Xinyun Chen,Tu Vu,Yuexin Wu,Wuyang Chen,Albert Webson,Yun
xuan Li,Vincent Y Zhao,Hongkun Yu,Kurt Keutzer,Trevor Darrell,Denny Zhou
Mixture-of-Experts Meets Instruction Tuning: A Winning Combination for Large Lan
guage Models
Sparse Mixture-of-Experts (MoE) is a neural architecture design that adds learna
ble parameters to Large Language Models (LLMs) without increasing computational
complexity (FLOPs). Instruction tuning is a technique for training LLMs to follo
w instructions. We advocate combining these two approaches, as we find that MoE
models benefit more from instruction tuning than dense models. In particular, we
 conduct empirical studies across three experimental setups: (i) Direct finetuni
ng on individual downstream tasks devoid of instruction tuning; (ii) Instruction
 tuning followed by in-context few-shot or zero-shot generalization on downstrea
m tasks; and (iii) Instruction tuning supplemented by further finetuning on indi
vidual downstream tasks. In the first scenario, MoE models overall underperform
dense models of identical computational capacity. This narrative, however, drama
tically changes with the introduction of instruction tuning (in the second and t
hird scenarios), used independently or in conjunction with task-specific finetun
ing. Our most powerful model, FLAN-MoE-32B, surpasses the performance of Flan-Pa
LM-62B on four benchmark tasks, while using only a third of the FLOPs. The advan
cements embodied by FLAN-MoE inspire a reevaluation of the design principles of
large-scale, high-performance language models in the framework of task-agnostic
learning.
****
Dennis Wu,Jerry Yao-Chieh Hu,Weijian Li,Bo-Yu Chen,Han Liu
STanHop: Sparse Tandem Hopfield Model for Memory-Enhanced Time Series Prediction
We present STanHop-Net (Sparse Tandem Hopfield Network) for multivariate time se
ries prediction with memory-enhanced capabilities. At the heart of our approach
is STanHop, a novel Hopfield-based neural network block, which sparsely learns a
nd stores both temporal and cross-series representations in a data-dependent fas
hion. In essence, STanHop sequentially learn temporal representation and cross-s
eries representation using two tandem sparse Hopfield layers. In addition, StanH
op incorporates two additional external memory modules: a Plug-and-Play module a
nd a Tune-and-Play module for train-less and task-aware memory-enhancements, re
spectively. They allow StanHop-Net to fastly respond to certain sudden events. M
ethodologically, we construct the StanHop-Net by stacking STanHop blocks in a hi
erarchical fashion, enabling multi-resolution feature extraction with resolution
-specific sparsity. Theoretically, we introduce a sparse extension of the modern
 Hopfield model and show that it endows a tighter memory retrieval error compare
d to the dense counterpart without sacrificing memory capacity. Empirically, we
validate the efficacy of our framework on both synthetic and real-world settings
.
****
Hanni Cheng,Ya Cong,Weihao Jiang,Shiliang Pu
Learning to solve Class-Constrained Bin Packing Problems via Encoder-Decoder Mod
el
Neural methods have shown significant merit in solving combinatorial optimizatio
n (CO) problems, including the Bin Packing Problem (BPP). However, most existing
 ML-based approaches focus on geometric BPP like 3DBPP, neglecting complex vecto
r BPP. In this study, we introduce a vector BPP variant called Class-Constrained
 Bin Packing Problem (CCBPP), dealing with items of both classes and sizes, and
the objective is to pack the items in the least amount of bins respecting the bi
n capacity and the number of different classes that it can hold. To enhance the
efficiency and practicality of solving CCBPP, we propose a learning-based Encode
r-Decoder Model. The Encoder employs a Graph Convolution Network (GCN) to genera
te a heat-map, representing probabilities of different items packing together. T
he Decoder decodes and fine-tunes the solution through Cluster Decode and Activ
e Search methods, thereby producing high-quality solutions for CCBPP instances.
Extensive experiments demonstrate that our proposed method consistently yields h

igh-quality solutions for various kinds of CCBPP with a very small gap from the optimal. Moreover, our Encoder-Decoder Model also shows promising performance on one practical application of CCBPP, the *Manufacturing Order Consolidation Problem* (OCP).
****
Minsu Kim,Taeyoung Yun,Emmanuel Bengio,Dinghuai Zhang,Yoshua Bengio,Sungsoo Ahn,Jinkyoo Park
Local Search GFlowNets
Generative Flow Networks (GFlowNets) are amortized sampling methods that learn a distribution over discrete objects proportional to their rewards. GFlowNets exhibit a remarkable ability to generate diverse samples, yet occasionally struggle to consistently produce samples with high rewards due to over-exploration on wide sample space.
This paper proposes to train GFlowNets with local search, which focuses on exploiting high-rewarded sample space to resolve this issue. Our main idea is to explore the local neighborhood via backtracking and reconstruction guided by backward and forward policies, respectively. This allows biasing the samples toward high-reward solutions, which is not possible for a typical GFlowNet solution generation scheme, which uses the forward policy to generate the solution from scratch. Extensive experiments demonstrate a remarkable performance improvement in several biochemical tasks. Source code is available: \url{https://github.com/dbsxodud-11/ls_gfn}.
****
Zhaowei Zhu,Jialu Wang,Hao Cheng,Yang Liu
Unmasking and Improving Data Credibility: A Study with Datasets for Training Harmless Language Models
Language models have shown promise in various tasks but can be affected by undesired data during training, fine-tuning, or alignment. For example, if some unsafe conversations are wrongly annotated as safe ones, the model fine-tuned on these samples may be harmful. Therefore, the correctness of annotations, i.e., the credibility of the dataset, is important. This study focuses on the credibility of real-world datasets, including the popular benchmarks Jigsaw Civil Comments, Anthropic Harmless & Red Team, PKU BeaverTails & SafeRLHF, that can be used for training a harmless language model. Given the cost and difficulty of cleaning these datasets by humans, we introduce a systematic framework for evaluating the credibility of datasets, identifying label errors, and evaluating the influence of noisy labels in the curated language data, specifically focusing on unsafe comments and conversation classification. With the framework, we find and fix an average of **6.16\%** label errors in **11** datasets constructed from the above benchmarks. The data credibility and downstream learning performance can be remarkably improved by directly fixing label errors, indicating the significance of cleaning existing real-world datasets. Code is available at https://github.com/Docta-ai/docta.
****
Yuzhen Mao,Martin Ester,Ke Li
IceFormer: Accelerated Inference with Long-Sequence Transformers on CPUs
One limitation of existing transformer-based models is that they cannot handle very long sequences as input since their self-attention operations exhibit quadratic time and space complexity. This problem becomes especially acute when transformers are deployed on hardware platforms equipped only with CPUs. To address this issue, we propose a novel method for accelerating self-attention at inference time that works with pretrained transformer models out-of-the-box without requiring retraining. We experiment using our method to accelerate various long-sequence transformers on various benchmarks and demonstrate a greater speedup compared to the baselines.
****
Yukang Chen,Shengju Qian,Haotian Tang,Xin Lai,Zhijian Liu,Song Han,Jiaya Jia
LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models
We present LongLoRA, an efficient fine-tuning approach that extends the context sizes of pre-trained large language models (LLMs), with limited computation cost

.
Typically, training LLMs with long context sizes is computationally expensive, r
equiring extensive training hours and GPU resources. For example, training on th
e context length of 8192 needs 16x computational costs in self-attention layers
as that of 2048. In this paper, we speed up the context extension of LLMs in two
 aspects. On the one hand, although dense global attention is needed during infe
rence, fine-tuning the model can be effectively and efficiently done by sparse l
ocal attention. The proposed shifted sparse attention effectively enables contex
t extension, leading to non-trivial computation saving with similar performance
to fine-tuning with vanilla attention. Particularly, it can be implemented with
only two lines of code in training, while being optional in inference. On the ot
her hand, we revisit the parameter-efficient fine-tuning regime for context expa
nsion. Notably, we find that LoRA for context extension works well under the pre
mise of trainable embedding and normalization. LongLoRA combines this improved L
oRA with S^2-Attn. LongLoRA demonstrates strong empirical results on various tas
ks on Llama2 models from 7B/13B to 70B. LongLoRA extends Llama2 7B from 4k conte
xt to 100k, or Llama2 70B to 32k on a single 8x A100 machine. LongLoRA extends m
odels' context while retaining their original architectures, and is compatible w
ith most existing techniques, like Flash-Attention2. In addition, we further con
duct supervised fine-tuning with LongLoRA and our long instruction-following Lon
gAlpaca dataset. All our code, models, dataset, and demo are available at https:
//github.com/dvlab-research/LongLoRA.
****
Hang Xu,Kai Li,Haobo Fu,QIANG FU,Junliang Xing,Jian Cheng
Dynamic Discounted Counterfactual Regret Minimization
Counterfactual regret minimization (CFR) is a family of iterative algorithms sho
wing promising results in solving imperfect-information games. Recent novel CFR
variants (e.g., CFR+, DCFR) have significantly improved the convergence rate of
the vanilla CFR. The key to these CFR variants’ performance is weighting each it
eration non-uniformly, i.e., discounting earlier iterations. However, these algo
rithms use a fixed, manually-specified scheme to weight each iteration, which en
ormously limits their potential. In this work, we propose Dynamic Discounted CFR
 (DDCFR), the first equilibrium-finding framework that discounts prior iteration
s using a dynamic, automatically-learned scheme. We formalize CFR’s iteration pr
ocess as a carefully designed Markov decision process and transform the discount
ing scheme learning problem into a policy optimization problem within it. The le
arned discounting scheme dynamically weights each iteration on the fly using inf
ormation available at runtime. Experimental results across multiple games demons
trate that DDCFR’s dynamic discounting scheme has a strong generalization abilit
y and leads to faster convergence with improved performance. The code is availa
ble at https://github.com/rpSebastian/DDCFR.
****
PengFei Zheng,Yonggang Zhang,Zhen Fang,Tongliang Liu,Defu Lian,Bo Han
NoiseDiffusion: Correcting Noise for Image Interpolation with Diffusion Models
 beyond Spherical Linear Interpolation
Image interpolation based on diffusion models is promising in creating fresh and
 interesting images.
Advanced interpolation methods mainly focus on spherical linear interpolation, w
here images are encoded into the noise space and then interpolated for denoising
 to images.
However, existing methods face challenges in effectively interpolating natural i
mages (not generated by diffusion models), thereby restricting their practical a
pplicability.
Our experimental investigations reveal that these challenges stem from the inval
idity of the encoding noise, which may no longer obey the expected noise distrib
ution, e.g., a normal distribution.
To address these challenges, we propose a novel approach to correct noise for im
age interpolation, NoiseDiffusion. Specifically, NoiseDiffusion approaches the i
nvalid noise to the expected distribution by introducing subtle Gaussian noise a
nd introduces a constraint to suppress noise with extreme values. In this contex

t, promoting noise validity contributes to mitigating image artifacts, but the c onstraint and introduced exogenous noise typically lead to a reduction in signal -to-noise ratio, i.e., loss of original image information. Hence, NoiseDiffusion performs interpolation within the noisy image space and injects raw images into these noisy counterparts to address the challenge of information loss. Conseque ntly, NoiseDiffusion enables us to interpolate natural images without causing ar tifacts or information loss, thus achieving the best interpolation results.
****
Yijue Dai,Wenzhong Yan,Feng Yin
Graphical Multioutput Gaussian Process with Attention
Integrating information while recognizing dependence from multiple data sources and enhancing the predictive performance of the multi-output regression are chal lenging tasks. Multioutput Gaussian Process (MOGP) methods offer outstanding sol utions with tractable predictions and uncertainty quantification. However, their practical applications are hindered by high computational complexity and storag e demand. Additionally, there exist model mismatches in existing MOGP models whe n dealing with non-Gaussian data. To improve the model representation ability in terms of flexibility, optimality, and scalability, this paper introduces a nove l multi-output regression framework, termed Graphical MOGP (GMOGP), which is emp owered by: (i) Generating flexible Gaussian process priors consolidated from den tified parents, (ii) providing dependent processes with attention-based graphica l representations, and (iii) achieving Pareto optimal solutions of kernel hyperp arameters via a distributed learning framework. Numerical results confirm that t he proposed GMOGP significantly outperforms state-of-the-art MOGP alternatives i n predictive performance, as well as in time and memory efficiency, across vario us synthetic and real datasets.
****
Jin Su,Chenchen Han,Yuyang Zhou,Junjie Shan,Xibin Zhou,Fajie Yuan
SaProt: Protein Language Modeling with Structure-aware Vocabulary
Large-scale protein language models (PLMs), such as the ESM family, have achieve d remarkable performance in various downstream tasks related to protein structur e and function by undergoing unsupervised training on residue sequences. They ha ve become essential tools for researchers and practitioners in biology. However , a limitation of vanilla PLMs is their lack of explicit consideration for prote in structure information, which suggests the potential for further improvement. Motivated by this, we introduce the concept of a ``structure-aware vocabulary" t hat integrates residue tokens with structure tokens. The structure tokens ar e derived by encoding the 3D structure of proteins using Foldseek. We then pro pose SaProt, a large-scale general-purpose PLM trained on an extensive dataset c omprising approximately 40 million protein sequences and structures. Through ext ensive evaluation, our SaProt model surpasses well-established and renowned base lines across 10 significant downstream tasks, demonstrating its exceptional capa city and broad applicability. We have made the code, pre-trained model, and all relevant materials available at https://github.com/westlake-repl/SaProt.
****
Yazheng Yang,Yuqi Wang,Guang Liu,Ledell Wu,Qi Liu
UniTabE: A Universal Pretraining Protocol for Tabular Foundation Model in Data Science
Recent advancements in Natural Language Processing (NLP) have witnessed the grou ndbreaking impact of pretrained models, yielding impressive outcomes across vari ous tasks. This study seeks to extend the power of pretraining methodologies to facilitating the prediction over tables in data science, a domain traditionally overlooked, yet inherently challenging due to the plethora of table schemas intr insic to different tasks. The primary research questions underpinning this work revolve around the establishment of a universal pretraining protocol for tables with varied structures, the generalizability and transferability of learned know ledge across tasks, the adaptation to diverse downstream applications, and the i ncorporation of incremental columns over time. In response to these challenges, we introduce UniTabE, a straightforward yet effective method designed to process tables in a uniform manner, devoid of constraints imposed by specific table str

uctures. UniTabE's core concept relies on representing each basic table element with a module, termed TabUnit. This is subsequently followed by a Transformer en coder to refine the representation. Moreover, our model is designed to facilitat e pretraining and finetuning through the utilization of free-form prompts. In or der to implement the pretraining phase, we curated an expansive tabular dataset comprising approximately 13 billion samples, meticulously gathered from the Kagg le platform. This research primarily centers on classification and regression ta sks involving tabular data, and conducts rigorous experimental testing and analy ses to validate the effectiveness of our methodology. The experimental results d emonstrate UniTabE's superior performance against several baseline models across a multitude of benchmark datasets. This, therefore, underscores UniTabE's poten tial to significantly enhance the semantic representation of tabular data, there by marking a significant stride for tabular data analysis.
****
Enyi Jiang,Yibo Jacky Zhang,Sanmi Koyejo
Principled Federated Domain Adaptation: Gradient Projection and Auto-Weighting
Federated Domain Adaptation (FDA) describes the federated learning (FL) setting where source clients and a server work collaboratively to improve the performanc e of a target client where limited data is available. The domain shift between t he source and target domains, coupled with limited data of the target client, ma kes FDA a challenging problem, e.g., common techniques such as federated averagi ng and fine-tuning fail due to domain shift and data scarcity.
To theoretically understand the problem, we introduce new metrics that character ize the FDA setting and a theoretical framework with novel theorems for analyzin g the performance of server aggregation rules. Further, we propose a novel light weight aggregation rule, Federated Gradient Projection ($\texttt{FedGP}$), which significantly improves the target performance with domain shift and data scarci ty. Moreover, our theory suggests an $\textit{auto-weighting scheme}$ that finds the optimal combinations of the source and target gradients. This scheme improv es both $\texttt{FedGP}$ and a simpler heuristic aggregation rule. Extensive exp eriments verify the theoretical insights and illustrate the effectiveness of the proposed methods in practice.
****
Shruthi Gowda,Bahram Zonooz,Elahe Arani
Conserve-Update-Revise to Cure Generalization and Robustness Trade-off in Advers arial Training
Adversarial training improves the robustness of neural networks against adversar ial attacks, albeit at the expense of the trade-off between standard and robust generalization. To unveil the underlying factors driving this phenomenon, we exa mine the layer-wise learning capabilities of neural networks during the transiti on from a standard to an adversarial setting. Our empirical findings demonstrate that selectively updating specific layers while preserving others can substanti ally enhance the network's learning capacity. We, therefore, propose CURE, a nov el training framework that leverages a gradient prominence criterion to perform selective conservation, updating, and revision of weights. Importantly, CURE is designed to be dataset- and architecture-agnostic, ensuring its applicability ac ross various scenarios. It effectively tackles both memorization and overfitting issues, thus enhancing the trade-off between robustness and generalization and additionally, this training approach also aids in mitigating "robust overfitting ". Furthermore, our study provides valuable insights into the mechanisms of sele ctive adversarial training and offers a promising avenue for future research.
****
Renrui Zhang,Zhengkai Jiang,Ziyu Guo,Shilin Yan,Junting Pan,Hao Dong,Yu Qiao,Pen g Gao,Hongsheng Li
Personalize Segment Anything Model with One Shot
Driven by large-data pre-training, Segment Anything Model (SAM) has been demonst rated as a powerful promptable framework, revolutionizing the segmentation field . Despite the generality, customizing SAM for specific visual concepts without m an-powered prompting is under-explored, e.g., automatically segmenting your pet dog in numerous images. In this paper, we introduce a training-free Personalizat

ion approach for SAM, termed PerSAM. Given only one-shot data, i.e., a single im
age with a reference mask, we first obtain a positive-negative location prior fo
r the target concept in new images. Then, aided by target visual semantics, we e
mpower SAM for personalized object segmentation via two proposed techniques: tar
get-guided attention and target-semantic prompting. In this way, we can effectiv
ely customize the general-purpose SAM for private use without any training. To f
urther alleviate the ambiguity of segmentation scales, we present an efficient o
ne-shot fine-tuning variant, PerSAM-F. Freezing the entire SAM, we introduce a s
cale-aware fine-tuning to aggregate multi-scale masks, which only tunes 2 parame
ters within 10 seconds for improved performance. To demonstrate our efficacy, we
 construct a new dataset, PerSeg, for the evaluation of personalized object segm
entation, and also test our methods on various one-shot image and video segmenta
tion benchmarks. Besides, we propose to leverage PerSAM to improve DreamBooth fo
r personalized text-to-image synthesis. By mitigating the disturbance of trainin
g-set backgrounds, our approach showcases better target appearance generation an
d higher fidelity to the input text prompt.
****

Tongzhou Mu,Minghua Liu,Hao Su
Learning Reusable Dense Rewards for Multi-Stage Tasks
The success of many RL techniques heavily relies on human-engineered dense rewar
ds, which typically demands substantial domain expertise and extensive trial and
 error. In our work, we propose **DrS** (**D**ense **r**eward learning from **S*
*tages), a novel approach for learning *reusable* dense rewards for multi-stage
tasks in a data-driven manner. By leveraging the stage structures of the task, D
rS learns a high-quality dense reward from sparse rewards and demonstrations if
given. The learned rewards can be *reused* in unseen tasks, thus reducing the hu
man effort for reward engineering. Extensive experiments on three physical robot
 manipulation task families with 1000+ task variants demonstrate that our learne
d rewards can be reused in unseen tasks, resulting in improved performance and s
ample efficiency of RL algorithms. The learned rewards even achieve comparable p
erformance to human-engineered rewards on some tasks. See our [project page](htt
ps://sites.google.com/view/iclr24drs) for videos.
****

Jiseok Chae,Kyuwon Kim,Donghwan Kim
Two-timescale Extragradient for Finding Local Minimax Points
Minimax problems are notoriously challenging to optimize. However, we present th
at two-timescale extragradient can be a viable solution. By utilizing dynamical
systems theory, we show that it converges to points that satisfy the second-orde
r necessary condition of local minimax points, under mild conditions. This work
provably improves upon all previous results as we eliminate a crucial assumption
 that the Hessian, with respect to the maximization variable, is nondegenerate.
****

Anirudh Buvanesh,Rahul Chand,Jatin Prakash,Bhawna Paliwal,Mudit Dhawan,Neelabh M
adan,Deepesh Hada,Vidit Jain,SONU MEHTA,Yashoteja Prabhu,Manish Gupta,Ramachandr
an Ramjee,Manik Varma
Enhancing Tail Performance in Extreme Classifiers by Label Variance Reduction
Extreme Classification (XC) architectures, which utilize a massive one-vs-all cl
assifier layer at the output, have demonstrated remarkable performance on proble
ms with large label sets. Nonetheless, these architectures falter on tail labels
 with few representative samples. This phenomenon has been attributed to factors
 such as classifier over-fitting and missing label bias, and solutions involving
 regularization and loss re-calibration have been developed. This paper explores
 the impact of label variance, a previously unexamined factor, on the tail perfo
rmance in extreme classifiers. It also presents a method to systematically reduc
e label variance in XC by effectively utilizing the capabilities of an additiona
l, tail-robust teacher model. For this purpose, it proposes a principled knowled
ge distillation framework, LEVER, which enhances the tail performance in extreme
 classifiers with formal guarantees on generalization. Comprehensive experiments
 are conducted on a diverse set of XC datasets, demonstrating that LEVER can enh
ance tail performance by around 5% and 6% points in PSP and coverage metrics, re

spectively, when integrated with leading extreme classifiers. Moreover, it estab
lishes a new state-of-the-art when added to the top-performing Ren ■ee classifie
r. Extensive ablations and analyses substantiate the efficacy of our design choi
ces. Another significant contribution is the release of two new XC datasets that
 are more challenging than the existing benchmark datasets and enable a more tho
rough algorithmic evaluation. Code for LEVER is available at: https://aka.ms/lev
er.
****
Jie Xiao,Ruili Feng,Han Zhang,Zhiheng Liu,Zhantao Yang,Yurui Zhu,Xueyang Fu,Kai
Zhu,Yu Liu,Zheng-Jun Zha
DreamClean: Restoring Clean Image Using Deep Diffusion Prior
Image restoration poses a garners substantial interest due to the exponential su
rge in demands for recovering high-quality images from diverse mobile camera dev
ices, adverse lighting conditions, suboptimal shooting environments, and frequen
t image compression for efficient transmission purposes. Yet this problem gather
s significant challenges as people are blind to the type of restoration the imag
es suffer, which, is usually the case in real-day scenarios and is most urgent t
o solve for this field. Current research, however, heavily relies on prior knowl
edge of the restoration type, either explicitly through rules or implicitly thro
ugh the availability of degraded-clean image pairs to define the restoration pro
cess, and consumes considerable effort to collect image pairs of vast degradatio
n types. This paper introduces DreamClean, a training-free method that needs no
degradation prior knowledge but yields high-fidelity and generality towards vari
ous types of image degradation. DreamClean embeds the degraded image back to the
 latent of pre-trained diffusion models and re-sample it through a carefully des
igned diffusion process that mimics those generating clean images. Thanks to the
 rich image prior in diffusion models and our novel Variance Preservation Sampli
ng (VPS) technique, DreamClean manages to handle various different degradation t
ypes at one time and reaches far more satisfied final quality than previous comp
etitors. DreamClean relies on elegant theoretical supports to assure its converg
ence to clean image when VPS has appropriate parameters, and also enjoys superio
r experimental performance over various challenging tasks that could be overwhel
ming for previous methods when degradation prior is unavailable.
****
Marien Renaud,Jiaming Liu,Valentin De Bortoli,Andres Almansa,Ulugbek Kamilov
Plug-and-Play Posterior Sampling under Mismatched Measurement and Prior Models
Posterior sampling has been shown to be a powerful Bayesian approach for solving
 imaging inverse problems. The recent plug-and-play unadjusted Langevin algorith
m (PnP-ULA) has emerged as a promising method for Monte Carlo sampling and minim
um mean squared error (MMSE) estimation by combining physical measurement models
 with deep-learning priors specified using image denoisers. However, the intrica
te relationship between the sampling distribution of PnP-ULA and the mismatched
data-fidelity and denoiser has not been theoretically analyzed. We address this
gap by proposing a posterior-L_2 pseudometric and using it to quantify an expl
icit error bound for PnP-ULA under mismatched posterior distribution. We numeric
ally validate our theory on several inverse problems such as sampling from Gauss
ian mixture models and image deblurring. Our results suggest that the sensitivit
y of the sampling distribution of PnP-ULA to a mismatch in the measurement model
 and the denoiser can be precisely characterized.
****
Jang-Hyun Kim,Junyoung Yeom,Sangdoo Yun,Hyun Oh Song
Compressed Context Memory for Online Language Model Interaction
This paper presents a context key/value compression method for Transformer langu
age models in online scenarios, where the context continually expands. As the co
ntext lengthens, the attention process demands increasing memory and computation
s, which in turn reduces the throughput of the language model. To address this c
hallenge, we propose a compressed context memory system that continually compres
ses the accumulating attention key/value pairs into a compact memory space, faci
litating language model inference in a limited memory space of computing environ
ments. Our compression process involves integrating a lightweight conditional Lo

RA into the language model's forward pass during inference, without the need for fine-tuning the model's entire set of weights. We achieve efficient training by modeling the recursive compression process as a single parallelized forward computation. Through evaluations on conversation, personalization, and multi-task learning, we demonstrate that our approach achieves the performance level of a full context model with $5\times$ smaller context memory size. We further demonstrate the applicability of our approach in a streaming setting with an unlimited context length, outperforming the sliding window approach. Codes are available at https://github.com/snu-mllab/context-memory.
****
Eric J Bigelow,Ekdeep Singh Lubana,Robert P. Dick,Hidenori Tanaka,Tomer Ullman
In-Context Learning Dynamics with Random Binary Sequences
Large language models (LLMs) trained on huge corpora of text datasets demonstrate complex, emergent capabilities, achieving state-of-the-art performance on tasks they were not explicitly trained for. The precise nature of LLM capabilities is often unclear, and different prompts can elicit different capabilities through in-context learning. We propose a Cognitive Interpretability framework that enables us to analyze in-context learning dynamics to understand latent concepts in LLMs underlying behavioral patterns. This provides a more nuanced understanding than success-or-failure evaluation benchmarks, but does not require observing internal activations as a mechanistic interpretation of circuits would require. Inspired by the cognitive science of human randomness perception, we use random binary sequences as context and study dynamics of in-context learning by manipulating properties of context data, such as sequence length. In the latest GPT-3.5+ models, we find emergent abilities to generate pseudo-random numbers and learn basic formal languages, with striking in-context learning dynamics where model outputs transition sharply from pseudo-random behaviors to deterministic repetition.
****
Gen Li,Lu Yin,Jie Ji,Wei Niu,Minghai Qin,Bin Ren,Linke Guo,Shiwei Liu,Xiaolong Ma
NeurRev: Train Better Sparse Neural Network Practically via Neuron Revitalization
Dynamic Sparse Training (DST) employs a greedy search mechanism to identify an optimal sparse subnetwork by periodically pruning and growing network connections during training. To guarantee effectiveness, DST algorithms rely on high search frequency, which consequently, requires large learning rate and batch size to enforce stable neuron learning. Such settings demand extreme memory consumption, as well as generating significant system overheads that limit the wide deployment of deep learning-based applications on resource-constraint platforms. To reconcile such, we propose $\underline{Neur}$on $\underline{Rev}$italization framework for DST (NeurRev), based on an innovative finding that dormant neurons exist with the presence of weight sparsity, and cannot be revitalized (i.e., activated for learning) even with high sparse mask search frequency. These dormant neurons produce a large quantity of zeros during training, which contribute relatively little to the outputs of succeeding layers or to the final results. Different from most existing DST algorithms that spare no effort designing weight growing criteria, NeurRev focuses on optimizing the long-neglected pruning part, which awakes dormant neurons by pruning and incurs no additional computation costs. As such, NeurRev advances more effective neuron learning, which not only achieves outperforming accuracy in a variety of networks and datasets, but also promoting a low-cost dynamism at system-level. Systematical evaluations on training speed and system overhead are conducted on the mobile devices, where the proposed NeurRev framework consistently outperforms representative state-of-the-arts. Code will be released.
****
Jijin Hu,Ke Li,Yonggang Qi,Yi-Zhe Song
Scale-Adaptive Diffusion Model for Complex Sketch Synthesis
While diffusion models have revolutionized generative AI, their application to human sketch generation, especially in the creation of complex yet concise and re

cognizable sketches, remains largely unexplored. Existing efforts have primarily focused on vector-based sketches, limiting their ability to handle intricate sketch data. This paper introduces an innovative extension of diffusion models to pixellevel sketch generation, addressing the challenge of dynamically optimizing the guidance scale for classifier-guided diffusion. Our approach achieves a delicate balance between recognizability and complexity in generated sketches through scale-adaptive classifier-guided diffusion models, a scaling indicator, and the concept of a residual sketch. We also propose a three-phase sampling strategy to enhance sketch diversity and quality. Experiments on the QuickDraw dataset showcase the potential of diffusion models to push the boundaries of sketch generation, particularly in complex scenarios unattainable by vector-based methods.
****
Thomas P Zollo,Todd Morrill,Zhun Deng,Jake Snell,Toniann Pitassi,Richard Zemel
Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models
With the explosion of the zero-shot capabilities of (and thus interest in) pre-trained large language models, there has come accompanying interest in how best to prompt a language model to perform a given task. While it may be tempting to choose a prompt based on empirical results on a validation set, this can lead to a deployment where an unexpectedly high loss occurs. To mitigate this prospect, we propose a lightweight framework, Prompt Risk Control, for selecting a prompt based on rigorous upper bounds on families of informative risk measures. We provide and compare different methods for producing bounds on a diverse set of risk metrics like mean, CVaR, and the Gini coefficient of the loss distribution. In addition, we extend the underlying statistical bounding techniques to accommodate the possibility of distribution shifts in deployment. Extensive experiments on high-impact applications like chatbots, medical question answering, and news summarization highlight why such a framework is necessary to reduce exposure to the worst outcomes.
****
Mikhail Khodak,Edmond Chow,Maria Florina Balcan,Ameet Talwalkar
Learning to Relax: Setting Solver Parameters Across a Sequence of Linear System Instances
Solving a linear system ${\bf Ax}={\bf b}$ is a fundamental scientific computing primitive, and numerous solvers and preconditioners have been developed. These come with parameters whose optimal values depend on the system being solved but are often impossible or too expensive to identify; thus in practice sub-optimal heuristics are used instead. We consider the common setting in which many related linear systems are solved, e.g. during a single numerical simulation. In this scenario, can we sequentially choose parameters that attain a near-optimal overall number of iterations, without extra matrix computations? We answer in the affirmative for Successive Over-Relaxation~(SOR), a standard solver whose parameter ω has a strong impact on its runtime. For this method, we prove that a bandit algorithm—using only the number of iterations as feedback—can select parameters for a sequence of instances such that the overall cost is almost as good as that the best fixed ω would have obtained. Furthermore, when given additional structural information, we show that a {\em contextual} bandit method approaches the performance of the {\em instance-optimal} policy, which selects the best ω for each instance. Our work provides the first learning-theoretic treatment of high-precision linear system solvers and the first end-to-end guarantees for data-driven scientific computing, demonstrating theoretically the potential to speed up numerical methods using well-understood learning algorithms.
****
Chunjin Song,Bastian Wandt,Helge Rhodin
Pose Modulated Avatars from Video
It is now possible to reconstruct dynamic human motion and shape from a sparse set of cameras using Neural Radiance Fields (NeRF) driven by an underlying skeleton. However, a challenge remains to model the deformation of cloth and skin in relation to skeleton pose. Unlike existing avatar models that are learned implicitly or rely on a proxy surface, our approach is motivated by the observation tha

t different poses necessitate unique frequency assignments. Neglecting this dist
inction yields noisy artifacts in smooth areas or blurs fine-grained texture and
 shape details in sharp regions. We develop a two-branch neural network that is
adaptive and explicit in the frequency domain. The first branch is a graph neura
l network that models correlations among body parts locally, taking skeleton pos
e as input. The second branch combines these correlation features to a set of gl
obal frequencies and then modulates the feature encoding. Our experiments demons
trate that our network outperforms state-of-the-art methods in terms of preservi
ng details and generalization capabilities. Our code is available at https://git
hub.com/ChunjinSong/PM-Avatars.
****
Yixuan He,Gesine Reinert,David Wipf,Mihai Cucuringu
Robust Angular Synchronization via Directed Graph Neural Networks
The angular synchronization problem aims to accurately estimate (up to a constan
t additive phase) a set of unknown angles $\theta_1, \dots, \theta_n\in[0, 2\pi)
$ from m noisy measurements of their offsets $\theta_i-\theta_j$ mod $2\pi.$ A
pplications include, for example, sensor network localization, phase retrieval,
and distributed clock synchronization.
An extension of the problem to the heterogeneous setting (dubbed k-synchroniza
tion) is to estimate k groups of angles simultaneously, given noisy observatio
ns (with unknown group assignment) from each group. Existing methods for angular
 synchronization usually perform poorly in high-noise regimes, which are common
in applications. In this paper, we leverage neural networks for the angular sync
hronization problem, and its heterogeneous extension, by proposing GNNSync, a th
eoretically-grounded end-to-end trainable framework using directed graph neural
networks. In addition, new loss functions are devised to encode synchronization
objectives. Experimental results on extensive data sets demonstrate that GNNSync
 attains competitive, and often superior, performance against a comprehensive se
t of baselines for the angular synchronization problem and its extension, valida
ting the robustness of GNNSync even at high noise levels.
****
Chuheng Zhang,Xiangsen Wang,Wei Jiang,Xianliang Yang,Siwei Wang,Lei Song,Jiang B
ian
Whittle Index with Multiple Actions and State Constraint for Inventory Managemen
t
Whittle index is a heuristic tool that leads to good performance for the restles
s bandits problem. In this paper, we extend Whittle index to a new multi-agent r
einforcement learning (MARL) setting with multiple discrete actions and a possib
ly changing constraint on the state space, resulting in WIMS (Whittle Index with
 Multiple actions and State constraint). This setting is common for inventory ma
nagement where each agent chooses a replenishing quantity level for the correspo
nding stock-keeping-unit (SKU) such that the total profit is maximized while the
 total inventory does not exceed a certain limit. Accordingly, we propose a deep
 MARL algorithm based on WIMS for inventory management. Empirically, our algorit
hm is evaluated on real large-scale inventory management problems with up to 230
7 SKUs and outperforms operation-research-based methods and baseline MARL algori
thms.
****
Jacek Karwowski,Oliver Hayman,Xingjian Bai,Klaus Kiendlhofer,Charlie Griffin,Joa
r Max Viktor Skalse
Goodhart's Law in Reinforcement Learning
Implementing a reward function that perfectly captures a complex task in the rea
l world is impractical. As a result, it is often appropriate to think of the rew
ard function as a *proxy* for the true objective rather than as its definition.
We study this phenomenon through the lens of *Goodhart’s law*, which predicts th
at increasing optimisation of an imperfect proxy beyond some critical point decr
eases performance on the true objective. First, we propose a way to *quantify* t
he magnitude of this effect and *show empirically* that optimising an imperfect
proxy reward often leads to the behaviour predicted by Goodhart’s law for a wide
 range of environments and reward functions. We then provide a *geometric explan

ation* for why Goodhart's law occurs in Markov decision processes. We use these theoretical insights to propose an *optimal early stopping method* that provably avoids the aforementioned pitfall and derive theoretical *regret bounds* for this method. Moreover, we derive a training method that maximises worst-case reward, for the setting where there is uncertainty about the true reward function. Finally, we evaluate our early stopping method experimentally. Our results support a foundation for a theoretically-principled study of reinforcement learning under reward misspecification.
****
Yury Nahshan,Joseph Kampeas,Emir Haleva
Linear Log-Normal Attention with Unbiased Concentration
Transformer models have achieved remarkable results in a wide range of applications. However, their scalability is hampered by the quadratic time and memory complexity of the self-attention mechanism concerning the sequence length. This limitation poses a substantial obstacle when dealing with long documents or high-resolution images. In this work, we study the self-attention mechanism by analyzing the distribution of the attention matrix and its concentration ability. Furthermore, we propose instruments to measure these quantities and introduce a novel self-attention mechanism, Linear Log-Normal Attention, designed to emulate the distribution and concentration behavior of the original self-attention. Our experimental results on popular natural language benchmarks reveal that our proposed Linear Log-Normal Attention outperforms other linearized attention alternatives, offering a promising avenue for enhancing the scalability of transformer models.
****
Firas Al-Hafez,Guoping Zhao,Jan Peters,Davide Tateo
Time-Efficient Reinforcement Learning with Stochastic Stateful Policies
Stateful policies play an important role in reinforcement learning, such as handling partially observable environments, enhancing robustness, or imposing an inductive bias directly into the policy structure. The conventional method for training stateful policies is Backpropagation Through Time (BPTT), which comes with significant drawbacks, such as slow training due to sequential gradient propagation and the occurrence of vanishing or exploding gradients. The gradient is often truncated to address these issues, resulting in a biased policy update. We present a novel approach for training stateful policies by decomposing the latter into a stochastic internal state kernel and a stateless policy, jointly optimized by following the stateful policy gradient. We introduce different versions of the stateful policy gradient theorem, enabling us to easily instantiate stateful variants of popular reinforcement learning and imitation learning algorithms. Furthermore, we provide a theoretical analysis of our new gradient estimator and compare it with BPTT. We evaluate our approach on complex continuous control tasks, e.g. humanoid locomotion, and demonstrate that our gradient estimator scales effectively with task complexity while offering a faster and simpler alternative to BPTT.
****
Emanuele Aiello,LILI YU,Yixin Nie,Armen Aghajanyan,Barlas Oguz
Jointly Training Large Autoregressive Multimodal Models
In recent years, advances in the large-scale pretraining of language and text-to-image models have revolutionized the field of machine learning. Yet, integrating these two modalities into a single, robust model capable of generating seamless multimodal outputs remains a significant challenge. To address this gap, we present the Joint Autoregressive Mixture (JAM) framework, a modular approach that systematically fuses existing text and image generation models. We also introduce a specialized, data-efficient instruction-tuning strategy, tailored for mixed-modal generation tasks. Our final instruct-tuned model demonstrates unparalleled performance in generating high-quality multimodal outputs and represents the first model explicitly designed for this purpose.
****
Grigory Khromov,Sidak Pal Singh
Some Intriguing Aspects about Lipschitz Continuity of Neural Networks

Lipschitz continuity is a crucial functional property of any predictive model, t
hat naturally governs its robustness, generalisation, as well as adversarial vul
nerability. Contrary to other works that focus on obtaining tighter bounds and d
eveloping different practical strategies to enforce certain Lipschitz properties
, we aim to thoroughly examine and characterise the Lipschitz behaviour of Neura
l Networks. Thus, we carry out an empirical investigation in a range of differen
t settings (namely, architectures, datasets, label noise, and more) by exhaustin
g the limits of the simplest and the most general lower and upper bounds. As a h
ighlight of this investigation, we showcase a remarkable fidelity of the lower L
ipschitz bound, identify a striking Double Descent trend in both upper and lower
 bounds to the Lipschitz and explain the intriguing effects of label noise on fu
nction smoothness and generalisation.
****
Haozhe Chen,Junfeng Yang,Carl Vondrick,Chengzhi Mao
INViTE: INterpret and Control Vision-Language Models with Text Explanations
Large-scale pre-trained vision foundation models, such as CLIP, have become de f
acto backbones for various vision tasks. However, due to their black-box nature,
 understanding the underlying rules behind these models’ predictions and control
ling model behaviors have remained open challenges. We present INViTE: a framewo
rk for INterpreting Vision Transformer’s latent tokens with Text Explanations. G
iven a latent token, INViTE retains its semantic information to the final layer
using transformer’s local operations and retrieves the closest text for explanat
ion. INViTE enables understanding of model visual reasoning procedure without ne
eding additional model training or data collection. Based on the obtained interp
retations, INViTE allows for model editing that controls model reasoning behavio
rs and improves model robustness against biases and spurious correlations. Our c
ode is available at https://github.com/tonychenxyz/vit-interpret.
****
Rui Yang,Han Zhong,Jiawei Xu,Amy Zhang,Chongjie Zhang,Lei Han,Tong Zhang
Towards Robust Offline Reinforcement Learning under Diverse Data Corruption
Offline reinforcement learning (RL) presents a promising approach for learning r
einforced policies from offline datasets without the need for costly or unsafe i
nteractions with the environment. However, datasets collected by humans in real-
world environments are often noisy and may even be maliciously corrupted, which
can significantly degrade the performance of offline RL. In this work, we first
investigate the performance of current offline RL algorithms under comprehensive
 data corruption, including states, actions, rewards, and dynamics. Our extensiv
e experiments reveal that implicit Q-learning (IQL) demonstrates remarkable resi
lience to data corruption among various offline RL algorithms. Furthermore, we c
onduct both empirical and theoretical analyses to understand IQL's robust perfor
mance, identifying its supervised policy learning scheme as the key factor. Desp
ite its relative robustness, IQL still suffers from heavy-tail targets of Q func
tions under dynamics corruption. To tackle this challenge, we draw inspiration f
rom robust statistics to employ the Huber loss to handle the heavy-tailedness an
d utilize quantile estimators to balance penalization for corrupted data and lea
rning stability. By incorporating these simple yet effective modifications into
IQL, we propose a more robust offline RL approach named Robust IQL (RIQL). Exten
sive experiments demonstrate that RIQL exhibits highly robust performance when s
ubjected to diverse data corruption scenarios.
****
Yuxian Gu,Li Dong,Furu Wei,Minlie Huang
MiniLLM: Knowledge Distillation of Large Language Models
Knowledge Distillation (KD) is a promising technique for reducing the high compu
tational demand of large language models (LLMs). However, previous KD methods ar
e primarily applied to white-box classification models or training small models
to imitate black-box model APIs like ChatGPT. How to effectively distill the kno
wledge of white-box LLMs into small models is still under-explored, which become
s more important with the prosperity of open-source LLMs. In this work, we propo
se a KD approach that distills LLMs into smaller language models. We first repla
ce the forward Kullback-Leibler divergence (KLD) objective in the standard KD ap

proaches with reverse KLD, which is more suitable for KD on generative language models, to prevent the student model from overestimating the low-probability reg
ions of the teacher distribution. Then, we derive an effective optimization appr
oach to learn this objective. The student models are named MiniLLM. Extensive ex
periments in the instruction-following setting show that MiniLLM generates more
precise responses with higher overall quality, lower exposure bias, better calib
ration, and higher long-text generation performance than the baselines. Our meth
od is scalable for different model families
with 120M to 13B parameters. Our code, data, and model checkpoints can be found
in https://github.com/microsoft/LMOps/tree/main/minillm.
****
Rohan Deb,Yikun Ban,Shiliang Zuo,Jingrui He,Arindam Banerjee
Contextual Bandits with Online Neural Regression
Recent works have shown a reduction from contextual bandits to online regression
 under a realizability assumption \citep{foster2020beyond,foster2021efficient}.
In this work, we investigate the use of neural networks for such online regressi
on and associated Neural Contextual Bandits (NeuCBs). Using existing results for
 wide networks, one can readily show a ${\mathcal{O}}(\sqrt{T})$ regret for onl
ine regression with square loss, which via the reduction implies a ${\mathcal{O}
}(\sqrt{K} T^{3/4})$ regret for NeuCBs. Departing from this standard approach, w
e first show a $\mathcal{O}(\log T)$ regret for online regression with almost co
nvex losses that satisfy QG (Quadratic Growth) condition, a generalization of th
e PL (Polyak-\L ojasiewicz) condition, and that have a unique minima. Although n
ot directly applicable to wide networks since they do not have unique minima, we
 show that adding a suitable small random perturbation to the network prediction
s surprisingly makes the loss satisfy QG with unique minima. Based on such a per
turbed prediction, we show a ${\mathcal{O}}(\log T)$ regret for online regressio
n with both squared loss and KL loss, and subsequently convert these respectivel
y to $\tilde{\mathcal{O}}(\sqrt{KT})$ and $\tilde{\mathcal{O}}(\sqrt{KL^*} + K)$
 regret for NeuCB, where L^* is the loss of the best policy. Separately, we al
so show that existing regret bounds for NeuCBs are $\Omega(T)$ or assume i.i.d.
contexts, unlike this work. Finally, our experimental results on various dataset
s demonstrate that our algorithms, especially the one based on KL loss, persiste
ntly outperform existing algorithms.
****
Yuanhao Xiong,Long Zhao,Boqing Gong,Ming-Hsuan Yang,Florian Schroff,Ting Liu,Cho
-Jui Hsieh,Liangzhe Yuan
Structured Video-Language Modeling with Temporal Grouping and Spatial Grounding
Existing video-language pre-training methods primarily focus on instance-level a
lignment between video clips and captions via global contrastive learning but ne
glect rich fine-grained local information in both videos and text, which is of i
mportance to downstream tasks requiring temporal localization and semantic reaso
ning. A powerful model is expected to be capable of capturing region-object corr
espondences and recognizing scene changes in a video clip, reflecting spatial an
d temporal granularity, respectively. To strengthen model's understanding into s
uch fine-grained details, we propose a simple yet effective video-language model
ing framework, S-ViLM, by exploiting the intrinsic structures of these two modal
ities. It includes two novel designs, inter-clip spatial grounding and intra-cli
p temporal grouping, to promote learning region-object alignment and temporal-aw
are features, simultaneously. Comprehensive evaluations demonstrate that S-ViLM
performs favorably against existing approaches in learning more expressive repre
sentations. Specifically, S-ViLM surpasses the state-of-the-art methods substant
ially on four representative downstream tasks, covering text-video retrieval, vi
deo question answering, video action recognition, and temporal action localizati
on.
****
Bhaskar Mukhoty,Hilal AlQuabeh,Giulia De Masi,Huan Xiong,Bin Gu
Certified Adversarial Robustness for Rate Encoded Spiking Neural Networks
The spiking neural networks are inspired by the biological neurons that employ b
inary spikes to propagate information in the neural network. It has garnered con

siderable attention as the next-generation neural network, as the spiking activi ty simplifies the computation burden of the network to a large extent and is kno wn for its low energy deployment enabled by specialized neuromorphic hardware. O ne popular technique to feed a static image to such a network is rate encoding, where each pixel is encoded into random binary spikes, following a Bernoulli dis tribution that uses the pixel intensity as bias. By establishing a novel connect ion between rate-encoding and randomized smoothing, we give the first provable r obustness guarantee for spiking neural networks against adversarial perturbation of inputs bounded under l_1-norm. We introduce novel adversarial training alg orithms for rate-encoded models that significantly improve the state-of-the-art empirical robust accuracy result. Experimental validation of the method is perfo rmed across various static image datasets, including CIFAR-10, CIFAR-100 and Ima geNet-100. The code is available at \url{https://github.com/BhaskarMukhoty/Certi fiedSNN}.
****
Yabo Zhang,Yuxiang Wei,Dongsheng Jiang,XIAOPENG ZHANG,Wangmeng Zuo,Qi Tian
ControlVideo: Training-free Controllable Text-to-video Generation
Text-driven diffusion models have unlocked unprecedented abilities in image gene ration, whereas their video counterpart lags behind due to the excessive trainin g cost.
To avert the training burden, we propose a training-free ControlVideo to produce high-quality videos based on the provided text prompts and motion sequences.
Specifically, ControlVideo adapts a pre-trained text-to-image model (i.e., Contr olNet) for controllable text-to-video generation.
To generate continuous videos without flicker effect, we propose an interleaved-frame smoother to smooth the intermediate frames.
In particular, interleaved-frame smoother splits the whole videos with successiv e three-frame clips, and stabilizes each clip by updating the middle frame with the interpolation among other two frames in latent space.
Furthermore, a fully cross-frame interaction mechanism have been exploited to fu rther enhance the frame consistency, while a hierarchical sampler is employed to produce long videos efficiently.
Extensive experiments demonstrate that our ControlVideo outperforms the state-of -the-arts both quantitatively and qualitatively.
It is worthy noting that, thanks to the efficient designs, ControlVideo could ge nerate both short and long videos within several minutes using one NVIDIA 2080Ti .
Code and videos are available at [this link](https://github.com/YBYBZhang/Contro lVideo).
****
Tianze Luo,Zhanfeng Mo,Sinno Jialin Pan
Learning Adaptive Multiresolution Transforms via Meta-Framelet-based Graph Convo lutional Network
Graph Neural Networks are popular tools in graph representation learning that ca pture the graph structural properties. However, most GNNs employ single-resoluti on graph feature extraction, thereby failing to capture micro-level local patter ns (high resolution) and macro-level graph cluster and community patterns (low r esolution) simultaneously. Many multiresolution methods have been developed to c apture graph patterns at multiple scales, but most of them depend on predefined and handcrafted multiresolution transforms that remain fixed throughout the trai ning process once formulated. Due to variations in graph instances and distribut ions, fixed handcrafted transforms can not effectively tailor multiresolution re presentations to each graph instance. To acquire multiresolution representation suited to different graph instances and distributions, we introduce the Multires olution Meta-Framelet-based Graph Convolutional Network (MM-FGCN), facilitating comprehensive and adaptive multiresolution analysis across diverse graphs. Exten sive experiments demonstrate that our MM-FGCN achieves SOTA performance on vario us graph learning tasks.
****
Hyunjin Seo,Jihun Yun,Eunho Yang

TEDDY: Trimming Edges with Degree-based Discrimination Strategy
Since the pioneering work on the lottery ticket hypothesis for graph neural networks (GNNs) was proposed in Chen et al. (2021), the study on finding graph lottery tickets (GLT) has become one of the pivotal focus in the GNN community, inspiring researchers to discover sparser GLT while achieving comparable performance to original dense networks. In parallel, the graph structure has gained substantial attention as a crucial factor in GNN training dynamics, also elucidated by several recent studies. Despite this, contemporary studies on GLT, in general, have not fully exploited inherent pathways in the graph structure and identified tickets in an iterative manner, which is time-consuming and inefficient. To address these limitations, we introduce **TEDDY**, a one-shot edge sparsification framework that leverages structural information by incorporating *edge-degree* statistics. Following the edge sparsification, we encourage the parameter sparsity during training via simple projected gradient descent on the ℓ_0 ball. Given the target sparsity levels for both the graph structure and the model parameters, our TEDDY facilitates efficient and rapid realization of GLT within a *single* training. Remarkably, our experimental results demonstrate that TEDDY significantly surpasses conventional iterative approaches in generalization, even when conducting one-shot sparsification that solely utilizes graph structures, without taking feature information into account.
****
Yidong Wang,Zhuohao Yu,Wenjin Yao,Zhengran Zeng,Linyi Yang,Cunxiang Wang,Hao Chen,Chaoya Jiang,Rui Xie,Jindong Wang,Xing Xie,Wei Ye,Shikun Zhang,Yue Zhang
PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization
Instruction tuning large language models (LLMs) remains a challenging task, owing to the complexity of hyperparameter selection and the difficulty involved in evaluating the tuned models. To determine the optimal hyperparameters, an automatic, robust, and reliable evaluation benchmark is essential. However, establishing such a benchmark is not a trivial task due to the challenges associated with evaluation accuracy and privacy protection. In response to these challenges, we introduce a judge large language model, named PandaLM, which is trained to distinguish the superior model given several LLMs. PandaLM's focus extends beyond just the objective correctness of responses, which is the main focus of traditional evaluation datasets. It addresses vital subjective factors such as relative conciseness, clarity, adherence to instructions, comprehensiveness, and formality. To ensure the reliability of PandaLM, we collect a diverse human-annotated test dataset, where all contexts are generated by humans and labels are aligned with human preferences. Our findings reveal that PandaLM-7B offers a performance comparable to both GPT-3.5 and GPT-4. Impressively, PandaLM-70B surpasses their performance. PandaLM enables the evaluation of LLM to be fairer but with less cost, evidenced by significant improvements achieved by models tuned through PandaLM compared to their counterparts trained with default Alpaca's hyperparameters. In addition, PandaLM does not depend on API-based evaluations, thus avoiding potential data leakage.
****
Yuchuan Tian,Hanting Chen,Xutao Wang,Zheyuan Bai,QINGHUA ZHANG,Ruifeng Li,Chao Xu,Yunhe Wang
Multiscale Positive-Unlabeled Detection of AI-Generated Texts
Recent releases of Large Language Models (LLMs), e.g. ChatGPT, are astonishing at generating human-like texts, but they may impact the authenticity of texts. Previous works proposed methods to detect these AI-generated texts, including simple ML classifiers, pretrained-model-based zero-shot methods, and finetuned language classification models. However, mainstream detectors always fail on short texts, like SMSes, Tweets, and reviews. In this paper, a Multiscale Positive-Unlabeled (MPU) training framework is proposed to address the difficulty of short-text detection without sacrificing long-texts. Firstly, we acknowledge the human-resemblance property of short machine texts, and rephrase AI text detection as a partial Positive-Unlabeled (PU) problem by regarding these short machine texts as partially "unlabeled". Then in this PU context, we propose the length-sensitive

Multiscale PU Loss, where a recurrent model in abstraction is used to estimate positive priors of scale-variant corpora. Additionally, we introduce a Text Multiscaling module to enrich training corpora. Experiments show that our MPU method augments detection performance on long AI-generated texts, and significantly improves short-text detection of language model detectors. Language Models trained with MPU could outcompete existing detectors on various short-text and long-text detection benchmarks. The codes are available at https://github.com/mindspore-lab/mindone/tree/master/examples/detect_chatgpt and https://github.com/YuchuanTian/AIGC_text_detector.
****
Haozhe Zhao,Zefan Cai,Shuzheng Si,Xiaojian Ma,Kaikai An,Liang Chen,Zixuan Liu,Sheng Wang,Wenjuan Han,Baobao Chang
MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning
Since the resurgence of deep learning, vision-language models (VLMs) enhanced by large language models (LLMs) have grown exponentially in popularity.
However, while LLMs can utilize extensive background knowledge and task information with in-context learning, most VLMs still struggle with understanding complex multi-modal prompts with multiple images, making VLMs less effective in downstream vision-language tasks.
In this paper, we address the limitation above by 1) introducing vision-language Model with **M**ulti-**M**odal **I**n-**C**ontext **L**earning(MMICL), a new approach to allow the VLM to deal with multi-modal inputs efficiently; 2) proposing a novel context scheme to augment the in-context learning ability of the VLM; 3) constructing the Multi-modal In-Context Learning (MIC) dataset, designed to enhance the VLM's ability to understand complex multi-modal prompts.
Our experiments confirm that MMICL achieves new state-of-the-art zero-shot performance on a wide range of general vision-language tasks, especially for complex benchmarks, including MME and MMBench. Our analysis demonstrates that MMICL effectively tackles the challenge of complex multi-modal prompt understanding and emerges the impressive ICL ability. Furthermore, we observe that MMICL successfully alleviates language bias in VLMs, a common issue for VLMs that often leads to hallucination when faced with extensive textual context.
Our code, dataset, dataset tool, and model are available at https://github.com/PKUnlp-icler/MIC.
****
Panagiotis Dimitrakopoulos,Giorgos Sfikas,Christophoros Nikou
Implicit Neural Representation Inference for Low-Dimensional Bayesian Deep Learning
Bayesian inference is the standard for providing full predictive distributions with well calibrated uncertainty estimates.
■However, scaling to a modern, overparameterized deep learning setting
■typically comes at the cost of severe and restrictive approximations, sacrificing model predictive strength.
■With our approach, we factor model parameters as a function of deterministic and probabilistic components;
■the model is solved by combining maximum a posteriori estimation of the former,
■with inference over a low-dimensional, Implicit Neural Representation of the latter.
■This results in a solution that combines both predictive accuracy and good calibration,
■as it entails inducing stochasticity over the full set of model weights while being comparatively cheap to compute.
■Experimentally, our approach compares favorably to the state of the art,
■including much more expensive methods as well as less expressive posterior approximations over full network parameters.
****
Dongwon Son,Jaehyung Kim,Sanghyeon Son,Beomjoon Kim
An Intuitive Multi-Frequency Feature Representation for SO(3)-Equivariant Networks
The usage of 3D vision algorithms, such as shape reconstruction, remains limited

 because they require inputs to be at a fixed canonical rotation. Recently, a si
mple equivariant network, Vector Neuron (VN) has been proposed that can be easil
y used with the state-of-the-art 3D neural network (NN) architectures. However,
its performance is limited because it is designed to use only three-dimensional
features, which is insufficient to capture the details present in 3D data. In th
is paper, we introduce an equivariant feature representation for mapping a 3D po
int to a high-dimensional feature space. Our feature can discern multiple freque
ncies present in 3D data, which, as shown by Tancik et al. (2020), is the key to
 designing an expressive feature for 3D vision tasks. Our representation can be
used as an input to VNs, and the results demonstrate that with our feature repre
sentation, VN captures more details, overcoming the limitation raised in its ori
ginal paper.
****
Biao Zhang,Zhongtao Liu,Colin Cherry,Orhan Firat
When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Meth
od
While large language models (LLMs) often adopt finetuning to unlock their capabi
lities for downstream applications, our understanding on the inductive biases (e
specially the scaling properties) of different finetuning methods is still limit
ed. To fill this gap, we conduct systematic experiments studying whether and how
 different scaling factors, including LLM model size, pretraining data size, new
 finetuning parameter size and finetuning data size, affect the finetuning perfo
rmance. We consider two types of finetuning – full-model tuning (FMT) and parame
ter efficient tuning (PET, including prompt tuning and LoRA), and explore their
scaling behaviors in the data-limited regime where the LLM model size substantia
lly outweighs the finetuning data size. Based on two sets of pretrained bilingua
l LLMs from 1B to 16B and experiments on bilingual machine translation and multi
lingual summarization benchmarks, we find that 1) LLM finetuning follows a power
based multiplicative joint scaling law between finetuning data size and each oth
er scaling factor; 2) LLM finetuning benefits more from LLM model scaling than p
retraining data scaling, and PET parameter scaling is generally ineffective; and
 3) the optimal finetuning method is highly task- and finetuning data-dependent.
 We hope our findings could shed light on understanding, selecting and developin
g LLM finetuning methods.
****
Yannis Kalantidis,Mert Bülent Sar■y■ld■z,Rafael S. Rezende,Philippe Weinzaepfel,
Diane Larlus,Gabriela Csurka
Weatherproofing Retrieval for Localization with Generative AI and Geometric Cons
istency
State-of-the-art visual localization approaches generally rely on a first image
retrieval step whose role is crucial. Yet, retrieval often struggles when facing
 varying conditions, due to e.g. weather or time of day, with dramatic consequen
ces on the visual localization accuracy. In this paper, we improve this retrieva
l step and tailor it to the final localization task. Among the several changes w
e advocate for, we propose to synthesize variants of the training set images, ob
tained from generative text-to-image models, in order to automatically expand th
e training set towards a number of nameable variations that particularly hurt vi
sual localization. After expanding the training set, we propose a training appro
ach that leverages the specificities and the underlying geometry of this mix of
real and synthetic images. We experimentally show that those changes translate i
nto large improvements for the most challenging visual localization datasets.
****
Cassidy Laidlaw,Banghua Zhu,Stuart Russell,Anca Dragan
The Effective Horizon Explains Deep RL Performance in Stochastic Environments
Reinforcement learning (RL) theory has largely focused on proving minimax sample
 complexity bounds. These require strategic exploration algorithms that use rela
tively limited function classes for representing the policy or value function. O
ur goal is to explain why deep RL algorithms often perform well in practice, des
pite using random exploration and much more expressive function classes like neu
ral networks. Our work arrives at an explanation by showing that many stochastic

MDPs can be solved by performing only a few steps of value iteration on the random policy's Q function and then acting greedily. When this is true, we find that it is possible to separate the exploration and learning components of RL, making it much easier to analyze. We introduce a new RL algorithm, SQIRL, that iteratively learns a near-optimal policy by exploring randomly to collect rollouts and then performing a limited number of steps of fitted-Q iteration over those rol- outs. We find that any regression algorithm that satisfies basic in-distribution generalization properties can be used in SQIRL to efficiently solve common MDPs. This can explain why deep RL works with complex function approximators like neural networks, since it is empirically established that neural networks generalize well in-distribution. Furthermore, SQIRL explains why random exploration works well in practice, since we show many environments can be solved by effectively estimating the random policy's Q-function and then applying zero or a few steps of value iteration. We leverage SQIRL to derive instance-dependent sample complexity bounds for RL that are exponential only in an "effective horizon" of lookahead—which is typically much smaller than the full horizon—and on the complexity of the class used for function approximation. Empirically, we also find that SQIRL performance strongly correlates with PPO and DQN performance in a variety of stochastic environments, supporting that our theoretical analysis is predictive of practical performance. Our code and data are available at https://github.com/cassidylaidlaw/effective-horizon.
****

Hongpeng Cao,Yanbing Mao,Lui Sha,Marco Caccamo
Physics-Regulated Deep Reinforcement Learning: Invariant Embeddings
This paper proposes the Phy-DRL: a physics-regulated deep reinforcement learning (DRL) framework for safety-critical autonomous systems. The Phy-DRL has three distinguished invariant-embedding designs: i) residual action policy (i.e., integrating data-driven-DRL action policy and physics-model-based action policy), ii) automatically constructed safety-embedded reward, and iii) physics-model-guided neural network (NN) editing, including link editing and activation editing. Theoretically, the Phy-DRL exhibits 1) a mathematically provable safety guarantee and 2) strict compliance of critic and actor networks with physics knowledge about the action-value function and action policy. Finally, we evaluate the Phy-DRL on a cart-pole system and a quadruped robot. The experiments validate our theoretical results and demonstrate that Phy-DRL features guaranteed safety compared to purely data-driven DRL and solely model-based design while offering remarkably fewer learning parameters and fast training towards safety guarantee.
****

Yossi Gandelsman,Alexei A Efros,Jacob Steinhardt
Interpreting CLIP's Image Representation via Text-Based Decomposition
We investigate the CLIP image encoder by analyzing how individual model components affect the final representation. We decompose the image representation as a sum across individual image patches, model layers, and attention heads, and use CLIP's text representation to interpret the summands. Interpreting the attention heads, we characterize each head's role by automatically finding text representations that span its output space, which reveals property-specific roles for many heads (e.g. location or shape). Next, interpreting the image patches, we uncover an emergent spatial localization within CLIP. Finally, we use this understanding to remove spurious features from CLIP and to create a strong zero-shot image segmenter. Our results indicate that scalable understanding of transformer models is attainable and can be used to repair and improve models.
****

Yongchao Du,Min Wang,Wengang Zhou,Shuping Hui,Houqiang Li
Image2Sentence based Asymmetrical Zero-shot Composed Image Retrieval
The task of composed image retrieval (CIR) aims to retrieve images based on the query image and the text describing the users' intent.
Existing methods have made great progress with the advanced large vision-language (VL) model in CIR task, however, they generally suffer from two main issues: lack of labeled triplets for model training and difficulty of deployment on resource-restricted environments when deploying the large vision-language model. To t

ackle the above problems, we propose Image2Sentence based Asymmetric zero-shot composed image retrieval (ISA), which takes advantage of the VL model and only relies on unlabeled images for composition learning. In the framework, we propose a new adaptive token learner that maps an image to a sentence in the word embedding space of VL model. The sentence adaptively captures discriminative visual information and is further integrated with the text modifier. An asymmetric structure is devised for flexible deployment, in which the lightweight model is adopted for the query side while the large VL model is deployed on the gallery side. The global contrastive distillation and the local alignment regularization are adopted for the alignment between the light model and the VL model for CIR task. Our experiments demonstrate that the proposed ISA could better cope with the real retrieval scenarios and further improve retrieval accuracy and efficiency.
****
Hu Xu,Saining Xie,Xiaoqing Tan,Po-Yao Huang,Russell Howes,Vasu Sharma,Shang-Wen Li,Gargi Ghosh,Luke Zettlemoyer,Christoph Feichtenhofer
Demystifying CLIP Data
Contrastive Language-Image Pre-training (CLIP) is an approach that has advanced research and applications in computer vision, fueling modern recognition systems and generative models. We believe that the main ingredient to the success of CLIP is its \textit{data} and \textit{not} the \textit{model} architecture or pre-training {objective}. However, CLIP only provides very limited information about its data and how it has been collected, leading to works that aim to reproduce CLIP's data by filtering with its model parameters. In this work, we intend to reveal CLIP's data curation approach and in our pursuit of making it open to the community introduce Metadata-Curated Language-Image Pre-training (MetaCLIP). MetaCLIP takes a raw data pool and metadata (derived from CLIP's concepts) and yields a balanced subset over the metadata distribution. Our experimental study rigorously isolates the model and training settings, concentrating solely on data. MetaCLIP applied to CommonCrawl with 400M image-text data pairs outperforms CLIP's data on multiple standard benchmarks. In zero-shot ImageNet classification, MetaCLIP achieves 70.8\% accuracy, surpassing CLIP's 68.3\% on \mbox{ViT-B} models. Scaling to 1B data, while maintaining the same training budget, attains \textbf{72.4\%}. Our observations hold across various model sizes, exemplified by ViT-H achieving \textbf{80.5\%}, without any bells-and-whistles. Curation code and training data distribution over metadata will be made available.
****
Lorenzo Pacchiardi,Alex James Chan,Sören Mindermann,Ilan Moscovitz,Alexa Yue Pan,Yarin Gal,Owain Evans,Jan M. Brauner
How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions
Large language models (LLMs) can “lie”, which we define as outputting false statements when incentivised to, despite “knowing” the truth in a demonstrable sense. LLMs might “lie”, for example, when instructed to output misinformation. Here, we develop a simple lie detector that requires neither access to the LLM’s activations (black-box) nor ground-truth knowledge of the fact in question. The detector works by asking a predefined set of unrelated follow-up questions after a suspected lie, and feeding the LLM’s yes/no answers into a logistic regression classifier. Despite its simplicity, this lie detector is highly accurate and surprisingly general. When trained on examples from a single setting—prompting GPT-3.5 to lie about factual questions—the detector generalises out-of-distribution to (1) other LLM architectures, (2) LLMs fine-tuned to lie, (3) sycophantic lies, and (4) lies emerging in real-life scenarios such as sales. These results indicate that LLMs have distinctive lie-related behavioural patterns, consistent across architectures and contexts, which could enable general-purpose lie detection.
****
Sagar Shrestha,Xiao Fu
Towards Identifiable Unsupervised Domain Translation: A Diversified Distribution Matching Approach
Unsupervised domain translation (UDT) aims to find functions that convert samples from one domain (e.g., sketches) to another domain (e.g., photos) without chan

ging the high-level semantic meaning (also referred to as "content"). The transl
ation functions are often sought by probability distribution matching of the tra
nsformed source domain and target domain. CycleGAN stands as arguably the most r
epresentative approach among this line of work. However, it was noticed in the l
iterature that CycleGAN and variants could fail to identify the desired translat
ion functions and produce content-misaligned translations.
This limitation arises due to the presence of multiple translation functions---r
eferred to as ``measure-preserving automorphism" (MPA)---in the solution space o
f the learning criteria. Despite awareness of such identifiability issues, solut
ions have remained elusive. This study delves into the core identifiability inqu
iry and introduces an MPA elimination theory. Our analysis shows that MPA is unl
ikely to exist, if multiple pairs of diverse cross-domain conditional distributi
ons are matched by the learning function.
Our theory leads to a UDT learner using distribution matching over auxiliary var
iable-induced subsets of the domains---other than over the entire data domains a
s in the classical approaches. The proposed framework is the first to rigorousl
y establish translation identifiability under reasonable UDT settings, to our be
st knowledge.
Experiments corroborate with our theoretical claims.
****
Haoxuan Li,Chunyuan Zheng,Sihao Ding,Peng Wu,Zhi Geng,Fuli Feng,Xiangnan He
Be Aware of the Neighborhood Effect: Modeling Selection Bias under Interference for Recommendation
The interaction between users and recommender systems is not only affected by se
lection bias but also the neighborhood effect, i.e., the interaction between a u
ser and an item is affected by the interactions between other users and other it
ems, or between the same user and other items, or between other users and the sa
me item. Many previous studies have focused on addressing selection bias to achi
eve unbiased learning of the prediction model, but the lack of consideration of
neighborhood effects can lead to biased estimates and suboptimal performance of
the prediction model. In this paper, we formally formulate the neighborhood effe
ct as an interference problem from the perspective of causal inference and intro
duce a treatment representation to capture the neighborhood effect. On this basi
s, we propose a novel ideal loss that can be used to deal with selection bias in
 the presence of neighborhood effects. In addition, we further develop two novel
 estimators for the ideal loss. We theoretically establish the connection betwee
n the proposed methods and previous methods ignoring the neighborhood effect and
 show that the proposed methods achieve unbiased learning when both selection bi
as and neighborhood effects are present, while the existing methods are biased.
Extensive semi-synthetic and real-world experiments are conducted to demonstrate
 the effectiveness of the proposed methods.
****
Xavier Puig,Eric Undersander,Andrew Szot,Mikael Dallaire Cote,Tsung-Yen Yang,Rus
lan Partsey,Ruta Desai,Alexander Clegg,Michal Hlavac,So Yeon Min,Vladimír Vondru
š,Theophile Gervet,Vincent-Pierre Berges,John M Turner,Oleksandr Maksymets,Zsolt
 Kira,Mrinal Kalakrishnan,Jitendra Malik,Devendra Singh Chaplot,Unnat Jain,Dhruv
 Batra,Akshara Rai,Roozbeh Mottaghi
Habitat 3.0: A Co-Habitat for Humans, Avatars, and Robots
We present Habitat 3.0: a simulation platform for studying collaborative human-r
obot tasks in home environments. Habitat 3.0 offers contributions across three d
imensions: (1) Accurate humanoid simulation: addressing challenges in modeling c
omplex deformable bodies and diversity in appearance and motion, all while ensur
ing high simulation speed. (2) Human-in-the-loop infrastructure: enabling real h
uman interaction with simulated robots via mouse/keyboard or a VR interface, fac
ilitating evaluation of robot policies with human input. (3) Collaborative tasks
: studying two collaborative tasks, Social Navigation and Social Rearrangement.
Social Navigation investigates a robot's ability to locate and follow humanoid a
vatars in unseen environments, whereas Social Rearrangement addresses collaborat
ion between a humanoid and robot while rearranging a scene. These contributions
allow us to study end-to-end learned and heuristic baselines for human-robot col

laboration in-depth, as well as evaluate them with humans in the loop. Our exper iments demonstrate that learned robot policies lead to efficient task completion when collaborating with unseen humanoid agents and human partners that might ex hibit behaviors that the robot has not seen before. Additionally, we observe eme rgent behaviors during collaborative task execution, such as the robot yielding space when obstructing a humanoid agent, thereby allowing the effective completi on of the task by the humanoid agent. Furthermore, our experiments using the hum an-in-the-loop tool demonstrate that our automated evaluation with humanoids can provide an indication of the relative ordering of different policies when evalu ated with real human collaborators. Habitat 3.0 unlocks interesting new features in simulators for Embodied AI, and we hope it paves the way for a new frontier of embodied human-AI interaction capabilities. For more details and visualizatio ns, visit: https://aihabitat.org/habitat3.
****
Hien Dang,Tho Tran Huu,Tan Minh Nguyen,Nhat Ho
Beyond Vanilla Variational Autoencoders: Detecting Posterior Collapse in Conditi onal and Hierarchical Variational Autoencoders
The posterior collapse phenomenon in variational autoencoder (VAE), where the va riational posterior distribution closely matches the prior distribution, can hin der the quality of the learned latent variables. As a consequence of posterior c ollapse, the latent variables extracted by the encoder in VAE preserve less info rmation from the input data and thus fail to produce meaningful representations as input to the reconstruction process in the decoder. While this phenomenon has been an actively addressed topic related to VAE performance, the theory for pos terior collapse remains underdeveloped, especially beyond the standard VAE. In t his work, we advance the theoretical understanding of posterior collapse to two important and prevalent yet less studied classes of VAE: conditional VAE and hie rarchical VAE. Specifically, via a non-trivial theoretical analysis of linear co nditional VAE and hierarchical VAE with two levels of latent, we prove that the cause of posterior collapses in these models includes the correlation between th e input and output of the conditional VAE and the effect of learnable encoder va riance in the hierarchical VAE. We empirically validate our theoretical findings for linear conditional and hierarchical VAE and demonstrate that these results are also predictive for non-linear cases with extensive experiments.
****
Steeven JANNY,Madiha Nadri,Julie Digne,Christian Wolf
Space and time continuous physics simulation from partial observations
Modern techniques for physical simulations rely on numerical schemes and mesh-re finement methods to address trade-offs between precision and complexity, but the se handcrafted solutions are tedious and require high computational power. Data-driven methods based on large-scale machine learning promise high adaptivity by integrating long-range dependencies more directly and efficiently. In this work, we focus on computational fluid dynamics and address the shortcomings of a larg e part of the literature, which are based on fixed support for computations and predictions in the form of regular or irregular grids. We propose a novel setup to perform predictions in a continuous spatial and temporal domain while being t rained on sparse observations. We formulate the task as a double observation pro blem and propose a solution with two interlinked dynamical systems defined on, r espectively, the sparse positions and the continuous domain, which allows to for ecast and interpolate a solution from the initial condition. Our practical imple mentation involves recurrent GNNs and a spatio-temporal attention observer capab le of interpolating the solution at arbitrary locations. Our model not only gene ralizes to new initial conditions (as standard auto-regressive models do) but al so performs evaluation at arbitrary space and time locations. We evaluate on thr ee standard datasets in fluid dynamics and compare to strong baselines, which ar e outperformed in classical settings and the extended new task requiring continu ous predictions.
****
Wenwen Si,Sangdon Park,Insup Lee,Edgar Dobriban,Osbert Bastani
PAC Prediction Sets Under Label Shift

Prediction sets capture uncertainty by predicting sets of labels rather than individual labels, enabling downstream decisions to conservatively account for all plausible outcomes. Conformal inference algorithms construct prediction sets guaranteed to contain the true label with high probability. These guarantees fail to hold in the face of distribution shift, which is precisely when reliable uncertainty quantification can be most useful. We propose a novel algorithm for constructing prediction sets with PAC guarantees in the label shift setting, where the probabilities of labels can differ between the source and target distributions. Our algorithm relies on constructing confidence intervals for importance weights by propagating uncertainty through a Gaussian elimination algorithm. We evaluate our approach on four datasets: the CIFAR-10 and ChestX-Ray image datasets, the tabular CDC Heart Dataset, and the AGNews text dataset. Our algorithm satisfies the PAC guarantee while producing smaller prediction set sizes compared to several baselines.
****
William F Whitney,Tatiana Lopez-Guevara,Tobias Pfaff,Yulia Rubanova,Thomas Kipf,Kim Stachenfeld,Kelsey R Allen
Learning 3D Particle-based Simulators from RGB-D Videos
Realistic simulation is critical for applications ranging from robotics to animation. Traditional analytic simulators sometimes struggle to capture sufficiently realistic simulation which can lead to problems including the well known "sim-to-real" gap in robotics. Learned simulators have emerged as an alternative for better capturing real-world physical dynamics, but require access to privileged ground truth physics information such as precise object geometry or particle tracks. Here we propose a method for learning simulators directly from observations. Visual Particle Dynamics (VPD) jointly learns a latent particle-based representation of 3D scenes, a neural simulator of the latent particle dynamics, and a renderer that can produce images of the scene from arbitrary views. VPD learns end to end from posed RGB-D videos and does not require access to privileged information. Unlike existing 2D video prediction models, we show that VPD's 3D structure enables scene editing and long-term predictions. These results pave the way for downstream applications ranging from video editing to robotic planning.
****
Ilyass Hammouamri,Ismail Khalfaoui-Hassani,Timothée Masquelier
Learning Delays in Spiking Neural Networks using Dilated Convolutions with Learnable Spacings
Spiking Neural Networks (SNNs) are a promising research direction for building power-efficient information processing systems, especially for temporal tasks such as speech recognition. In SNNs, delays refer to the time needed for one spike to travel from one neuron to another. These delays matter because they influence the spike arrival times, and it is well-known that spiking neurons respond more strongly to coincident input spikes. More formally, it has been shown theoretically that plastic delays greatly increase the expressivity in SNNs. Yet, efficient algorithms to learn these delays have been lacking. Here, we propose a new discrete-time algorithm that addresses this issue in deep feedforward SNNs using backpropagation, in an offline manner. To simulate delays between consecutive layers, we use 1D convolutions across time. The kernels contain only a few non-zero weights – one per synapse – whose positions correspond to the delays. These positions are learned together with the weights using the recently proposed Dilated Convolution with Learnable Spacings (DCLS). We evaluated our method on three datasets: the Spiking Heidelberg Dataset (SHD), the Spiking Speech Commands (SSC) and its non spiking version Google Speech Commands v0.02 (GSC) benchmarks, which require detecting temporal patterns. We used feedforward SNNs with two or three hidden fully connected layers, and vanilla leaky integrate-and-fire neurons. We showed that fixed random delays help and that learning them helps even more. Furthermore, our method outperformed the state-of-the-art in the three datasets without using recurrent connections and with substantially fewer parameters. Our work demonstrates the potential of delay learning in developing accurate and precise models for temporal data processing. Our code is based on PyTorch / SpikingJelly and available at: https://github.com/Thvnvtos/SNN-delays

****
Nan Yin,Mengzhu Wang,Zhenghan Chen,Li Shen,Huan Xiong,Bin Gu,Xiao Luo
DREAM: Dual Structured Exploration with Mixup for Open-set Graph Domain Adaption
Recently, numerous graph neural network methods have been developed to tackle domain shifts in graph data. However, these methods presuppose that unlabeled target graphs belong to categories previously seen in the source domain. This assumption could not hold true for in-the-wild target graphs. In this paper, we delve deeper to explore a more realistic problem open-set graph domain adaptation. Our objective is to not only identify target graphs from new categories but also accurately classify remaining target graphs into their respective categories under domain shift and label scarcity. To address this challenging problem, we introduce a novel method named Dual Structured Exploration with Mixup (DREAM). DREAM incorporates a graph-level representation learning branch as well as a subgraph-enhanced branch, which jointly explores graph topological structures from both global and local viewpoints. To maximize the use of unlabeled target graphs, we train these two branches simultaneously using posterior regularization to enhance their inter-module consistency. To accommodate the open-set setting, we amalgamate dissimilar samples to generate virtual unknown samples belonging to novel classes. Moreover, to alleviate domain shift, we establish a k nearest neighbor-based graph-of-graphs and blend multiple neighbors of each sample to produce cross-domain virtual samples for inter-domain consistency learning. Extensive experiments validate the effectiveness of our proposed DREAM compared with various state-of-the-art approaches in different settings.
****
Qiwei Di,Heyang Zhao,Jiafan He,Quanquan Gu
Pessimistic Nonlinear Least-Squares Value Iteration for Offline Reinforcement Learning
Offline reinforcement learning (RL), where the agent aims to learn the optimal policy based on the data collected by a behavior policy, has attracted increasing attention in recent years. While offline RL with linear function approximation has been extensively studied with optimal results achieved under certain assumptions, many works shift their interest to offline RL with non-linear function approximation.
However, limited works on offline RL with non-linear function approximation have instance-dependent regret guarantees.
In this paper, we propose an oracle-efficient algorithm, dubbed Pessimistic Nonlinear Least-Square Value Iteration (PNLSVI), for offline RL with non-linear function approximation. Our algorithmic design comprises three innovative components: (1) a variance-based weighted regression scheme that can be applied to a wide range of function classes, (2) a subroutine for variance estimation, and (3) a planning phase that utilizes a pessimistic value iteration approach. Our algorithm enjoys a regret bound that has a tight dependency on the function class complexity and achieves minimax optimal instance-dependent regret when specialized to linear function approximation. Our work extends the previous instance-dependent results within simpler function classes, such as linear and differentiable function to a more general framework. To the best of our knowledge, this is the first statistically optimal algorithm for nonlinear offline RL.
****
Arijit Sehanobish,Krzysztof Marcin Choromanski,YUNFAN ZHAO,Kumar Avinava Dubey,Valerii Likhosherstov
Scalable Neural Network Kernels
We introduce the concept of scalable neural network kernels (SNNKs), the replacements of regular feedforward layers (FFLs), capable of approximating the latter, but with favorable computational properties. SNNKs effectively disentangle the inputs from the parameters of the neural network in the FFL, only to connect them in the final computation via the dot-product kernel.
They are also strictly more expressive, as allowing to model complicated relationships beyond the functions of the dot-products of parameter-input vectors. We also introduce the neural network bundling process that applies SNNKs to compactify deep neural network architectures, resulting in additional compression gains.

In its extreme version, it leads to the fully bundled network whose optimal parameters can be expressed via explicit formulae for several loss functions (e.g. mean squared error), opening a possibility to bypass backpropagation. As a by-product of our analysis, we introduce the mechanism of the universal random features (or URFs), applied to instantiate several SNNK variants, and interesting on its own in the context of scalable kernel methods. We provide rigorous theoretical analysis of all these concepts as well as an extensive empirical evaluation, ranging from point-wise kernel estimation to Transformers' fine-tuning with novel adapter layers inspired by SNNKs. Our mechanism provides up to 5x reduction in the number of trainable parameters, while maintaining competitive accuracy.
****
Xinyu Yuan,Yan Qiao
Diffusion-TS: Interpretable Diffusion for General Time Series Generation
Denoising diffusion probabilistic models (DDPMs) are becoming the leading paradigm for generative models. It has recently shown breakthroughs in audio synthesis, time series imputation and forecasting. In this paper, we propose Diffusion-TS, a novel diffusion-based framework that generates multivariate time series samples of high quality by using an encoder-decoder transformer with disentangled temporal representations, in which the decomposition technique guides Diffusion-TS to capture the semantic meaning of time series while transformers mine detailed sequential information from the noisy model input. Different from existing diffusion-based approaches, we train the model to directly reconstruct the sample instead of the noise in each diffusion step, combining a Fourier-based loss term. Diffusion-TS is expected to generate time series satisfying both interpretablity and realness. In addition, it is shown that the proposed Diffusion-TS can be easily extended to conditional generation tasks, such as forecasting and imputation, without any model changes. This also motivates us to further explore the performance of Diffusion-TS under irregular settings. Finally, through qualitative and quantitative experiments, results show that Diffusion-TS achieves the state-of-the-art results on various realistic analyses of time series.
****
Michael Zhang,Kush Bhatia,Hermann Kumbong,Christopher Re
The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry
Linear attentions have shown promise for improving Transformer efficiency, reducing attention's quadratic complexity to linear in sequence length. This holds exciting promise for (1) training linear Transformers from scratch, (2) `inetuned-conversion of task-specific Transformers into linear versions that recover task performance, and (3) pretrained-conversion of Transformers, such as language models, into linear versions readily finetunable on downstream tasks. However, linear attentions often underperform compared to standard softmax attention. To close this performance gap, we study the behaviors of softmax and linear attentions in various train-from-scratch and finetuned-conversion settings. We find prior linear attentions lack key properties of softmax attention tied to good performance: low-entropy (or spiky) weights and dot-product monotonicity. We further observe surprisingly simple feature maps that retain these properties match softmax performance, but are inefficient to compute in linear attention. We thus propose Hedgehog, a learnable linear attention that retains the spiky and monotonic properties of softmax attention while maintaining linear complexity. Hedgehog uses simple, trainable MLPs to produce attention weights mimicking softmax attention. Experiments show Hedgehog recovers over 99\% of standard Transformer performance in train-from-scratch and finetuned-conversion settings, outperforming prior linear attentions by up to 6 perplexity points on WikiText-103 when training causal GPT models from scratch, and up to 8.7 GLUE score points when converting finetuned bidirectional BERT models. Hedgehog also enables pretrained-conversion. Converting a pretrained GPT-2 into a linear attention variant achieves state-of-the-art 16.7 perplexity on WikiText-103 for 125M subquadratic decoder models. We finally turn a pretrained Llama-2 7B into a viable linear attention Llama. With low-rank adaptation, Hedgehog-Llama-2 7B achieves 28.1 higher ROUGE-1 points over the base standard attention model, where prior linear attentions lead to 16.5 point drops.

****
Charlotte Nicks,Eric Mitchell,Rafael Rafailov,Archit Sharma,Christopher D Manning,Chelsea Finn,Stefano Ermon
Language Model Detectors Are Easily Optimized Against
The fluency and general applicability of large language models (LLMs) has motivated significant interest in detecting whether a piece of text was written by a language model. While both academic and commercial detectors have been deployed in some settings, particularly education, other research has highlighted the fragility of these systems. In this paper, we demonstrate a data-efficient attack that fine-tunes language models to confuse existing detectors, leveraging recent developments in reinforcement learning of language models. We use the 'human-ness' score (often just a log probability) of various open-source and commercial detectors as a reward function for reinforcement learning, subject to a KL-divergence constraint that the resulting model does not differ significantly from the original. For a 7B parameter Llama-2 model, fine-tuning for under a day reduces the AUROC of the OpenAI RoBERTa-Large detector from 0.84 to 0.62, while perplexity on OpenWebText increases from 8.7 to only 9.0; with a larger perplexity budget, we reduce AUROC to 0.30 (worse than random), with a perplexity increase to 9.9. Similar to traditional adversarial attacks, we find that this increase in 'detector evasion' generalizes to other detectors not used during training. In light of our empirical results, we advise against continued reliance on LLM-generated text detectors.
****
Yiting Chen,Zhanpeng Zhou,Junchi Yan
Going Beyond Neural Network Feature Similarity: The Network Feature Complexity and Its Interpretation Using Category Theory
The behavior of neural networks still remains opaque, and a recently widely noted phenomenon is that networks often achieve similar performance when initialized with different random parameters. This phenomenon has attracted significant attention in measuring the similarity between features learned by distinct networks. However, feature similarity could be vague in describing the same feature since equivalent features hardly exist. In this paper, we expand the concept of equivalent feature and provide the definition of what we call *functionally equivalent features*. These features produce equivalent output under certain transformations.
Using this definition, we aim to derive a more intrinsic metric for the so-called *feature complexity* regarding the redundancy of features learned by a neural network at each layer. We offer a formal interpretation of our approach through the lens of category theory, a well-developed area in mathematics. To quantify the feature complexity, we further propose an efficient algorithm named Iterative Feature Merging. Our experimental results validate our ideas and theories from various perspectives. We empirically demonstrate that the functionally equivalence widely exists among different features learned by the same neural network and we could reduce the number of parameters of the network without affecting the performance. We have also drawn several interesting empirical findings, including:
1) the larger the network, the more redundant features it learns; 2) in particular, we show how to prune the networks based on our finding using direct equivalent feature merging, without fine-tuning which is often needed in peer network pruning methods; 3) same structured networks with higher feature complexity achieve better performance; 4) through the layers of a neural network, the feature complexity first increase then decrease; 5) for the image classification task, a group of functionally equivalent features may correspond to a specific semantic meaning. Source code will be made publicly available.
****
Siyu Ren,Zhiyong Wu,Kenny Q. Zhu
EMO: EARTH MOVER DISTANCE OPTIMIZATION FOR AUTO-REGRESSIVE LANGUAGE MODELING
Neural language models are probabilistic models of human text. They are predominantly trained using maximum likelihood estimation (MLE), which is equivalent to minimizing the forward cross-entropy between the empirical data distribution and

the model distribution. However, various degeneration phenomena are still widely observed when decoding from the distributions learned by such models. We establish that the forward cross-entropy is suboptimal as a distance metric for aligning human and model distribution due to its (1) recall-prioritization (2) negative diversity ignorance and (3) train-test mismatch. In this paper, we propose Earth Mover Distance Optimization (EMO) for auto-regressive language modeling. EMO capitalizes on the inherent properties of earth mover distance to address the aforementioned challenges. Due to the high complexity of direct computation, we further introduce a feasible upper bound for EMO to ease end-to-end training. Upon extensive evaluation of language models trained using EMO and MLE. We find that EMO demonstrates a consistently better language modeling performance than MLE across domains. Moreover, EMO demonstrates noteworthy enhancements in downstream performance with minimal fine-tuning on merely 25,000 sentences. This highlights the tremendous potential of EMO as a lightweight calibration method for enhancing large-scale pre-trained language models.
****
Yujia Wang,Yuanpu Cao,Jingcheng Wu,Ruoyu Chen,Jinghui Chen
Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration
Asynchronous federated learning, which enables local clients to send their model update asynchronously to the server without waiting for others, has recently emerged for its improved efficiency and scalability over traditional synchronized federated learning. In this paper, we study how the asynchronous delay affects the convergence of asynchronous federated learning under non-i.i.d. distributed data across clients. Through the theoretical convergence analysis of one representative asynchronous federated learning algorithm under standard nonconvex stochastic settings, we show that the asynchronous delay can largely slow down the convergence, especially with high data heterogeneity. To further improve the convergence of asynchronous federated learning under heterogeneous data distributions, we propose a novel asynchronous federated learning method with a cached update calibration. Specifically, we let the server cache the latest update for each client and reuse these variables for calibrating the global update at each round. We theoretically prove the convergence acceleration for our proposed method under nonconvex stochastic settings. Extensive experiments on several vision and language tasks demonstrate our superior performances compared to other asynchronous federated learning baselines.
****
Yash Chandak,Shiv Shankar,Vasilis Syrgkanis,Emma Brunskill
Adaptive Instrument Design for Indirect Experiments
Indirect experiments provide a valuable framework for estimating treatment effects in situations where conducting randomized control trials (RCTs) is impractical or unethical. Unlike RCTs, indirect experiments estimate treatment effects by leveraging (conditional) instrumental variables, enabling estimation through encouragement and recommendation rather than strict treatment assignment. However, the sample efficiency of such estimators depends not only on the inherent variability in outcomes but also on the varying compliance levels of users with the instrumental variables and the choice of estimator being used, especially when dealing with numerous instrumental variables. While adaptive experiment design has a rich literature for \textit{direct} experiments, in this paper we take the initial steps towards enhancing sample efficiency for \textit{indirect} experiments by adaptively designing a data collection policy over instrumental variables. Our main contribution is a practical computational procedure that utilizes influence functions to search for an optimal data collection policy, minimizing the mean-squared error of the desired (non-linear) estimator. Through experiments conducted in various domains inspired by real-world applications, we showcase how our method can significantly improve the sample efficiency of indirect experiments.
****
Siqi Liu,Luke Marris,Georgios Piliouras,Ian Gemp,Nicolas Heess
NfgTransformer: Equivariant Representation Learning for Normal-form Games

Normal-form games (NFGs) are the fundamental model of *strategic interaction*. We study their representation using neural networks. We describe the inherent equivariance of NFGs --- any permutation of strategies describes an equivalent game --- as well as the challenges this poses for representation learning. We then propose the NfgTransformer architecture that leverages this equivariance, leading to state-of-the-art performance in a range of game-theoretic tasks including equilibrium-solving, deviation gain estimation and ranking, with a common approach to NFG representation. We show that the resulting model is interpretable and versatile, paving the way towards deep learning systems capable of game-theoretic reasoning when interacting with humans and with each other.
****

Zhangir Azerbayev,Hailey Schoelkopf,Keiran Paster,Marco Dos Santos,Stephen Marcus McAleer,Albert Q. Jiang,Jia Deng,Stella Biderman,Sean Welleck
Llemma: An Open Language Model for Mathematics
We present Llemma, a large language model for mathematics. We continue pretraining Code Llama on the Proof-Pile-2, a mixture of scientific papers, web data containing mathematics, and mathematical code, yielding Llemma. On the MATH benchmark Llemma outperforms all known openly released models, as well as the unreleased Minerva model suite on an equi-parameter basis. Moreover, Llemma is capable of tool use and formal theorem proving without any finetuning. We openly release all artifacts, including 7 billion and 34 billion parameter models, the Proof-Pile-2, and code to replicate our experiments.
****

Zhepeng Cen,Zuxin Liu,Zitong Wang,Yihang Yao,Henry Lam,Ding Zhao
Learning from Sparse Offline Datasets via Conservative Density Estimation
Offline reinforcement learning (RL) offers a promising direction for learning policies from pre-collected datasets without requiring further interactions with the environment. However, existing methods struggle to handle out-of-distribution (OOD) extrapolation errors, especially in sparse reward or scarce data settings. In this paper, we propose a novel training algorithm called Conservative Density Estimation (CDE), which addresses this challenge by explicitly imposing constraints on the state-action occupancy stationary distribution. CDE overcomes the limitations of existing approaches, such as the stationary distribution correction method, by addressing the support mismatch issue in marginal importance sampling. Our method achieves state-of-the-art performance on the D4RL benchmark. Notably, CDE consistently outperforms baselines in challenging tasks with sparse rewards or insufficient data, demonstrating the advantages of our approach in addressing the extrapolation error problem in offline RL.
****

Zahra Babaiee,Peyman Kiasari,Daniela Rus,Radu Grosu
Unveiling the Unseen: Identifiable Clusters in Trained Depthwise Convolutional Kernels
Recent advances in depthwise-separable convolutional neural networks (DS-CNNs) have led to novel architectures, that surpass the performance of classical CNNs, by a considerable scalability and accuracy margin. This paper reveals another striking property of DS-CNN architectures: discernible and explainable patterns emerge in their trained depthwise convolutional kernels in all layers. Through an extensive analysis of millions of trained filters, with different sizes and from various models, we employed unsupervised clustering with autoencoders, to categorize these filters. Astonishingly, the patterns converged into a few main clusters, each resembling the difference of Gaussian (DoG) functions, and their first and second-order derivatives. Notably, we classify over 95\% and 90\% of the filters from state-of-the-art ConvNeXtV2 and ConvNeXt models, respectively. This finding is not merely a technological curiosity; it echoes the foundational models neuroscientists have long proposed for the vision systems of mammals. Our results thus deepen our understanding of the emergent properties of trained DS-CNNs and provide a bridge between artificial and biological visual processing systems. More broadly, they pave the way for more interpretable and biologically-inspired neural network designs in the future.
****

YongKyung Oh,Dongyoung Lim,Sungil Kim
Stable Neural Stochastic Differential Equations in Analyzing Irregular Time Series Data
Irregular sampling intervals and missing values in real-world time series data present challenges for conventional methods that assume consistent intervals and complete data. Neural Ordinary Differential Equations (Neural ODEs) offer an alternative approach, utilizing neural networks combined with ODE solvers to learn continuous latent representations through parameterized vector fields. Neural Stochastic Differential Equations (Neural SDEs) extend Neural ODEs by incorporating a diffusion term, although this addition is not trivial, particularly when addressing irregular intervals and missing values. Consequently, careful design of drift and diffusion functions is crucial for maintaining stability and enhancing performance, while incautious choices can result in adverse properties such as the absence of strong solutions, stochastic destabilization, or unstable Euler discretizations, significantly affecting Neural SDEs' performance. In this study, we propose three stable classes of Neural SDEs: Langevin-type SDE, Linear Noise SDE, and Geometric SDE. Then, we rigorously demonstrate their robustness in maintaining excellent performance under distribution shift, while effectively preventing overfitting. To assess the effectiveness of our approach, we conduct extensive experiments on four benchmark datasets for interpolation, forecasting, and classification tasks, and analyze the robustness of our methods with 30 public datasets under different missing rates. Our results demonstrate the efficacy of the proposed method in handling real-world irregular time series data.
****
Gen Li,Yuting Wei,Yuxin Chen,Yuejie Chi
Towards Non-Asymptotic Convergence for Diffusion-Based Generative Models
Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. In this work, we develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models in discrete time, assuming access to ℓ_2-accurate estimates of the (Stein) score functions. For a popular deterministic sampler (based on the probability flow ODE), we establish a convergence rate proportional to $1/T$ (with T the total number of steps), improving upon past results; for another mainstream stochastic sampler (i.e., a type of the denoising diffusion probabilistic model), we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results characterize how ℓ_2 score estimation errors affect the quality of the data generation process. In contrast to prior works, our theory is developed based on an elementary yet versatile non-asymptotic approach without resorting to toolboxes for SDEs and ODEs.
****
Joo Chan Lee,Daniel Rho,Seungtae Nam,Jong Hwan Ko,Eunbyung Park
Coordinate-Aware Modulation for Neural Fields
Neural fields, mapping low-dimensional input coordinates to corresponding signals, have shown promising results in representing various signals. Numerous methodologies have been proposed, and techniques employing MLPs and grid representations have achieved substantial success. MLPs allow compact and high expressibility, yet often suffer from spectral bias and slow convergence speed. On the other hand, methods using grids are free from spectral bias and achieve fast training speed, however, at the expense of high spatial complexity. In this work, we propose a novel way for exploiting both MLPs and grid representations in neural fields. Unlike the prevalent methods that combine them sequentially (extract features from the grids first and feed them to the MLP), we inject spectral bias-free grid representations into the intermediate features in the MLP. More specifically, we suggest a Coordinate-Aware Modulation (CAM), which modulates the intermediate features using scale and shift parameters extracted from the grid representations. This can maintain the strengths of MLPs while mitigating any remaining pote

ntial biases, facilitating the rapid learning of high-frequency components. In a
ddition, we empirically found that the feature normalizations, which have not be
en successful in neural filed literature, proved to be effective when applied in
 conjunction with the proposed CAM. Experimental results demonstrate that CAM en
hances the performance of neural representation and improves learning stability
across a range of signals. Especially in the novel view synthesis task, we achie
ved state-of-the-art performance with the least number of parameters and fast tr
aining speed for dynamic scenes and the best performance under 1MB memory for st
atic scenes. CAM also outperforms the best-performing video compression methods
using neural fields by a large margin. Our project page is available at https://
maincold2.github.io/cam/.
****

Federico Barbero,Ameya Velingker,Amin Saberi,Michael M. Bronstein,Francesco Di G
iovanni

Locality-Aware Graph Rewiring in GNNs

Graph Neural Networks (GNNs) are popular models for machine learning on graphs t
hat typically follow the message-passing paradigm, whereby the feature of a node
 is updated recursively upon aggregating information over its neighbors. While e
xchanging messages over the input graph endows GNNs with a strong inductive bias
, it can also make GNNs susceptible to over-squashing, thereby preventing them f
rom capturing long-range interactions in the given graph. To rectify this issue,
 graph rewiring techniques have been proposed as a means of improving informatio
n flow by altering the graph connectivity. In this work, we identify three desid
erata for graph-rewiring: (i) reduce over-squashing, (ii) respect the locality o
f the graph, and
(iii) preserve the sparsity of the graph. We highlight fundamental trade-offs th
at occur between spatial and spectral rewiring techniques; while the former ofte
n satisfy (i) and (ii) but not (iii), the latter generally satisfy (i) and (iii)
 at the expense of (ii). We propose a novel rewiring framework that satisfies al
l of (i)--(iii) through a locality-aware sequence of rewiring operations. We the
n discuss a specific instance of such rewiring framework and
validate its effectiveness on several real-world benchmarks, showing that it eit
her matches or significantly outperforms existing rewiring approaches.
****

Xiangyu Dong,Xingyi Zhang,Sibo Wang

Rayleigh Quotient Graph Neural Networks for Graph-level Anomaly Detection

Graph-level anomaly detection has gained significant attention as it finds appli
cations in various domains, such as cancer diagnosis and enzyme prediction. Howe
ver, existing methods fail to capture the spectral properties of graph anomalies
, resulting in unexplainable framework design and unsatisfying performance. In t
his paper, we re-investigate the spectral differences between anomalous and norm
al graphs. Our main observation shows a significant disparity in the accumulated
 spectral energy between these two classes. Moreover, we prove that the accumula
ted spectral energy of the graph signal can be represented by its Rayleigh Quoti
ent, indicating that the Rayleigh Quotient is a driving factor behind the anomal
ous properties of graphs. Motivated by this, we propose Rayleigh Quotient Graph
Neural Network (RQGNN), the first spectral GNN that explores the inherent spectr
al features of anomalous graphs for graph-level anomaly detection. Specifically,
 we introduce a novel framework with two components: the Rayleigh Quotient learn
ing component (RQL) and Chebyshev Wavelet GNN with RQ-pooling (CWGNN-RQ). RQL ex
plicitly captures the Rayleigh Quotient of graphs and CWGNN-RQ implicitly explor
es the spectral space of graphs. Extensive experiments on 10 real-world datasets
 show that RQGNN outperforms the best rival by 6.74% in Macro-F1 score and 1.44%
 in AUC, demonstrating the effectiveness of our framework. Our code is available
 at https://github.com/xydong127/RQGNN.
****

Luke Nicholas Darlow,Qiwen Deng,Ahmed Hassan,Martin Asenov,Rajkarn Singh,Artjom
Joosen,Adam Barker,Amos Storkey

DAM: Towards a Foundation Model for Forecasting

It is challenging to scale time series forecasting models such that they forecas

t accurately for multiple distinct domains and datasets, all with potentially different underlying collection procedures (e.g., sample resolution), patterns (e.g., periodicity), and prediction requirements (e.g., reconstruction vs. forecasting). We call this general task universal forecasting. Existing methods usually assume that input data is regularly sampled, and they forecast to pre-determined horizons, resulting in failure to generalise outside of the scope of their training. We propose the DAM -- a neural model that takes randomly sampled histories and outputs an adjustable basis composition as a continuous function of time for forecasting to non-fixed horizons. It involves three key components: (1) a flexible approach for using randomly sampled histories from a long-tail distribution, that enables an efficient global perspective of the underlying temporal dynamics while retaining focus on the recent history; (2) a transformer backbone that is trained on these actively sampled histories to produce, as representational output, (3) the basis coefficients of a continuous function of time. We show that a single univariate DAM, trained on 25 time series datasets, either outperformed or closely matched existing SoTA models at multivariate long-term forecasting across 18 datasets, including 8 held-out for zero-shot transfer, even though these models were trained to specialise for each dataset-horizon combination. This single DAM excels at zero-shot transfer and very-long-term forecasting, performs well at imputation, is interpretable via basis function composition and attention, can be tuned for different inference-cost requirements, is robust to missing and irregularly sampled data by design.
****

Tanvir Mahmud,Saeed Amizadeh,Kazuhito Koishida,Diana Marculescu
Weakly-supervised Audio Separation via Bi-modal Semantic Similarity
Conditional sound separation in multi-source audio mixtures without having access to single source sound data during training is a long standing challenge. Existing mix-and-separate based methods suffer from significant performance drop with multi-source training mixtures due to the lack of supervision signal for single source separation cases during training. However, in the case of language-conditional audio separation, we do have access to corresponding text descriptions for each audio mixture in our training data, which can be seen as (rough) representations of the audio samples in the language modality. That raises the curious question of how to generate supervision signal for single-source audio extraction by leveraging the fact that single-source sounding language entities can be easily extracted from the text description. To this end, in this paper, we propose a generic bi-modal separation framework which can enhance the existing unsupervised frameworks to separate single-source signals in a target modality (i.e., audio) using the easily separable corresponding signals in the conditioning modality (i.e., language), without having access to single-source samples in the target modality during training. We empirically show that this is well within reach if we have access to a pretrained joint embedding model between the two modalities (i.e., CLAP). Furthermore, we propose to incorporate our framework into two fundamental scenarios to enhance separation performance. First, we show that our proposed methodology significantly improves the performance of purely unsupervised baselines by reducing the distribution shift between training and test samples. In particular, we show that our framework can achieve 71% boost in terms of Signal-to-Distortion Ratio (SDR) over the baseline, reaching 97.5% of the supervised learning performance. Second, we show that we can further improve the performance of the supervised learning itself by 17% if we augment it by our proposed weakly-supervised framework. Our framework achieves this by making large corpora of unsupervised data available to the supervised learning model as well as utilizing a natural, robust regularization mechanism through weak supervision from the language modality, and hence enabling a powerful semi-supervised framework for audio separation.
****

Ziqi Gao,Tao Feng,Jiaxuan You,Chenyi Zi,Yan Zhou,Chen Zhang,Jia Li
Deep Reinforcement Learning for Modelling Protein Complexes
Structure prediction of large protein complexes (a.k.a., protein multimer mod-elling, PMM) can be achieved through the one-by-one assembly using provided

dimer structures and predicted docking paths. However, existing PMM methods struggle with vast search spaces and generalization challenges: (1) The assembly of a N -chain multimer can be depicted using graph structured data, with each chain represented as a node and assembly actions as edges. Thus the assembly graph can be arbitrary acyclic undirected connected graph, leading to the combinatorial optimization space of N^(N −2) for the PMM problem. (2) Knowledge transfer in the PMM task is non-trivial. The gradually limited data availability as
the chain number increases necessitates PMM models that can generalize across multimers of various chains. To address these challenges, we propose GAPN, a Generative Adversarial Policy Network powered by domain-specific rewards and adversarial loss through policy gradient for automatic PMM prediction. Specifically, GAPN learns to efficiently search through the immense assembly space and optimize the direct docking reward through policy gradient. Importantly, we design a adversarial reward function to enhance the receptive field of our model. In
this way, GAPN will simultaneously focus on a specific batch of multimers and the global assembly rules learned from multimers with varying chain numbers. Empirically, we have achieved both significant accuracy (measured by RMSD and TM-Score) and efficiency improvements compared to leading complex modeling software. GAPN outperforms the state-of-the-art method (MoLPC) with up to 27% improvement in TM-Score, with a speed-up of 600×.
****
Zilong Wang,Hao Zhang,Chun-Liang Li,Julian Martin Eisenschlos,Vincent Perot,Zifeng Wang,Lesly Miculicich,Yasuhisa Fujii,Jingbo Shang,Chen-Yu Lee,Tomas Pfister
Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding
Table-based reasoning with large language models (LLMs) is a promising direction to tackle many table understanding tasks, such as table-based question answering and fact verification. Compared with generic reasoning, table-based reasoning requires the extraction of underlying semantics from both free-form questions and semi-structured tabular data. Chain-of-Thought and its similar approaches incorporate the reasoning chain in the form of textual context, but it is still an open question how to effectively leverage tabular data in the reasoning chain. We propose the Chain-of-Table framework, where tabular data is explicitly used in the reasoning chain as a proxy for intermediate thoughts. Specifically, we guide LLMs using in-context learning to iteratively generate operations and update the table to represent a tabular reasoning chain. LLMs can therefore dynamically plan the next operation based on the results of the previous ones. This continuous evolution of the table forms a chain, showing the reasoning process for a given tabular problem. The chain carries structured information of the intermediate results, enabling more accurate and reliable predictions. Chain-of-Table achieves new state-of-the-art performance on WikiTQ, FeTaQA, and TabFact benchmarks across multiple LLM choices.
****
Zhang-Wei Hong,Idan Shenfeld,Tsun-Hsuan Wang,Yung-Sung Chuang,Aldo Pareja,James R. Glass,Akash Srivastava,Pulkit Agrawal
Curiosity-driven Red-teaming for Large Language Models
Large language models (LLMs) hold great potential for many natural language applications but risk generating incorrect or toxic content. To probe when an LLM generates unwanted content, the current paradigm is to recruit a $\textit{red team}$ of human testers to design input prompts (i.e., test cases) that elicit undesirable responses from LLMs.
However, relying solely on human testers is expensive and time-consuming. Recent works automate red teaming by training a separate red team LLM with reinforcement learning (RL) to generate test cases that maximize the chance of eliciting undesirable responses from the target LLM. However, current RL methods are only able to generate a small number of effective test cases resulting in a low coverage of the span of prompts that elicit undesirable responses from the target LLM. To overcome this limitation, we draw a connection between the problem of increasing the coverage of generated test cases and the well-studied approach of curios

ity-driven exploration that optimizes for novelty.
Our method of curiosity-driven red teaming (CRT) achieves greater coverage of test cases while mantaining or increasing their effectiveness compared to existing methods.
Our method, CRT successfully provokes toxic responses from LLaMA2 model that has been heavily fine-tuned using human preferences to avoid toxic outputs. Code is available at https://github.com/Improbable-AI/curiosity_redteam.
****
Atharva Sehgal,Arya Grayeli,Jennifer J. Sun,Swarat Chaudhuri
Neurosymbolic Grounding for Compositional Generalization
We introduce Cosmos, a framework for object-centric world modeling that is designed for compositional generalization (CG), i.e., high performance on unseen input scenes obtained through the composition of known visual "atoms." The central insight behind Cosmos is the use of a novel form of neurosymbolic grounding. Specifically, the framework introduces two new tools: (i) neurosymbolic scene encodings, which represent each entity in a scene using a real vector computed using a neural encoder, as well as a vector of composable symbols describing attributes of the entity, and (ii) a neurosymbolic attention mechanism that binds these entities to learned rules of interaction. Cosmos is end-to-end differentiable; also, unlike traditional neurosymbolic methods that require representations to be manually mapped to symbols, it computes an entity's symbolic attributes using vision-language foundation models. Through an evaluation that considers two different forms of CG on an established blocks-pushing domain, we show that the framework establishes a new state-of-the-art for CG in world modeling.
****
Cheng Shi,Sibei Yang
The Devil is in the Object Boundary: Towards Annotation-free Instance Segmentation using Foundation Models
Foundation models, pre-trained on a large amount of data have demonstrated impressive zero-shot capabilities in various downstream tasks. However, in object detection and instance segmentation, two fundamental computer vision tasks heavily reliant on extensive human annotations, foundation models such as SAM and DINO struggle to achieve satisfactory performance.
In this study, we reveal that the devil is in the object boundary, $\textit{i.e.}$, these foundation models fail to discern boundaries between individual objects.
For the first time, we probe that CLIP, which has never accessed any instance-level annotations, can provide a highly beneficial and strong instance-level boundary prior in the clustering results of its particular intermediate layer. Following this surprising observation, we propose $\textbf{\textit{Zip}}$ which $\textbf{Z}$ips up CL$\textbf{ip}$ and SAM in a novel classification-first-then-discovery pipeline, enabling annotation-free, complex-scene-capable, open-vocabulary object detection and instance segmentation.
Our Zip significantly boosts SAM's mask AP on COCO dataset by 12.5\% and establishes state-of-the-art performance in various settings, including training-free, self-training, and label-efficient finetuning. Furthermore, annotation-free Zip even achieves comparable performance to the best-performing open-vocabulary object detecters using base annotations. Code is released at https://github.com/ChengShiest/Zip-Your-CLIP
****
Hao Liu,Jiarui Feng,Lecheng Kong,Ningyue Liang,Dacheng Tao,Yixin Chen,Muhan Zhang
One For All: Towards Training One Graph Model For All Classification Tasks
Designing a single model to address multiple tasks has been a long-standing objective in artificial intelligence. Recently, large language models have demonstrated exceptional capability in solving different tasks within the language domain. However, a unified model for various graph tasks remains underexplored, primarily due to the challenges unique to the graph learning domain. First, graph data from different areas carry distinct attributes and follow different distributions. Such discrepancy makes it hard to represent graphs in a single representatio

n space. Second, tasks on graphs diversify into node, link, and graph tasks, req
uiring distinct embedding strategies. Finally, an appropriate graph prompting pa
radigm for in-context learning is unclear. We propose **One for All (OFA)**, the
 first general framework that can use a single graph model to address the above
challenges. Specifically, OFA proposes text-attributed graphs to unify different
 graph data by describing nodes and edges with natural language and uses languag
e models to encode the diverse and possibly cross-domain text attributes to feat
ure vectors in the same embedding space. Furthermore, OFA introduces the concept
 of nodes-of-interest to standardize different tasks with a single task represen
tation. For in-context learning on graphs, OFA introduces a novel graph promptin
g paradigm that appends prompting substructures to the input graph, which enable
s it to address varied tasks without fine-tuning. We train the OFA model using g
raph data from multiple domains (including citation networks, molecular graphs,
knowledge graphs, etc.) simultaneously and evaluate its ability in supervised, f
ew-shot, and zero-shot learning scenarios. OFA performs well across different ta
sks, making it the first general-purpose across-domains classification model on
graphs.
****
Harsh Chaudhari,Giorgio Severi,Alina Oprea,Jonathan Ullman
Chameleon: Increasing Label-Only Membership Leakage with Adaptive Poisoning
The integration of Machine Learning (ML) in numerous critical applications intro
duces a range of privacy concerns for individuals who provide their datasets for
 ML training purposes. One such privacy risk is Membership Inference (MI), in wh
ich an adversary seeks to determine whether a particular data point was included
 in the training dataset of a model. Current state-of-the-art MI approaches capi
talize on access to the model’s predicted confidence scores to successfully perf
orm membership inference, and employ data poisoning to further enhance their eff
ectiveness.
In this work, we focus on the less explored and more realistic label-only setti
ng, where the model provides only the predicted label as output. We show that ex
isting label-only attacks are ineffective at inferring membership in the low Fal
se Positive Rate (FPR) regime. To address this challenge, we propose a new atta
ck Chameleon that leverages a novel data poisoning strategy and an efficient que
ry selection method to achieve significantly more accurate membership inference
than existing label-only attacks, especially for low FPRs.
****
Hengrui Zhang,Jiani Zhang,Zhengyuan Shen,Balasubramaniam Srinivasan,Xiao Qin,Chr
istos Faloutsos,Huzefa Rangwala,George Karypis
Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space
Recent advances in tabular data generation have greatly enhanced synthetic data
quality. However, extending diffusion models to tabular data is challenging due
to the intricately varied distributions and a blend of data types of tabular dat
a. This paper introduces TabSyn, a methodology that synthesizes tabular data by
leveraging a diffusion model within a variational autoencoder (VAE) crafted late
nt space. The key advantages of the proposed Tabsyn include (1) Generality: the
ability to handle a broad spectrum of data types by converting them into a singl
e unified space and explicitly capturing inter-column relations; (2) Quality: op
timizing the distribution of latent embeddings to enhance the subsequent trainin
g of diffusion models, which helps generate high-quality synthetic data; (3) Spe
ed: much fewer number of reverse steps and faster synthesis speed than existing
diffusion-based methods. Extensive experiments on six datasets with five metrics
 demonstrate that Tabsyn outperforms existing methods. Specifically, it reduces
the error rates by 86% and 67% for column-wise distribution and pair-wise column
 correlation estimations compared with the most competitive baselines. The code
has been made available at https://github.com/amazon-science/tabsyn.
****
Rocio P Diaz Martin,Ivan Vladimir Medri,Yikun Bai,Xinran Liu,Kangbai Yan,Gustavo
 Rohde,Soheil Kolouri
LCOT: Linear Circular Optimal Transport
The optimal transport problem for measures supported on non-Euclidean spaces has

recently gained ample interest in diverse applications involving representation learning. In this paper, we focus on circular probability measures, i.e., probability measures supported on the unit circle, and introduce a new computationally efficient metric for these measures, denoted as Linear Circular Optimal Transport (LCOT). The proposed metric comes with an explicit linear embedding that allows one to apply Machine Learning (ML) algorithms to the embedded measures and seamlessly modify the underlying metric for the ML algorithm to LCOT. We show that the proposed metric is rooted in the Circular Optimal Transport (COT) and can be considered the linearization of the COT metric with respect to a fixed reference measure. We provide a theoretical analysis of the proposed metric and derive the computational complexities for pairwise comparison of circular probability measures. Lastly, through a set of numerical experiments, we demonstrate the benefits of LCOT in learning representations from circular measures.
****
Haozhe Ji,Pei Ke,Hongning Wang,Minlie Huang
Language Model Decoding as Direct Metrics Optimization
Despite the remarkable advances in language modeling, current mainstream decoding methods still struggle to generate texts that align with human texts across different aspects. In particular, sampling-based methods produce less-repetitive texts which are often disjunctive in discourse, while search-based methods maintain topic coherence at the cost of increased repetition. Overall, these methods fall short in achieving holistic alignment across a broad range of aspects. In this work, we frame decoding from a language model as an optimization problem with the goal of strictly matching the expected performance with human texts measured by multiple metrics of desired aspects simultaneously. The resulting decoding distribution enjoys an analytical solution that scales the input language model distribution via a sequence-level energy function defined by these metrics. And most importantly, we prove that this induced distribution is guaranteed to improve the perplexity on human texts, which suggests a better approximation to the underlying distribution of human texts. To facilitate tractable sampling from this globally normalized distribution, we adopt the Sampling-Importance-Resampling technique. Experiments on various domains and model scales demonstrate the superiority of our method in metrics alignment with human texts and human evaluation over strong baselines.
****
Lijia Yu,Xiao-Shan Gao,Lijun Zhang
OPTIMAL ROBUST MEMORIZATION WITH RELU NEURAL NETWORKS
Memorization with neural networks is to study the expressive power of neural networks to interpolate a finite classification data set, which is closely related to the generalizability of deep learning. However, the important problem of robust memorization has not been thoroughly studied. In this paper, several basic problems about robust memorization are solved. First, we prove that it is NP-hard to compute neural networks with certain simple structures, which are robust memorization. A network hypothesis space is called optimal robust memorization for a data set if it can achieve robust memorization for any budget less than half the separation bound of the data set. Second, we explicitly construct neural networks with O(N n) parameters for optimal robust memorization of any data set with dimension n and size N . We also give a lower bound for the width of networks to achieve optimal robust memorization. Finally, we explicitly construct neural networks with
O(N n log n) parameters for optimal robust memorization of any binary classification data set by controlling the Lipschitz constant of the network.
****
Haiquan Qiu,Yongqi Zhang,Yong Li,quanming yao
Understanding Expressivity of GNN in Rule Learning
Rule learning is critical to improving knowledge graph (KG) reasoning due to their ability to provide logical and interpretable explanations. Recently, Graph Neural Networks (GNNs) with tail entity scoring achieve the state-of-the-art performance on KG reasoning. However, the theoretical understandings for these GNNs are either lacking or focusing on single-relational graphs, leaving what the kind

of rules these GNNs can learn an open problem. We propose to fill the above gap in this paper. Specifically, GNNs with tail entity scoring are unified into a common framework. Then, we analyze their expressivity by formally describing the rule structures they can learn and theoretically demonstrating their superiority. These results further inspire us to propose a novel labeling strategy to learn more rules in KG reasoning. Experimental results are consistent with our theoretical findings and verify the effectiveness of our proposed method. The code is publicly available at https://github.com/LARS-research/Rule-learning-expressivity.
****
Yiwei Zhang,Guo Lu,Yunuo Chen,Shen Wang,Yibo Shi,Jing Wang,Li Song
Neural Rate Control for Learned Video Compression
The learning-based video compression method has made significant progress in recent years, exhibiting promising compression performance compared with traditional video codecs. However, prior works have primarily focused on advanced compression architectures while neglecting the rate control technique. Rate control can precisely control the coding bitrate with optimal compression performance, which is a critical technique in practical deployment. To address this issue, we present a fully neural network-based rate control system for learned video compression methods. Our system accurately encodes videos at a given bitrate while enhancing the rate-distortion performance. Specifically, we first design a rate allocation model to assign optimal bitrates to each frame based on their varying spatial and temporal characteristics. Then, we propose a deep learning-based rate implementation network to perform the rate-parameter mapping, precisely predicting coding parameters for a given rate. Our proposed rate control system can be easily integrated into existing learning-based video compression methods. The extensive experimental results show that the proposed method achieves accurate rate control on several baseline methods while also improving overall rate-distortion performance.
****
Allan Jabri,Sjoerd van Steenkiste,Emiel Hoogeboom,Mehdi S. M. Sajjadi,Thomas Kipf
DORSal: Diffusion for Object-centric Representations of Scenes $\textit{et al.}$
Recent progress in 3D scene understanding enables scalable learning of representations across large datasets of diverse scenes. As a consequence, generalization to unseen scenes and objects, rendering novel views from just a single or a handful of input images, and controllable scene generation that supports editing, is now possible. However, training jointly on a large number of scenes typically compromises rendering quality when compared to single-scene optimized models such as NeRFs. In this paper, we leverage recent progress in diffusion models to equip 3D scene representation learning models with the ability to render high-fidelity novel views, while retaining benefits such as object-level scene editing to a large degree. In particular, we propose DORSal, which adapts a video diffusion architecture for 3D scene generation conditioned on frozen object-centric slot-based representations of scenes. On both complex synthetic multi-object scenes and on the real-world large-scale Street View dataset, we show that DORSal enables scalable neural rendering of 3D scenes with object-level editing and improves upon existing approaches.
****
Maxwell Xu,Alexander Moreno,Hui Wei,Benjamin Marlin,James Matthew Rehg
REBAR: Retrieval-Based Reconstruction for Time-series Contrastive Learning
The success of self-supervised contrastive learning hinges on identifying positive data pairs, such that when they are pushed together in embedding space, the space encodes useful information for subsequent downstream tasks. Constructing positive pairs is non-trivial as the pairing must be similar enough to reflect a shared semantic meaning, but different enough to capture within-class variation. Classical approaches in vision use augmentations to exploit well-established invariances to construct positive pairs, but invariances in the time-series domain are much less obvious. In our work, we propose a novel method of using a learned measure for identifying positive pairs. Our Retrieval-Based Reconstruction (REB

AR) measure measures the similarity between two sequences as the reconstruction error that results from reconstructing one sequence with retrieved information from the other. Then, if the two sequences have high REBAR similarity, we label them as a positive pair. Through validation experiments, we show that the REBAR error is a predictor of mutual class membership. Once integrated into a contrastive learning framework, our REBAR method learns an embedding that achieves state-of-the-art performance on downstream tasks across various modalities.
****
Rishabh Agarwal,Nino Vieillard,Yongchao Zhou,Piotr Stanczyk,Sabela Ramos Garea,Matthieu Geist,Olivier Bachem
On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes
Knowledge distillation (KD) is widely used for compressing a teacher model to reduce its inference cost and memory footprint, by training a smaller student model. However, current KD methods for auto-regressive sequence models suffer from distribution mismatch between output sequences seen during training and those generated by the student during inference. To address this issue, we introduce Generalized Knowledge Distillation (GKD). Instead of solely relying on a fixed set of output sequences, GKD trains the student on its self-generated output sequences by leveraging feedback from the teacher on such sequences. Unlike supervised KD approaches, GKD also offers the flexibility to employ alternative loss functions between the student and teacher, which can be useful when the student lacks the expressivity to mimic the teacher's distribution. Furthermore, GKD facilitates the seamless integration of distillation with RL fine-tuning (RLHF). We demonstrate the efficacy of GKD for distilling auto-regressive T5 language models on summarization, translation, and arithmetic reasoning tasks.
****
Sung Moon Ko,Sumin Lee,Dae-Woong Jeong,Woohyung Lim,Sehui Han
Geometrically Aligned Transfer Encoder for Inductive Transfer in Regression Tasks
Transfer learning is a crucial technique for handling a small amount of data that is potentially related to other abundant data. However, most of the existing methods are focused on classification tasks using images and language datasets. Therefore, in order to expand the transfer learning scheme to regression tasks, we propose a novel transfer technique based on differential geometry, namely the Geometrically Aligned Transfer Encoder (${\it GATE}$). In this method, we interpret the latent vectors from the model to exist on a Riemannian curved manifold. We find a proper diffeomorphism between pairs of tasks to ensure that every arbitrary point maps to a locally flat coordinate in the overlapping region, allowing the transfer of knowledge from the source to the target data. This also serves as an effective regularizer for the model to behave in extrapolation regions. In this article, we demonstrate that ${\it GATE}$ outperforms conventional methods and exhibits stable behavior in both the latent space and extrapolation regions for various molecular graph datasets.
****
Archibald Felix Fraikin,Adrien Bennetot,Stephanie Allassonniere
T-Rep: Representation Learning for Time Series using Time-Embeddings
Multivariate time series present challenges to standard machine learning techniques, as they are often unlabeled, high dimensional, noisy, and contain missing data. To address this, we propose T-Rep, a self-supervised method to learn time series representations at a timestep granularity. T-Rep learns vector embeddings of time alongside its feature extractor, to extract temporal features such as trend, periodicity, or distribution shifts from the signal. These time-embeddings are leveraged in pretext tasks, to incorporate smooth and fine-grained temporal dependencies in the representations, as well as reinforce robustness to missing data. We evaluate T-Rep on downstream classification, forecasting, and anomaly detection tasks. It is compared to existing self-supervised algorithms for time series, which it outperforms in all three tasks. We test T-Rep in missing data regimes, where it proves more resilient than its counterparts. Finally, we provide latent space visualisation experiments, highlighting the interpretability of the learned representations.

****
Yohann Benchetrit,Hubert Banville,Jean-Remi King
Brain decoding: toward real-time reconstruction of visual perception
In the past five years, the use of generative and foundational AI systems has gr
eatly improved the decoding of brain activity. Visual perception, in particular,
 can now be decoded from functional Magnetic Resonance Imaging (fMRI) with remar
kable fidelity. This neuroimaging technique, however, suffers from a limited tem
poral resolution ($\approx$0.5\,Hz) and thus fundamentally constrains its real-t
ime usage. Here, we propose an alternative approach based on magnetoencephalogra
phy (MEG), a neuroimaging device capable of measuring brain activity with high t
emporal resolution ($\approx$5,000 Hz). For this, we develop an MEG decoding mod
el trained with both contrastive and regression objectives and consisting of thr
ee modules: i) pretrained embeddings obtained from the image, ii) an MEG module
trained end-to-end and iii) a pretrained image generator. Our results are threef
old: Firstly, our MEG decoder shows a 7X improvement of image-retrieval over cla
ssic linear decoders. Second, late brain responses to images are best decoded wi
th DINOv2, a recent foundational image model. Third, image retrievals and genera
tions both suggest that high-level visual features can be decoded from MEG signa
ls, although the same approach applied to 7T fMRI also recovers better low-level
 features. Overall, these results, while preliminary, provide an important step
towards the decoding - in real-time - of the visual processes continuously unfol
ding within the human brain.
****
Hong Liu,Zhiyuan Li,David Leo Wright Hall,Percy Liang,Tengyu Ma
Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-trai
ning
Given the massive cost of language model pre-training, a non-trivial improvement
 of the optimization algorithm would lead to a material reduction on the time an
d cost of training. Adam and its variants have been state-of-the-art for years,
and more sophisticated second-order (Hessian-based) optimizers often incur too m
uch per-step overhead. In this paper, we propose Sophia, a simple scalable secon
d-order optimizer that uses a light-weight estimate of the diagonal Hessian as t
he pre-conditioner. The update is the moving average of the gradients divided by
 the moving average of the estimated Hessian, followed by element-wise clipping.
 The clipping controls the worst-case update size and tames the negative impact
of non-convexity and rapid change of Hessian along the trajectory. Sophia only e
stimates the diagonal Hessian every handful of iterations, which has negligible
average per-step time and memory overhead. On language modeling with GPT models
of sizes ranging from 125M to 1.5B, Sophia achieves a 2x speed-up compared to Ad
am in the number of steps, total compute, and wall-clock time, achieving the sam
e perplexity with 50\% fewer steps, less total compute, and reduced wall-clock t
ime.
****
Jacob Mitchell Springer,Vaishnavh Nagarajan,Aditi Raghunathan
Sharpness-Aware Minimization Enhances Feature Quality via Balanced Learning
Sharpness-Aware Minimization (SAM) has emerged as a promising alternative to sto
chastic gradient descent (SGD) for minimizing the loss objective in neural netwo
rk training. While the motivation behind SAM is to bias models towards flatter m
inima that are believed to generalize better, recent studies have shown conflict
ing evidence on the relationship between flatness and (in-distribution) generali
zation, leaving the mechanism behind SAM's performance improvement unclear. In t
his work, we present a complementary effect that cannot be explained by in-distr
ibution improvements alone: we argue that SAM can enhance the quality of feature
s in datasets containing redundant or spurious features. We explain how SAM can
induce feature diversity by investigating a controlled setting. Our results impl
y that one mechanism by which SAM improves the quality of features is by adaptiv
ely suppressing well-learned features which gives remaining features opportunity
 to be learned.
****
Ishita Mediratta,Qingfei You,Minqi Jiang,Roberta Raileanu

The Generalization Gap in Offline Reinforcement Learning
Despite recent progress in offline learning, these methods are still trained and tested on the same environment. In this paper, we compare the generalization abilities of widely used online and offline learning methods such as online reinforcement learning (RL), offline RL, sequence modeling, and behavioral cloning. Our experiments show that offline learning algorithms perform worse on new environments than online learning ones. We also introduce the first benchmark for evaluating generalization in offline learning, collecting datasets of varying sizes and skill-levels from Procgen (2D video games) and WebShop (e-commerce websites). The datasets contain trajectories for a limited number of game levels or natural language instructions and at test time, the agent has to generalize to new levels or instructions. Our experiments reveal that existing offline learning algorithms struggle to match the performance of online RL on both train and test environments. Behavioral cloning is a strong baseline, outperforming state-of-the-art offline RL and sequence modeling approaches when trained on data from multiple environments and tested on new ones. Finally, we find that increasing the diversity of the data, rather than its size, improves performance on new environments for all offline learning algorithms. Our study demonstrates the limited generalization of current offline learning algorithms highlighting the need for more research in this area.
****
Enming Liang,Minghua Chen
Generative Learning for Solving Non-Convex Problem with Multi-Valued Input-Solution Mapping
By employing neural networks (NN) to learn input-solution mappings and passing a new input through the learned mapping to obtain a solution instantly, recent studies have shown remarkable speed improvements over iterative algorithms for solving optimization problems. Meanwhile, they also highlight methodological challenges to be addressed. In particular, general non-convex problems often present multiple optimal solutions for identical inputs, signifying a complex, multi-valued input-solution mapping. Conventional learning techniques, primarily tailored to learn single-valued mappings, struggle to train NNs to accurately decipher multi-valued ones, leading to inferior solutions. We address this fundamental issue by developing a generative learning approach using a rectified flow (RectFlow) model built upon ordinary differential equations. In contrast to learning input-solution mapping, we learn the mapping from input to solution distribution, exploiting the universal approximation capability of the RectFlow model. Upon receiving a new input, we employ the trained RectFlow model to sample high-quality solutions from the input-dependent distribution it has learned. Our approach outperforms conceivable GAN and Diffusion models in terms of training stability and run-time complexity. We provide a detailed characterization of the optimality loss and runtime complexity associated with our generative approach. Simulation results for solving non-convex problems show that our method achieves significantly better solution optimality than recent NN schemes, with comparable feasibility and speedup performance.
****
Zhentao Tan,Xiaodan Li,Yue Wu,Qi Chu,Le Lu,Nenghai Yu,Jieping Ye
Boosting Vanilla Lightweight Vision Transformers via Re-parameterization
Large-scale Vision Transformers have achieved promising performance on downstream tasks through feature pre-training. However, the performance of vanilla lightweight Vision Transformers (ViTs) is still far from satisfactory compared to that of recent lightweight CNNs or hybrid networks. In this paper, we aim to unlock the potential of vanilla lightweight ViTs by exploring the adaptation of the widely-used re-parameterization technology to ViTs for improving learning ability during training without increasing the inference cost. The main challenge comes from the fact that CNNs perfectly complement with re-parameterization over convolution and batch normalization, while vanilla Transformer architectures are mainly comprised of linear and layer normalization layers. We propose to incorporate the nonlinear ensemble into linear layers by expanding the depth of the linear layers with batch normalization and fusing multiple linear features with hierarch

ical representation ability through a pyramid structure. We also discover and solve a new transformer-specific distribution rectification problem caused by multi-branch re-parameterization. Finally, we propose our Two-Dimensional Re-parameterized Linear module (TDRL) for ViTs. Under the popular self-supervised pre-training and supervised fine-tuning strategy, our TDRL can be used in these two stages to enhance both generic and task-specific representation. Experiments demonstrate that our proposed method not only boosts the performance of vanilla Vit-Tiny on various vision tasks to new state-of-the-art (SOTA) but also shows promising generality ability on other networks. Code will be available.
****
Jianlang Chen,Xuhong Ren,Qing Guo,Felix Juefei-Xu,Di Lin,Wei Feng,Lei Ma,Jianjun Zhao
LRR: Language-Driven Resamplable Continuous Representation against Adversarial Tracking Attacks
Visual object tracking plays a critical role in visual-based autonomous systems, as it aims to estimate the position and size of the object of interest within a live video. Despite significant progress made in this field, state-of-the-art (SOTA) trackers often fail when faced with adversarial perturbations in the incoming frames. This can lead to significant robustness and security issues when these trackers are deployed in the real world. To achieve high accuracy on both clean and adversarial data, we propose building a spatial-temporal continuous representation using the semantic text guidance of the object of interest. This novel continuous representation enables us to reconstruct incoming frames to maintain semantic and appearance consistency with the object of interest and its clean counterparts. As a result, our proposed method successfully defends against different SOTA adversarial tracking attacks while maintaining high accuracy on clean data. In particular, our method significantly increases tracking accuracy under adversarial attacks with around 90% relative improvement on UAV123, which is even higher than the accuracy on clean data.
****
Xianghong Fang,Jian Li,Qiang Sun,Benyou Wang
Rethinking the Uniformity Metric in Self-Supervised Learning
Uniformity plays a crucial role in the assessment of learned representations, contributing to a deeper comprehension of self-supervised learning. The seminal work by \citet{Wang2020UnderstandingCR} introduced a uniformity metric that quantitatively measures the collapse degree of learned representations. Directly optimizing this metric together with alignment proves to be effective in preventing constant collapse. However, we present both theoretical and empirical evidence revealing that this metric lacks sensitivity to dimensional collapse, highlighting its limitations. To address this limitation and design a more effective uniformity metric, this paper identifies five fundamental properties, some of which the existing uniformity metric fails to meet. We subsequently introduce a novel uniformity metric that satisfies all of these desiderata and exhibits sensitivity to dimensional collapse. When applied as an auxiliary loss in various established self-supervised methods, our proposed uniformity metric consistently enhances their performance in downstream tasks.
****
Qihan Ren,Jiayang Gao,Wen Shen,Quanshi Zhang
Where We Have Arrived in Proving the Emergence of Sparse Interaction Primitives in DNNs
This study aims to prove the emergence of symbolic concepts (or more precisely, sparse primitive inference patterns) in well-trained deep neural networks (DNNs). Specifically, we prove the following three conditions for the emergence. (i) The high-order derivatives of the network output with respect to the input variables are all zero. (ii) The DNN can be used on occluded samples, and when the input sample is less occluded, the DNN will yield higher confidence. (iii) The confidence of the DNN does not significantly degrade on occluded samples. These conditions are quite common, and we prove that under these conditions, the DNN will only encode a relatively small number of sparse interactions between input variables. Moreover, we can consider such interactions as symbolic primitive inferenc

e patterns encoded by a DNN, because we show that inference scores of the DNN on an exponentially large number of randomly masked samples can always be well mimicked by numerical effects of just a few interactions.
****
Yang Liu,Jiashun Cheng,Haihong Zhao,Tingyang Xu,Peilin Zhao,Fugee Tsung,Jia Li,Yu Rong
SEGNO: Generalizing Equivariant Graph Neural Networks with Physical Inductive Biases
Graph Neural Networks (GNNs) with equivariant properties have emerged as powerful tools for modeling complex dynamics of multi-object physical systems. However, their generalization ability is limited by the inadequate consideration of physical inductive biases: (1) Existing studies overlook the continuity of transitions among system states, opting to employ several discrete transformation layers to learn the direct mapping between two adjacent states; (2) Most models only account for first-order velocity information, despite the fact that many physical systems are governed by second-order motion laws. To incorporate these inductive biases, we propose the Second-order Equivariant Graph Neural Ordinary Differential Equation (SEGNO). Specifically, we show how the second-order continuity can be incorporated into GNNs while maintaining the equivariant property. Furthermore, we offer theoretical insights into SEGNO, highlighting that it can learn a unique trajectory between adjacent states, which is crucial for model generalization. Additionally, we prove that the discrepancy between this learned trajectory of SEGNO and the true trajectory is bounded. Extensive experiments on complex dynamical systems including molecular dynamics and motion capture demonstrate that our model yields a significant improvement over the state-of-the-art baselines.
****
Dipendra Misra,Akanksha Saran,Tengyang Xie,Alex Lamb,John Langford
Towards Principled Representation Learning from Videos for Reinforcement Learning
We study pre-training representations for decision-making using video data, which is abundantly available for tasks such as game agents and software testing. Even though significant empirical advances have been made on this problem, a theoretical understanding remains absent. We initiate the theoretical investigation into principled approaches for representation learning and focus on learning the latent state representations of the underlying MDP using video data. We study two types of settings: one where there is iid noise in the observation, and a more challenging setting where there is also the presence of exogenous noise, which is non-iid noise that is temporally correlated, such as the motion of people or cars in the background. We study three commonly used approaches: autoencoding, temporal contrastive learning, and forward modeling. We prove upper bounds for temporal contrastive learning and forward modeling in the presence of only iid noise. We show that these approaches can learn the latent state and use it to do efficient downstream RL with polynomial sample complexity. When exogenous noise is also present, we establish a lower bound result showing that the sample complexity of learning from video data can be exponentially worse than learning from action-labeled trajectory data. This partially explains why reinforcement learning with video pre-training is hard. We evaluate these representational learning methods in two visual domains, yielding results that are consistent with our theoretical findings.
****
Shuhai Zhang,Yiliao Song,Jiahao Yang,Yuanqing Li,Bo Han,Mingkui Tan
Detecting Machine-Generated Texts by Multi-Population Aware Optimization for Maximum Mean Discrepancy
Large language models (LLMs) such as ChatGPT have exhibited remarkable performance in generating human-like texts. However, machine-generated texts (MGTs) may carry critical risks, such as plagiarism issues and hallucination information. Therefore, it is very urgent and important to detect MGTs in many situations. Unfortunately, it is challenging to distinguish MGTs and human-written texts because the distributional discrepancy between them is often very subtle due to the remarkable performance of LLMS. In this paper, we seek to exploit \textit{maximum m

ean discrepancy} (MMD) to address this issue in the sense that MMD can well iden
tify distributional discrepancies. However, directly training a detector with M
MD using diverse MGTs will incur a significantly increased variance of MMD since
 MGTs may contain \textit{multiple text populations} due to various LLMs. This w
ill severely impair MMD's ability to measure the difference between two samples.
 To tackle this, we propose a novel \textit{multi-population} aware optimization
 method for MMD called MMD-MP, which can \textit{avoid variance increases} and t
hus improve the stability to measure the distributional discrepancy. Relying on
MMD-MP, we develop two methods for paragraph-based and sentence-based detection,
 respectively. Extensive experiments on various LLMs, \eg, GPT2 and ChatGPT, sh
ow superior detection performance of our MMD-MP.
****
Haiming Wang,Huajian Xin,Chuanyang Zheng,Zhengying Liu,Qingxing Cao,Yinya Huang,
Jing Xiong,Han Shi,Enze Xie,Jian Yin,Zhenguo Li,Xiaodan Liang
LEGO-Prover: Neural Theorem Proving with Growing Libraries
Despite the success of large language models (LLMs), the task of theorem proving
 still remains one of the hardest reasoning tasks that is far from being fully s
olved. Prior methods using language models have demonstrated promising results,
but they still struggle to prove even middle school level theorems. One common l
imitation of these methods is that they assume a fixed theorem library during th
e whole theorem proving process. However, as we all know, creating new useful th
eorems or even new theories is not only helpful but crucial and necessary for ad
vancing mathematics and proving harder and deeper results. In this work, we pres
ent LEGO-Prover, which employs a growing skill library containing verified lemma
s as skills to augment the capability of LLMs used in theorem proving. By constr
ucting the proof modularly, LEGO-Prover enables LLMs to utilize existing skills
retrieved from the library and to create new skills during the proving process.
These skills are further evolved (by prompting an LLM) to enrich the library on
another scale. Modular and reusable skills are constantly added to the library t
o enable tackling increasingly intricate mathematical problems. Moreover, the le
arned library further bridges the gap between human proofs and formal proofs by
making it easier to impute missing steps. LEGO-Prover advances the state-of-the-
art pass rate on miniF2F-valid (48.0\% to 57.0\%) and miniF2F-test (45.5\% to 50
.0\%). During the proving process, LEGO-Prover also generates over 20,000 skills
 (theorems/lemmas) and adds them to the growing library. Our ablation study indi
cates that these newly added skills are indeed helpful for proving theorems, res
ulting in a 4.9\% improvement in success rate
****
Yifan Jiang,Hao Tang,Jen-Hao Rick Chang,Liangchen Song,Zhangyang Wang,Liangliang
 Cao
Efficient-3Dim: Learning a Generalizable Single-image Novel-view Synthesizer in
One Day
The task of novel view synthesis aims to generate unseen perspectives of an obje
ct or scene from a limited set of input images. Nevertheless, synthesizing novel
 views from a single image remains a significant challenge. Previous approaches
tackle this problem by adopting mesh prediction, multi-plane image construction,
 or more advanced techniques such as neural radiance fields. Recently, a pre-tra
ined diffusion model that is specifically designed for 2D image synthesis has de
monstrated its capability in producing photorealistic novel views, if sufficient
ly optimized with a 3D finetuning task. Despite greatly improved fidelity and ge
neralizability, training such a powerful diffusion model requires a vast volume
of training data and model parameters, resulting in a notoriously long time and
high computational costs. To tackle this issue, we propose Efficient-3DiM, a hig
hly efficient yet effective framework to learn a single-image novel-view synthes
izer. Motivated by our in-depth analysis of the diffusion model inference proce
ss, we propose several pragmatic strategies to reduce training overhead to a man
ageable scale, including a crafted timestep sampling strategy, a superior 3D fea
ture extractor, and an enhanced training scheme. When combined, our framework ca
n reduce the total training time from 10 days to less than 1 day, significantly
accelerating the training process on the same computational platform (an instanc

e with 8 Nvidia A100 GPUs). Comprehensive experiments are conducted to demonstrate the efficiency and generalizability of our proposed method.
****
Fan Wu,Huseyin A Inan,Arturs Backurs,Varun Chandrasekaran,Janardhan Kulkarni,Robert Sim
Privately Aligning Language Models with Reinforcement Learning
Positioned between pre-training and user deployment, aligning large language models (LLMs) through reinforcement learning (RL) has emerged as a prevailing strategy for training instruction following-models such as ChatGPT. In this work, we initiate the study of privacy-preserving alignment of LLMs through Differential Privacy (DP) in conjunction with RL. Following the influential work of Ziegler et al. (2020), we study two dominant paradigms: (i) alignment via RL without human in the loop (e.g., positive review generation) and (ii) alignment via RL from human feedback (RLHF) (e.g., summarization in a human-preferred way). We give a new DP framework to achieve alignment via RL, and prove its correctness. Our experimental results validate the effectiveness of our approach, offering competitive utility while ensuring strong privacy protections.
****
Sorawit Saengkyongam,Elan Rosenfeld,Pradeep Kumar Ravikumar,Niklas Pfister,Jonas Peters
Identifying Representations for Intervention Extrapolation
The premise of identifiable and causal representation learning is to improve the current representation learning paradigm in terms of generalizability or robustness. Despite recent progress in questions of identifiability, more theoretical results demonstrating concrete advantages of these methods for downstream tasks are needed. In this paper, we consider the task of intervention extrapolation: predicting how interventions affect an outcome, even when those interventions are not observed at training time, and show that identifiable representations can provide an effective solution to this task even if the interventions affect the outcome non-linearly. Our setup includes an outcome variable Y, observed features X, which are generated as a non-linear transformation of latent features Z, and exogenous action variables A, which influence Z. The objective of intervention extrapolation is then to predict how interventions on A that lie outside the training support of A affect Y. Here, extrapolation becomes possible if the effect of A on Z is linear and the residual when regressing Z on A has full support. As Z is latent, we combine the task of intervention extrapolation with identifiable representation learning, which we call $\texttt{Rep4Ex}$: we aim to map the observed features X into a subspace that allows for non-linear extrapolation in A. We show that the hidden representation is identifiable up to an affine transformation in Z-space, which, we prove, is sufficient for intervention extrapolation. The identifiability is characterized by a novel constraint describing the linearity assumption of A on Z. Based on this insight, we propose a flexible method that enforces the linear invariance constraint and can be combined with any type of autoencoder. We validate our theoretical findings through a series of synthetic experiments and show that our approach can indeed succeed in predicting the effects of unseen interventions.
****
Huaixiu Steven Zheng,Swaroop Mishra,Xinyun Chen,Heng-Tze Cheng,Ed H. Chi,Quoc V Le,Denny Zhou
Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models
We present STEP-BACK PROMPTING, a simple prompting technique that enables LLMs to do abstractions to derive high-level concepts and first principles from instances containing specific details. Using the concepts and principles to guide reasoning, LLMs significantly improve their abilities in following a correct reasoning path towards the solution. We conduct experiments of STEP-BACK PROMPTING with PaLM-2L, GPT-4 and Llama2-70B models, and observe substantial performance gains on various challenging reasoning-intensive tasks including STEM, Knowledge QA, and Multi-Hop Reasoning. For instance, STEP-BACK PROMPTING improves PaLM-2L performance on MMLU (Physics and Chemistry) by 7% and 11% respectively, TimeQA by 27%, and MuSiQue by 7%.

****
Christina Baek,J Zico Kolter,Aditi Raghunathan
Why is SAM Robust to Label Noise?
Sharpness-Aware Minimization (SAM) is most known for achieving state-of the-art performances on natural image and language tasks. However, its most pronounced improvements (of tens of percent) is rather in the presence of label noise. Understanding SAM's label noise robustness requires a departure from characterizing the robustness of minimas lying in ``flatter'' regions of the loss landscape. In particular, the peak performance occurs with early stopping, far before the loss converges. We decompose SAM's robustness into two effects: one induced by changes to the logit term and the other induced by changes to the network Jacobian. The first can be observed in linear logistic regression where SAM provably upweights the gradient contribution from clean examples. Although this explicit upweighting is also observable in neural networks, when we intervene and modify SAM to remove this effect, surprisingly, we see no visible degradation in performance. We infer that SAM's effect in deeper networks is instead explained entirely by the effect SAM has on the network Jacobian. We theoretically derive the explicit regularization induced by this Jacobian effect in two layer linear networks. Motivated by our analysis, we see that cheaper alternatives to SAM that explicitly induce these regularization effects largely recover the benefits even in deep networks trained on real-world datasets.
****
Tim Ruben Davidson,Veniamin Veselovsky,Michal Kosinski,Robert West
Evaluating Language Model Agency Through Negotiations
We introduce an approach to evaluate language model (LM) agency using negotiation games. This approach better reflects real-world use cases and addresses some of the shortcomings of alternative LM benchmarks. Negotiation games enable us to study multi-turn, and cross-model interactions, modulate complexity, and side-step accidental evaluation data leakage. We use our approach to test six widely used and publicly accessible LMs, evaluating performance and alignment in both self-play and cross-play settings. Noteworthy findings include: (i) only closed-source models tested here were able to complete these tasks; (ii) cooperative bargaining games proved to be most challenging to the models; and (iii) even the most powerful models sometimes "lose" to weaker opponents.
****
Dawei Zhu,Nan Yang,Liang Wang,Yifan Song,Wenhao Wu,Furu Wei,Sujian Li
PoSE: Efficient Context Window Extension of LLMs via Positional Skip-wise Training
Large Language Models (LLMs) are trained with a pre-defined context length, restricting their use in scenarios requiring long inputs. Previous efforts for adapting LLMs to a longer length usually requires fine-tuning with this target length (Full-length fine-tuning), suffering intensive training cost. To decouple train length from target length for efficient context window extension, we propose Positional Skip-wisE (PoSE) training that smartly simulates long inputs using a fixed context window. This is achieved by first dividing the original context window into several chunks, then designing distinct skipping bias terms to manipulate the position indices of each chunk. These bias terms and the lengths of each chunk are altered for every training example, allowing the model to adapt to all positions within target length. Experimental results show that PoSE greatly reduces memory and time overhead compared with Full-length fine-tuning, with minimal impact on performance. Leveraging this advantage, we have successfully extended the LLaMA model to 128k tokens using a 2k training context window. Furthermore, we empirically confirm that PoSE is compatible with all RoPE-based LLMs and position interpolation strategies. Notably, our method can potentially support infinite length, limited only by memory usage in inference. With ongoing progress for efficient inference, we believe PoSE can further scale the context window beyond 128k.
****
Donggyu Lee,Sangwon Jung,Taesup Moon
Continual Learning in the Presence of Spurious Correlations: Analyses and a Simp

le Baseline
Most continual learning (CL) algorithms have focused on tackling the stability-plasticity dilemma, that is, the challenge of preventing the forgetting of past tasks while learning new ones. However, we argue that they have overlooked the impact of knowledge transfer when the training dataset of a certain task is biased — namely, when the dataset contains some spurious correlations that can overly influence the prediction rule of a model. In that case, how would the dataset bias of a certain task affect prediction rules of a CL model for the future or past tasks? In this work, we carefully design systematic experiments using three benchmark datasets to answer the question from our empirical findings. Specifically, we first show through two-task CL experiments that standard CL methods, which are oblivious of the dataset bias, can transfer bias from one task to another, both forward and backward. Moreover, we find out this transfer is exacerbated depending on whether the CL methods focus on stability or plasticity. We then present that the bias is also transferred and even accumulates in longer task sequences. Finally, we offer a standardized experiment setup and a simple, yet strong plug-in baseline method, dubbed as Group-class Balanced Greedy Sampling (BGS). These resources can be utilized for the development of more advanced bias-aware CL methods.
****
Yite Wang,Jiahao Su,Hanlin Lu,Cong Xie,Tianyi Liu,Jianbo Yuan,Haibin Lin,Ruoyu Sun,Hongxia Yang
LEMON: Lossless model expansion
Scaling of deep neural networks, especially Transformers, is pivotal for their surging performance and has further led to the emergence of sophisticated reasoning capabilities in foundation models.
Such scaling generally requires training large models from scratch with random initialization, failing to leverage the knowledge acquired by their smaller counterparts, which are already resource-intensive to obtain.
To tackle this inefficiency, we present $\textbf{L}$ossl$\textbf{E}$ss $\textbf{MO}$del Expansio$\textbf{N}$ (LEMON), a recipe
to initialize scaled models using the weights of their smaller but pre-trained counterparts. This is followed by model training with an optimized learning rate scheduler tailored explicitly for the scaled models, substantially reducing the training time compared to training from scratch.
Notably, LEMON is versatile, ensuring compatibility with various network structures, including models like Vision Transformers and BERT.
Our empirical results demonstrate that LEMON reduces computational costs by 56.7\% for Vision Transformers and 33.2\% for BERT when compared to training from scratch.
****
Zhiyuan Cheng,Hongjun Choi,Shiwei Feng,James Chenhao Liang,Guanhong Tao,Dongfang Liu,Michael Zuzak,Xiangyu Zhang
Fusion Is Not Enough: Single Modal Attacks on Fusion Models for 3D Object Detection
Multi-sensor fusion (MSF) is widely used in autonomous vehicles (AVs) for perception, particularly for 3D object detection with camera and LiDAR sensors. The purpose of fusion is to capitalize on the advantages of each modality while minimizing its weaknesses. Advanced deep neural network (DNN)-based fusion techniques have demonstrated the exceptional and industry-leading performance. Due to the redundant information in multiple modalities, MSF is also recognized as a general defence strategy against adversarial attacks.
In this paper, we attack fusion models from the camera modality that is considered to be of lesser importance in fusion but is more affordable for attackers. We argue that the weakest link of fusion models depends on their most vulnerable modality and propose an attack framework that targets advanced camera-LiDAR fusion-based 3D object detection models through camera-only adversarial attacks.
Our approach employs a two-stage optimization-based strategy that first thoroughly evaluates vulnerable image areas under adversarial attacks, and then applies dedicated attack strategies for different fusion models to generate deployable p

atches. The evaluations with six advanced camera-LiDAR fusion models and one cam
era-only model indicate that our attacks successfully compromise all of them. Ou
r approach can either decrease the mean average precision (mAP) of detection per
formance from 0.824 to 0.353 or degrade the detection score of a target object f
rom 0.728 to 0.156, demonstrating the efficacy of our proposed attack framework.
 Code is available.
****
Weiyu Liu,Geng Chen,Joy Hsu,Jiayuan Mao,Jiajun Wu
Learning Planning Abstractions from Language
This paper presents a framework for learning state and action abstractions in se
quential decision-making domains. Our framework, planning abstraction from langu
age (PARL), utilizes language-annotated demonstrations to automatically discover
 a symbolic and abstract action space and induce a latent state abstraction base
d on it. PARL consists of three stages: 1) recovering object-level and action co
ncepts, 2) learning state abstractions, abstract action feasibility, and transit
ion models, and 3) applying low-level policies for abstract actions. During infe
rence, given the task description, PARL first makes abstract action plans using
the latent transition and feasibility functions, then refines the high-level pla
n using low-level policies. PARL generalizes across scenarios involving novel ob
ject instances and environments, unseen concept compositions, and tasks that req
uire longer planning horizons than settings it is trained on.
****
Qingqing Cao,Sewon Min,Yizhong Wang,Hannaneh Hajishirzi
BTR: Binary Token Representations for Efficient Retrieval Augmented Language Mod
els
Retrieval augmentation addresses many critical problems in large language models
 such as hallucination, staleness, and privacy leaks.
However, running retrieval-augmented language models (LMs) is slow and difficult
 to scale due to processing large amounts of retrieved text.
We introduce binary token representations (BTR), which use 1-bit vectors to prec
ompute every token in passages, significantly reducing computation during infere
nce.
Despite the potential loss of accuracy, our new calibration techniques and train
ing objectives restore performance. Combined with offline and runtime compressio
n, this only requires 127GB of disk space for encoding 3 billion tokens in Wikip
edia.
Our experiments show that on five knowledge-intensive NLP tasks, BTR accelerates
 state-of-the-art inference by up to 4x and reduces storage by over 100x while m
aintaining over 95% task performance. Our code is publicly available at https://
github.com/csarron/BTR.
****
Maksim Velikanov,Maxim Panov,Dmitry Yarotsky
Generalization error of spectral algorithms
The asymptotically precise estimation of the generalization of kernel methods ha
s recently received attention due to the parallels between neural networks and t
heir associated kernels. However, prior works derive such estimates for training
 by kernel ridge regression (KRR), whereas neural networks are typically trained
 with gradient descent (GD). In the present work, we consider the training of ke
rnels with a family of \emph{spectral algorithms} specified by profile $h(\lambd
a)$, and including KRR and GD as special cases. Then, we derive the generalizati
on error as a functional of learning profile $h(\lambda)$ for two data models: h
igh-dimensional Gaussian and low-dimensional translation-invariant model.
Under power-law assumptions on the spectrum of the kernel and target, we use our
 framework to (i) give full loss asymptotics for both noisy and noiseless observ
ations (ii) show that the loss localizes on certain spectral scales, giving a ne
w perspective on the KRR saturation phenomenon (iii) conjecture, and demonstrate
 for the considered data models, the universality of the loss w.r.t. non-spectra
l details of the problem, but only in case of noisy observation.
****
Noa Rubin,Inbar Seroussi,Zohar Ringel

Grokking as a First Order Phase Transition in Two Layer Networks
A key property of deep neural networks (DNNs) is their ability to learn new features during training. This intriguing aspect of deep learning stands out most clearly in recently reported Grokking phenomena. While mainly reflected as a sudden increase in test accuracy, Grokking is also believed to be a beyond lazy-learning/Gaussian Process (GP) phenomenon involving feature learning. Here we apply a recent development in the theory of feature learning, the adaptive kernel approach, to two teacher-student models with cubic-polynomial and modular addition teachers. We provide analytical predictions on feature learning and Grokking properties of these models and demonstrate a mapping between Grokking and the theory of phase transitions. We show that after Grokking, the state of the DNN is analogous to the mixed phase following a first-order phase transition. In this mixed phase, the DNN generates useful internal representations of the teacher that are sharply distinct from those before the transition.
****
Benjamin Schneider,Nils Lukas,Florian Kerschbaum
Universal Backdoor Attacks
Web-scraped datasets are vulnerable to data poisoning, which can be used for backdooring deep image classifiers during training. Since training on large datasets is expensive, a model is trained once and reused many times. Unlike adversarial examples, backdoor attacks often target specific classes rather than any class learned by the model. One might expect that targeting many classes through a naïve composition of attacks vastly increases the number of poison samples. We show this is not necessarily true and more efficient,
universal data poisoning attacks exist that allow controlling misclassifications from any source class into any target class with a slight increase in poison samples. Our idea is to generate triggers with salient characteristics that the model can learn. The triggers we craft exploit a phenomenon we call _inter-class poison transferability_, where learning a trigger from one class makes the model more vulnerable to learning triggers for other classes. We demonstrate the effectiveness and robustness of our universal backdoor attacks by controlling models with up to 6,000 classes while poisoning only 0.15% of the training dataset.
****
Hongjun Wang,Sagar Vaze,Kai Han
SPTNet: An Efficient Alternative Framework for Generalized Category Discovery with Spatial Prompt Tuning
Generalized Category Discovery (GCD) aims to classify unlabelled images from both ‘seen’ and ‘unseen’ classes by transferring knowledge from a set of labelled ‘seen’ class images. A key theme in existing GCD approaches is adapting large-scale pre-trained models for the GCD task. An alternate perspective, however, is to adapt the data representation itself for better alignment with the pre-trained model. As such, in this paper, we introduce a two-stage adaptation approach termed SPTNet, which iteratively optimizes model parameters (i.e., model-finetuning) and data parameters (i.e., prompt learning). Furthermore, we propose a novel spatial prompt tuning method (SPT) which considers the spatial property of image data, enabling the method to better focus on object parts, which can transfer between seen and unseen classes. We thoroughly evaluate our SPTNet on standard benchmarks and demonstrate that our method outperforms existing GCD methods. Notably, we find our method achieves an average accuracy of 61.4% on the SSB, surpassing prior state-of-the-art methods by approximately 10%. The improvement is particularly remarkable as our method yields extra parameters amounting to only 0.117% of those in the backbone architecture. Project page: https://visual-ai.github.io/sptnet.
****
Dante Everaert,Christopher Potts
GIO: Gradient Information Optimization for Training Dataset Selection
It is often advantageous to train models on a subset of the available train examples, because the examples are of variable quality or because one would like to train with fewer examples, without sacrificing performance. We present Gradient Information Optimization (GIO), a scalable, task-agnostic approach to this data

selection problem that requires only a small set of (unlabeled) examples representing a target distribution. GIO begins from a natural, information-theoretic objective that is intractable in practice. Our contribution is in showing that it can be made highly scalable through a simple relaxation of the objective and a highly efficient implementation. In experiments with machine translation, spelling correction, and image recognition, we show that GIO delivers outstanding results with very small train sets. These findings are robust to different representation models and hyperparameters for GIO itself. GIO is task- and domain-agnostic and can be applied out-of-the-box to new datasets and domains. We open source a pip-installable implementation of the algorithm as "pip install grad-info-opt".
****
Soobin Um,Suhyeon Lee,Jong Chul Ye
Don't Play Favorites: Minority Guidance for Diffusion Models
We explore the problem of generating minority samples using diffusion models. The minority samples are instances that lie on low-density regions of a data manifold. Generating a sufficient number of such minority instances is important, since they often contain some unique attributes of the data. However, the conventional generation process of the diffusion models mostly yields majority samples (that lie on high-density regions of the manifold) due to their high likelihoods, making themselves ineffective and time-consuming for the minority generating task. In this work, we present a novel framework that can make the generation process of the diffusion models focus on the minority samples. We first highlight that Tweedie's denoising formula yields favorable results for majority samples. The observation motivates us to introduce a metric that describes the uniqueness of a given sample. To address the inherent preference of the diffusion models w.r.t. the majority samples, we further develop *minority guidance*, a sampling technique that can guide the generation process toward regions with desired likelihood levels. Experiments on benchmark real datasets demonstrate that our minority guidance can greatly improve the capability of generating high-quality minority samples over existing generative samplers. We showcase that the performance benefit of our framework persists even in demanding real-world scenarios such as medical imaging, further underscoring the practical significance of our work. Code is available at https://github.com/soobin-um/minority-guidance.
****
Tim Lebailly,Thomas Stegmüller,Behzad Bozorgtabar,Jean-Philippe Thiran,Tinne Tuytelaars
CrIBo: Self-Supervised Learning via Cross-Image Object-Level Bootstrapping
Leveraging nearest neighbor retrieval for self-supervised representation learning has proven beneficial with object-centric images. However, this approach faces limitations when applied to scene-centric datasets, where multiple objects within an image are only implicitly captured in the global representation. Such global bootstrapping can lead to undesirable entanglement of object representations. Furthermore, even object-centric datasets stand to benefit from a finer-grained bootstrapping approach. In response to these challenges, we introduce a novel $\textbf{Cr}$oss-$\textbf{I}$mage Object-Level $\textbf{Bo}$otstrapping method tailored to enhance dense visual representation learning. By employing object-level nearest neighbor bootstrapping throughout the training, CrIBo emerges as a notably strong and adequate candidate for in-context learning, leveraging nearest neighbor retrieval at test time. CrIBo shows state-of-the-art performance on the latter task while being highly competitive in more standard downstream segmentation tasks. Our code and pretrained models are publicly available at https://github.com/tileb1/CrIBo.
****
Tingchen Fu,Lemao Liu,Deng Cai,Guoping Huang,Shuming Shi,Rui Yan
The Reasonableness Behind Unreasonable Translation Capability of Large Language Model
Multilingual large language models trained on non-parallel data yield impressive translation capabilities. Existing studies demonstrate that incidental sentence-level bilingualism within pre-training data contributes to the LLM's translation abilities. However, it has also been observed that LLM's translation capabilit

ies persist even when incidental sentence-level bilingualism are excluded from the training corpus.
In this study, we comprehensively investigate the unreasonable effectiveness and the underlying mechanism for LLM's translation abilities, specifically addressing the question why large language models learn to translate without parallel data, using the BLOOM model series as a representative example. Through extensive experiments, our findings suggest the existence of unintentional bilingualism in the pre-training corpus, especially word alignment data significantly contributes to the large language model's acquisition of translation ability. Moreover, the translation signal derived from word alignment data is comparable to that from sentence-level bilingualism. Additionally, we study the effects of monolingual data and parameter-sharing in assisting large language model to learn to translate. Together, these findings present another piece of the broader puzzle of trying to understand how large language models acquire translation capability.
****
Hanlin Zhu,Baihe Huang,Stuart Russell
On Representation Complexity of Model-based and Model-free Reinforcement Learning
We study the representation complexity of model-based and model-free reinforcement learning (RL) in the context of circuit complexity. We prove theoretically that there exists a broad class of MDPs such that their underlying transition and reward functions can be represented by constant depth circuits with polynomial size, while the optimal Q-function suffers an exponential circuit complexity in constant-depth circuits. By drawing attention to the approximation errors and building connections to complexity theory, our theory provides unique insights into why model-based algorithms usually enjoy better sample complexity than model-free algorithms from a novel representation complexity perspective: in some cases, the ground-truth rule (model) of the environment is simple to represent, while other quantities, such as Q-function, appear complex. We empirically corroborate our theory by comparing the approximation error of the transition kernel, reward function, and optimal Q-function in various Mujoco environments, which demonstrates that the approximation errors of the transition kernel and reward function are consistently lower than those of the optimal Q-function. To the best of our knowledge, this work is the first to study the circuit complexity of RL, which also provides a rigorous framework for future research.
****
Runzhe Wang,Sadhika Malladi,Tianhao Wang,Kaifeng Lyu,Zhiyuan Li
The Marginal Value of Momentum for Small Learning Rate SGD
Momentum is known to accelerate the convergence of gradient descent in strongly convex settings without stochastic gradient noise. In stochastic optimization, such as training neural networks, folklore suggests that momentum may help deep learning optimization by reducing the variance of the stochastic gradient update, but previous theoretical analyses do not find momentum to offer any provable acceleration. Theoretical results in this paper clarify the role of momentum in stochastic settings where the learning rate is small and gradient noise is the dominant source of instability, suggesting that SGD with and without momentum behave similarly in the short and long time horizons. Experiments show that momentum indeed has limited benefits for both optimization and generalization in practical training regimes where the optimal learning rate is not very large, including small- to medium-batch training from scratch on ImageNet and fine-tuning language models on downstream tasks.
****
Feng Hong,Jiangchao Yao,Yueming Lyu,Zhihan Zhou,Ivor Tsang,Ya Zhang,Yanfeng Wang
On Harmonizing Implicit Subpopulations
Machine learning algorithms learned from data with skewed distributions usually suffer from poor generalization, especially when minority classes matter as much as, or even more than majority ones. This is more challenging on class-balanced data that has some hidden imbalanced subpopulations, since prevalent techniques mainly conduct class-level calibration and cannot perform subpopulation-level adjustments without subpopulation annotations. Regarding implicit subpopulation i

mbalance, we reveal that the key to alleviating the detrimental effect lies in effective subpopulation discovery with proper rebalancing. We then propose a novel subpopulation-imbalanced learning method called Scatter and HarmonizE (SHE). Our method is built upon the guiding principle of optimal data partition, which involves assigning data to subpopulations in a manner that maximizes the predictive information from inputs to labels. With theoretical guarantees and empirical evidences, SHE succeeds in identifying the hidden subpopulations and encourages subpopulation-balanced predictions. Extensive experiments on various benchmark datasets show the effectiveness of SHE.
****
Yujie Mo,Feiping Nie,Ping Hu,Heng Tao Shen,Zheng Zhang,Xinchao Wang,Xiaofeng Zhu
Self-Supervised Heterogeneous Graph Learning: a Homophily and Heterogeneity View
Self-supervised heterogeneous graph learning has achieved promising results in various real applications, but it still suffers from the following issues: (i) meta-paths can be employed to capture the homophily in the heterogeneous graph, but meta-paths are human-defined, requiring substantial expert knowledge and computational costs; and (ii) the heterogeneity in the heterogeneous graph is usually underutilized, leading to the loss of task-related information. To solve these issues, this paper proposes to capture both homophily and heterogeneity in the heterogeneous graph without pre-defined meta-paths. Specifically, we propose to learn a self-expressive matrix to capture the homophily from the subspace and nearby neighbors. Meanwhile, we propose to capture the heterogeneity by aggregating the information of nodes from different types. We further design a consistency loss and a specificity loss, respectively, to extract the consistent information between homophily and heterogeneity and to preserve their specific task-related information. We theoretically analyze that the learned homophilous representations exhibit the grouping effect to capture the homophily, and considering both homophily and heterogeneity introduces more task-related information. Extensive experimental results verify the superiority of the proposed method on different downstream tasks.
****
Zhiyuan Li,Hong Liu,Denny Zhou,Tengyu Ma
Chain of Thought Empowers Transformers to Solve Inherently Serial Problems
Generating a sequence of intermediate steps, \emph{a.k.a.}, a chain of thought (CoT), is a highly effective method to improve the accuracy of large language models (LLMs) on arithmetics and symbolic reasoning tasks. However, the mechanism behind CoT remains unclear.
This work provides a theoretical understanding of the power of CoT for decoder-only transformers through the lens of expressiveness. Conceptually, CoT empowers the model with the ability to perform inherently serial computation, which is otherwise lacking in transformers, especially when depth is low. Given input length n, previous works have constant-depth transformers with finite precision $\mathsf{poly}(n)$ embedding size can only solve problems in TC^0 without CoT. We first show an even tighter expressiveness upper bound for constant-depth transformers with constant-bit precision, which can only solve problems in AC^0, a proper subset of $ \mathsf{TC}^0$. However, with T steps of CoT, constant-depth transformers using constant-bit precision and $O(\log n)$ embedding size can solve any problem solvable by boolean circuits of size T. Empirically, enabling CoT dramatically improves the accuracy for tasks that are hard for parallel computation, including the composition of permutation groups, iterated squaring, and circuit value problems, especially for low-depth transformers.
****
Yeongmin Kim,Byeonghu Na,Minsang Park,JoonHo Jang,Dongjun Kim,Wanmo Kang,Il-chul Moon
Training Unbiased Diffusion Models From Biased Dataset
With significant advancements in diffusion models, addressing the potential risks of dataset bias becomes increasingly important. Since generated outputs directly suffer from dataset bias, mitigating latent bias becomes a key factor in impr

oving sample quality and proportion. This paper proposes time-dependent importan
ce reweighting to mitigate the bias for the diffusion models. We demonstrate tha
t the time-dependent density ratio becomes more precise than previous approaches
, thereby minimizing error propagation in generative learning. While directly ap
plying it to score-matching is intractable, we discover that using the time-depe
ndent density ratio both for reweighting and score correction can lead to a trac
table form of the objective function to regenerate the unbiased data density. Fu
rthermore, we theoretically establish a connection with traditional score-matchi
ng, and we demonstrate its convergence to an unbiased distribution. The experime
ntal evidence supports the usefulness of the proposed method, which outperforms
baselines including time-independent importance reweighting on CIFAR-10, CIFAR-1
00, FFHQ, and CelebA with various bias settings. Our code is available at https:
//github.com/alsdudrla10/TIW-DSM.
****

Jing-Cheng Pang,Pengyuan Wang,Kaiyuan Li,Xiong-Hui Chen,Jiacheng Xu,Zongzhang Zh
ang,Yang Yu
Language Model Self-improvement by Reinforcement Learning Contemplation
Language model self-improvement (LMSI) techniques have recently gained significa
nt attention as they improve language models without requiring external supervis
ion. A common approach is reinforcement learning from AI feedback (RLAIF), which
 trains a reward model based on AI preference data and employs a reinforcement l
earning algorithm to train the language model.
However, RLAIF relies on the heuristic assumption that an AI model can provide e
ffective feedback and correct wrong answers, requiring a solid capability of the
 language model. This paper presents a novel LMSI method, Reinforcement Learning
 Contemplation (RLC). We disclose that it is simpler for language models to eval
uate a sentence than to generate it, even for small language models. Leveraging
the gap between the evaluation and generation, RLC evaluates generated answers a
nd updates language model parameters using reinforcement learning to maximize ev
aluation scores. Through testing on various challenging reasoning tasks and text
 summarization task, our experiments show that RLC effectively improves language
 model performance without external supervision, resulting in an answering accur
acy increase (from 31.23% to 37.09%) for BigBench-hard reasoning tasks, and a ri
se in BERTScore for CNN/Daily Mail summarization tasks. Furthermore, RLC can be
applied to models of different sizes, showcasing its broad applicability.
****

Iosif Sakos,Stefanos Leonardos,Stelios Andrew Stavroulakis,William Overman,Ioann
is Panageas,Georgios Piliouras
Beating Price of Anarchy and Gradient Descent without Regret in Potential Games
Arguably one of the thorniest problems in game theory is that of equilibrium sel
ection. Specifically, in the presence of multiple equilibria do self-interested
learning dynamics typically select the socially optimal ones? We study a rich cl
ass of continuous-time no-regret dynamics in potential games (PGs). Our class of
 dynamics, *Q-Replicator Dynamics* (QRD), include gradient descent (GD), log-bar
rier and replicator dynamics (RD) as special cases. We start by establishing *po
intwise convergence* of all QRD to Nash equilibria in almost all PGs. In the cas
e of GD, we show a tight average case performance within a factor of two of opti
mal, for a class of symmetric 2×2 potential games with unbounded Price of
 Anarchy (PoA). Despite this positive result, we show that GD is not always the
optimal choice even in this restricted setting. Specifically, GD outperforms RD,
 if and only if *risk-* and *payoff-dominance* equilibria coincide. Finally, we
experimentally show how these insights extend to all QRD dynamics and that unbou
nded gaps between average case performance and PoA analysis are common even in l
arger settings.
****

Peihao Wang,Shenghao Yang,Shu Li,Zhangyang Wang,Pan Li
Polynomial Width is Sufficient for Set Representation with High-dimensional Feat
ures
Set representation has become ubiquitous in deep learning for modeling the induc
tive bias of neural networks that are insensitive to the input order. DeepSets i

s the most widely used neural network architecture for set representation. It involves embedding each set element into a latent space with dimension L, followed by a sum pooling to obtain a whole-set embedding, and finally mapping the whole-set embedding to the output. In this work, we investigate the impact of the dimension L on the expressive power of DeepSets. Previous analyses either oversimplified high-dimensional features to be one-dimensional features or were limited to complex analytic activations, thereby diverging from practical use or resulting in L that grows exponentially with the set size N and feature dimension D. To investigate the minimal value of L that achieves sufficient expressive power, we present two set-element embedding layers: (a) linear + power activation (LP) and (b) linear + exponential activations (LE). We demonstrate that L being $\operatorname{poly}(N, D)$ is sufficient for set representation using both embedding layers. We also provide a lower bound of L for the LP embedding layer. Furthermore, we extend our results to permutation-equivariant set functions and the complex field.
****
Anastasios Nikolas Angelopoulos,Stephen Bates,Adam Fisch,Lihua Lei,Tal Schuster
Conformal Risk Control
We extend conformal prediction to control the expected value of any monotone loss function. The algorithm generalizes split conformal prediction together with its coverage guarantee. Like conformal prediction, the conformal risk control procedure is tight up to an $\mathcal{O}(1/n)$ factor. We also introduce extensions of the idea to distribution shift, quantile risk control, multiple and adversarial risk control, and expectations of U-statistics. Worked examples from computer vision and natural language processing demonstrate the usage of our algorithm to bound the false negative rate, graph distance, and token-level F1-score.
****
Dogyun Park,Sihyeon Kim,Sojin Lee,Hyunwoo J. Kim
DDMI: Domain-agnostic Latent Diffusion Models for Synthesizing High-Quality Implicit Neural Representations
Recent studies have introduced a new class of generative models for synthesizing implicit neural representations (INRs) that capture arbitrary continuous signals in various domains. These models opened the door for domain-agnostic generative models, but they often fail to achieve high-quality generation. We observed that the existing methods generate the weights of neural networks to parameterize INRs and evaluate the network with fixed positional embeddings (PEs). Arguably, this architecture limits the expressive power of generative models and results in low-quality INR generation. To address this limitation, we propose Domain-agnostic Latent Diffusion Model for INRs (DDMI) that generates adaptive positional embeddings instead of neural networks' weights. Specifically, we develop a Discrete-to-continuous space Variational AutoEncoder (D2C-VAE) that seamlessly connects discrete data and continuous signal functions in the shared latent space. Additionally, we introduce a novel conditioning mechanism for evaluating INRs with the hierarchically decomposed PEs to further enhance expressive power. Extensive experiments across four modalities, \eg, 2D images, 3D shapes, Neural Radiance Fields, and videos, with seven benchmark datasets, demonstrate the versatility of DDMI and its superior performance compared to the existing INR generative models. Code is available at \href{https://github.com/mlvlab/DDMI}{https://github.com/mlvlab/DDMI}.
****
Isaac Reid,Eli Berger,Krzysztof Marcin Choromanski,Adrian Weller
Repelling Random Walks
We present a novel quasi-Monte Carlo mechanism to improve graph-based sampling, coined repelling random walks. By inducing correlations between the trajectories of an interacting ensemble such that their marginal transition probabilities are unmodified, we are able to explore the graph more efficiently, improving the concentration of statistical estimators whilst leaving them unbiased. The mechanism has a trivial drop-in implementation. We showcase the effectiveness of repelling random walks in a range of settings including estimation of graph kernels, the PageRank vector and graphlet concentrations. We provide detailed experimenta

l evaluation and robust theoretical guarantees. To our knowledge, repelling random walks constitute the first rigorously studied quasi-Monte Carlo scheme correlating the directions of walkers on a graph, inviting new research in this exciting nascent domain.
****
Xuan Son Nguyen,Shuo Yang,Aymeric Histace
Matrix Manifold Neural Networks++
Deep neural networks (DNNs) on Riemannian manifolds have garnered increasing interest in various applied areas. For instance, DNNs on spherical and hyperbolic manifolds have been designed to solve a wide range of computer vision and nature language processing tasks. One of the key factors that contribute to the success of these networks is that spherical and hyperbolic manifolds have the rich algebraic structures of gyrogroups and gyrovector spaces. This enables principled and effective generalizations of the most successful DNNs to these manifolds. Recently, some works have shown that many concepts in the theory of gyrogroups and gyrovector spaces can also be generalized to matrix manifolds such as Symmetric Positive Definite (SPD) and Grassmann manifolds. As a result, some building blocks for SPD and Grassmann neural networks, e.g., isometric models and multinomial logistic regression (MLR) can be derived in a way that is fully analogous to their spherical and hyperbolic counterparts. Building upon these works, in this paper, we design fully-connected (FC) and convolutional layers for SPD neural networks. We also develop MLR on Symmetric Positive Semi-definite (SPSD) manifolds, and propose a method for performing backpropagation with the Grassmann logarithmic map in the projector perspective. We demonstrate the effectiveness of the proposed approach in the human action recognition and node classification tasks.
****
Robin Louiset,Edouard Duchesnay,Antoine Grigis,Pietro Gori
Separating common from salient patterns with Contrastive Representation Learning
Contrastive Analysis is a sub-field of Representation Learning that aims at separating 1) salient factors of variation - that only exist in the target dataset (i.e., diseased subjects) in contrast with 2) common factors of variation between target and background (i.e., healthy subjects) datasets. Despite their relevance, current models based on Variational Auto-Encoders have shown poor performance in learning semantically-expressive representations. On the other hand, Contrastive Representation Learning has shown tremendous performance leaps in various applications (classification, clustering, etc.). In this work, we propose to leverage the ability of Contrastive Learning to learn semantically expressive representations when performing Contrastive Analysis. Namely, we reformulate Contrastive Analysis under the lens of the InfoMax Principle and identify two Mutual Information terms to maximize and one to minimize. We decompose the two first terms into an Alignment and a Uniformity term, as commonly done in Contrastive Learning. Then, we motivate a novel Mutual Information minimization strategy to prevent information leakage between common and salient distributions. We validate our method on datasets designed to assess the pattern separation capability in Contrastive Analysis, including MNIST superimposed on CIFAR10, CelebA accessories, dSprites item superimposed on a digit grid, and three medical datasets.
****
Ziqi Wang,Le Hou,Tianjian Lu,Yuexin Wu,Yunxuan Li,Hongkun Yu,Heng Ji
Enabling Lanuguage Models to Implicitly Learn Self-Improvement
Large Language Models (LLMs) have demonstrated remarkable capabilities in open-ended text generation tasks. However, the inherent open-ended nature of these tasks implies that there is always room for improvement in the quality of model responses. To address this challenge, various approaches have been proposed to enhance the performance of LLMs. There has been a growing focus on enabling LLMs to self-improve their response quality, thereby reducing the reliance on extensive human annotation efforts for collecting diverse and high-quality training data. Recently, prompting-based methods have been widely explored among self-improvement methods owing to their effectiveness, efficiency, and convenience. However, those methods usually require explicitly and thoroughly written rubrics as inputs to LLMs. It is expensive and challenging to manually derive and provide all nec

essary rubrics with a real-world complex goal for improvement (e.g., being more helpfulness and less harmful). To this end, we propose an imPlicit self-ImprovemenT (PIT) framework that implicitly learns the improvement goal from human preference data. PIT only requires preference data that are used to train reward models with no extra human efforts. Specifically, we reformulate the training objective of reinforcement learning from human feedback (RLHF) -- instead of maximizing response quality for a given input, we maximize the quality gap of the response conditioned on a reference response. In this way, PIT is implicitly trained with the improvement goal of better aligning with human preferences. Experiments on two real-world datasets and one synthetic dataset show that our method significantly outperforms prompting-based methods.
****
Ziteng Gao,Zhan Tong,Limin Wang,Mike Zheng Shou
SparseFormer: Sparse Visual Recognition via Limited Latent Tokens
Human visual recognition is a sparse process, where only a few salient visual cues are attended to rather than every detail being traversed uniformly. However, most current vision networks follow a dense paradigm, processing every single visual unit (such as pixels or patches) in a uniform manner. In this paper, we challenge this dense paradigm and present a new method, coined SparseFormer, to imitate human's sparse visual recognition in an end-to-end manner. SparseFormer learns to represent images using a highly limited number of tokens (as low as 49) in the latent space with sparse feature sampling procedure instead of processing dense units in the original image space. Therefore, SparseFormer circumvents most of dense operations on the image space and has much lower computational costs. Experiments on the ImageNet-1K classification show that SparseFormer achieves performance on par with canonical or well-established models while offering more favorable accuracy-throughput tradeoff. Moreover, the design of our network can be easily extended to the video classification with promising performance at lower computational costs. We hope that our work can provide an alternative way for visual modeling and inspire further research on sparse vision architectures.
****
Huigen Ye,Hua Xu,Hongyan Wang
Light-MILPopt: Solving Large-scale Mixed Integer Linear Programs with Lightweight Optimizer and Small-scale Training Dataset
Machine Learning (ML)-based optimization approaches emerge as a promising technique for solving large-scale Mixed Integer Linear Programs (MILPs). However, existing ML-based frameworks suffer from high model computation complexity, weak problem reduction, and reliance on large-scale optimizers and large training datasets, resulting in performance bottlenecks for large-scale MILPs. This paper proposes Light-MILPopt, a lightweight large-scale optimization framework that only uses a lightweight optimizer and small training dataset to solve large-scale MILPs. Specifically, Light-MILPopt can be divided into four stages: Problem Formulation for problem division to reduce model computational costs, Model-based Initial Solution Prediction for predicting and constructing the initial solution using a small-scale training dataset, Problem Reduction for both variable and constraint reduction, and Data-driven Optimization for current solution improvement employing a lightweight optimizer. Experimental evaluations on four large-scale benchmark MILPs and a real-world case study demonstrate that Light-MILPopt, leveraging a lightweight optimizer and small training dataset, outperforms the state-of-the-art ML-based optimization framework and advanced large-scale solvers (e.g. Gurobi, SCIP). The results and further analyses substantiate the ML-based framework's feasibility and effectiveness in solving large-scale MILPs.
****
Haoxuan You,Haotian Zhang,Zhe Gan,Xianzhi Du,Bowen Zhang,Zirui Wang,Liangliang Cao,Shih-Fu Chang,Yinfei Yang
Ferret: Refer and Ground Anything Anywhere at Any Granularity
We introduce Ferret, a new Multimodal Large Language Model (MLLM) capable of understanding spatial referring of any shape or granularity within an image and accurately grounding open-vocabulary descriptions. To unify referring and grounding in the LLM paradigm, Ferret employs a novel and powerful hybrid region represen

tation that integrates discrete coordinates and continuous features jointly to r epresent a region in the image. To extract the continuous features of versatile regions, we propose a spatial-aware visual sampler, adept at handling varying s parsity across different shapes. Consequently, Ferret can accept diverse region inputs, such as points, bounding boxes, and free-form shapes. To bolster the des ired capability of Ferret, we curate GRIT, a comprehensive refer-and-ground inst ruction tuning dataset including 1.1M samples that contain rich hierarchical spa tial knowledge, with an additional 130K hard negative data to promote model robu stness. The resulting model not only achieves superior performance in classical referring and grounding tasks, but also greatly outperforms existing MLLMs in re gion-based and localization-demanded multimodal chatting. Our evaluations also r eveal a significantly improved capability of describing image details and a rema rkable alleviation in object hallucination.
****
Jiahao Li,Hao Tan,Kai Zhang,Zexiang Xu,Fujun Luan,Yinghao Xu,Yicong Hong,Kalyan Sunkavalli,Greg Shakhnarovich,Sai Bi
Instant3D: Fast Text-to-3D with Sparse-view Generation and Large Reconstruction Model
Text-to-3D with diffusion models have achieved remarkable progress in recent yea rs. However, existing methods either rely on score distillation-based optimizati on which suffer from slow inference, low diversity and Janus problems, or are fe ed-forward methods that generate low quality results due to the scarcity of 3D t raining data. In this paper, we propose Instant3D, a novel method that generates high-quality and diverse 3D assets from text prompts in a feed-forward manner. We adopt a two-stage paradigm, which first generates a sparse set of four struct ured and consistent views from text in one shot with a fine-tuned 2D text-to-ima ge diffusion model, and then directly regresses the NeRF from the generated imag es with a novel transformer-based sparse-view reconstructor. Through extensive e xperiments, we demonstrate that our method can generate high-quality, diverse an d Janus-free 3D assets within 20 seconds, which is two order of magnitude faster than previous optimization-based methods that can take 1 to 10 hours. Our proje ct webpage: https://instant-3d.github.io/.
****
Ben Eisner,Yi Yang,Todor Davchev,Mel Vecerik,Jonathan Scholz,David Held
Deep SE(3)-Equivariant Geometric Reasoning for Precise Placement Tasks
Many robot manipulation tasks can be framed as geometric reasoning tasks, where an agent must be able to precisely manipulate an object into a position that sat isfies the task from a set of initial conditions. Often, task success is defined based on the relationship between two objects - for instance, hanging a mug on a rack. In such cases, the solution should be equivariant to the initial positi on of the objects as well as the agent, and invariant to the pose of the camera. This poses a challenge for learning systems which attempt to solve this task by learning directly from high-dimensional demonstrations: the agent must learn to be both equivariant as well as precise, which can be challenging without any in ductive biases about the problem. In this work, we propose a method for precise relative pose prediction which is provably SE(3)-equivariant, can be learned fro m only a few demonstrations, and can generalize across variations in a class of objects. We accomplish this by factoring the problem into learning an SE(3) inva riant task-specific representation of the scene and then interpreting this repre sentation with novel geometric reasoning layers which are provably SE(3) equivar iant. We demonstrate that our method can yield substantially more precise placem ent predictions in simulated placement tasks than previous methods trained with the same amount of data, and can accurately represent relative placement relatio nships data collected from real-world demonstrations. Supplementary information and videos can be found at https://sites.google.com/view/reldist-iclr-2023.
****
Suresh Bishnoi,Jayadeva Jayadeva,Sayan Ranu,N M Anoop Krishnan
BroGNet: Momentum-Conserving Graph Neural Stochastic Differential Equation for L earning Brownian Dynamics
Neural networks (NNs) that exploit strong inductive biases based on physical law

s and symmetries have shown remarkable success in learning the dynamics of physical systems directly from their trajectory. However, these works focus only on the systems that follow deterministic dynamics, such as Newtonian or Hamiltonian. Here, we propose a framework, namely Brownian graph neural networks (BroGNet), combining stochastic differential equations (SDEs) and GNNs to learn Brownian dynamics directly from the trajectory. We modify the architecture of BroGNet to enforce linear momentum conservation of the system, which, in turn, provides superior performance on learning dynamics as revealed empirically. We demonstrate this approach on several systems, namely, linear spring, linear spring with binary particle types, and non-linear spring systems, all following Brownian dynamics at finite temperatures. We show that BroGNet significantly outperforms proposed baselines across all the benchmarked Brownian systems. In addition, we demonstrate zero-shot generalizability of BroGNet to simulate unseen system sizes that are two orders of magnitude larger and to different temperatures than those used during training. Finally, we show that BroGNet conserves the momentum of the system resulting in superior performance and data efficiency. Altogether, our study contributes to advancing the understanding of the intricate dynamics of Brownian motion and demonstrates the effectiveness of graph neural networks in modeling such complex systems.
****
Timothée Darcet,Maxime Oquab,Julien Mairal,Piotr Bojanowski
Vision Transformers Need Registers
Transformers have recently emerged as a powerful tool for learning visual representations. In this paper, we identify and characterize artifacts in feature maps of both supervised and self-supervised ViT networks. The artifacts correspond to high-norm tokens appearing during inference primarily in low-informative background areas of images, that are repurposed for internal computations. We propose a simple yet effective solution based on providing additional tokens to the input sequence of the Vision Transformer to fill that role. We show that this solution fixes that problem entirely for both supervised and self-supervised models, sets a new state of the art for self-supervised visual models on dense visual prediction tasks, enables object discovery methods with larger models, and most importantly leads to smoother feature maps and attention maps for downstream visual processing.
****
Aidan Scannell,Riccardo Mereu,Paul Edmund Chang,Ella Tamir,Joni Pajarinen,Arno Solin
Function-space Parameterization of Neural Networks for Sequential Learning
Sequential learning paradigms pose challenges for gradient-based deep learning due to difficulties incorporating new data and retaining prior knowledge. While Gaussian processes elegantly tackle these problems, they struggle with scalability and handling rich inputs, such as images. To address these issues, we introduce a technique that converts neural networks from weight space to function space, through a dual parameterization. Our parameterization offers: (*i*) a way to scale function-space methods to large data sets via sparsification, (*ii*) retention of prior knowledge when access to past data is limited, and (*iii*) a mechanism to incorporate new data without retraining. Our experiments demonstrate that we can retain knowledge in continual learning and incorporate new data efficiently. We further show its strengths in uncertainty quantification and guiding exploration in model-based RL. Further information and code is available on the project website.
****
Chaoqi Wang,Yibo Jiang,Chenghao Yang,Han Liu,Yuxin Chen
Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints
The increasing capabilities of large language models (LLMs) raise opportunities for artificial general intelligence but concurrently amplify safety concerns, such as potential misuse of AI systems, necessitating effective AI alignment. Reinforcement Learning from Human Feedback (RLHF) has emerged as a promising pathway towards AI alignment but brings forth challenges due to its complexity and depe

ndence on a separate reward model. Direct Preference Optimization (DPO) has been proposed as an alternative; and it remains equivalent to RLHF under the reverse KL regularization constraint. This paper presents f-DPO, a generalized approach to DPO by incorporating diverse divergence constraints. We show that under certain f-divergences, including Jensen-Shannon divergence, forward KL divergences and α-divergences, the complex relationship between the reward and optimal policy can also be simplified by addressing the Karush–Kuhn–Tucker conditions. This eliminates the need for estimating the normalizing constant in the Bradley-Terry model and enables a tractable mapping between the reward function and the optimal policy. Our approach optimizes LLMs to align with human preferences in a more efficient and supervised manner under a broad set of divergence constraints. Empirically, adopting these divergences ensures a balance between alignment performance and generation diversity. Importantly, our f-DPO outperforms PPO-based methods in divergence efficiency, and divergence constraints directly influence expected calibration error (ECE).
****
Yichuan Li,Xiyao Ma,Sixing Lu,Kyumin Lee,Xiaohu Liu,Chenlei Guo
MEND: Meta Demonstration Distillation for Efficient and Effective In-Context Learning
Large Language models (LLMs) have demonstrated impressive in-context learning (ICL) capabilities,
where a LLM makes predictions for a given test input together with a few input-output pairs (demonstrations).
Nevertheless, the inclusion of demonstrations poses a challenge, leading to a quadratic increase in the computational overhead of the self-attention mechanism.
Existing solutions attempt to condense lengthy demonstrations into compact vectors.
However, they often require task-specific retraining or compromise LLM's in-context learning performance.
To mitigate these challenges, we present Meta Demonstration Distillation (MEND), where a language model learns to distill any lengthy demonstrations into vectors without retraining for a new downstream task.
We exploit the knowledge distillation to enhance alignment between MEND and MEND, achieving both efficiency and effectiveness concurrently.
MEND is endowed with the meta-knowledge of distilling demonstrations through a two-stage training process, which includes meta-distillation pretraining and fine-tuning.
Comprehensive evaluations across seven diverse ICL settings using decoder-only (GPT-2) and encoder-decoder (T5) attest to MEND's prowess.
It not only matches but often outperforms the Vanilla ICL as well as other state-of-the-art distillation models, while significantly reducing the computational demands.
This innovation promises enhanced scalability and efficiency for the practical deployment of large language models.
****
Andreas Bergmeister,Karolis Martinkus,Nathanaël Perraudin,Roger Wattenhofer
Efficient and Scalable Graph Generation through Iterative Local Expansion
In the realm of generative models for graphs, extensive research has been conducted. However, most existing methods struggle with large graphs due to the complexity of representing the entire joint distribution across all node pairs and capturing both global and local graph structures simultaneously.
To overcome these issues, we introduce a method that generates a graph by progressively expanding a single node to a target graph. In each step, nodes and edges are added in a localized manner through denoising diffusion, building first the global structure, and then refining the local details. The local generation avoids modeling the entire joint distribution over all node pairs, achieving substantial computational savings with subquadratic runtime relative to node count while maintaining high expressivity through multiscale generation.
Our experiments show that our model achieves state-of-the-art performance on well-established benchmark datasets while successfully scaling to graphs with at le

ast 5000 nodes. Our method is also the first to successfully extrapolate to grap
hs outside of the training distribution, showcasing a much better generalization
 capability over existing methods.
****
Dongjin Kim,Donggoo Jung,Sungyong Baik,Tae Hyun Kim
sRGB Real Noise Modeling via Noise-Aware Sampling with Normalizing Flows
Noise poses a widespread challenge in signal processing, particularly when it co
mes to denoising images. Although convolutional neural networks (CNNs) have exhi
bited remarkable success in this field, they are predicated upon the belief that
 noise follows established distributions, which restricts their practicality whe
n dealing with real-world noise. To overcome this limitation, several efforts ha
ve been taken to collect noisy image datasets from the real world. Generative me
thods, employing techniques such as generative adversarial networks (GANs) and n
ormalizing flows (NFs), have emerged as a solution for generating realistic nois
y images. Recent works model noise using camera metadata, however requiring meta
data even for sampling phase. In contrast, in this work, we aim to estimate the
underlying camera settings, enabling us to improve noise modeling and generate d
iverse noise distributions. To this end, we introduce a new NF framework that al
lows us to both classify noise based on camera settings and generate various noi
sy images. Through experimental results, our model demonstrates exceptional nois
e quality and leads in denoising performance on benchmark datasets.
****
Runqi Lin,Chaojian Yu,Bo Han,Tongliang Liu
On the Over-Memorization During Natural, Robust and Catastrophic Overfitting
Overfitting negatively impacts the generalization ability of deep neural network
s (DNNs) in both natural and adversarial training. Existing methods struggle to
consistently address different types of overfitting, typically designing strateg
ies that focus separately on either natural or adversarial patterns. In this wor
k, we adopt a unified perspective by solely focusing on natural patterns to expl
ore different types of overfitting. Specifically, we examine the memorization ef
fect in DNNs and reveal a shared behaviour termed over-memorization, which impai
rs their generalization capacity. This behaviour manifests as DNNs suddenly beco
ming high-confidence in predicting certain training patterns and retaining a per
sistent memory for them. Furthermore, when DNNs over-memorize an adversarial pat
tern, they tend to simultaneously exhibit high-confidence prediction for the cor
responding natural pattern. These findings motivate us to holistically mitigate
different types of overfitting by hindering the DNNs from over-memorization trai
ning patterns. To this end, we propose a general framework, $\textit{Distraction
 Over-Memorization}$ (DOM), which explicitly prevents over-memorization by eithe
r removing or augmenting the high-confidence natural patterns. Extensive experim
ents demonstrate the effectiveness of our proposed method in mitigating overfitt
ing across various training paradigms.
****
Luca Eyring,Dominik Klein,Théo Uscidda,Giovanni Palla,Niki Kilbertus,Zeynep Akat
a,Fabian J Theis
Unbalancedness in Neural Monge Maps Improves Unpaired Domain Translation
In optimal transport (OT), a Monge map is known as a mapping that transports a s
ource distribution to a target distribution in the most cost-efficient way. Rece
ntly, multiple neural estimators for Monge maps have been developed and applied
in diverse unpaired domain translation tasks, e.g. in single-cell biology and co
mputer vision. However, the classic OT framework enforces mass conservation, whi
ch
makes it prone to outliers and limits its applicability in real-world scenarios.
 The latter can be particularly harmful in OT domain translation tasks, where th
e relative position of a sample within a distribution is explicitly taken into a
ccount. While unbalanced OT tackles this challenge in the discrete setting, its
integration into neural Monge map estimators has received limited attention. We
propose a theoretically
grounded method to incorporate unbalancedness into any Monge map estimator. We i
mprove existing estimators to model cell trajectories over time and to predict c

ellular responses to perturbations. Moreover, our approach seamlessly integrates with the OT flow matching (OT-FM) framework. While we show that OT-FM performs competitively in image translation, we further improve performance by incorporating unbalancedness (UOT-FM), which better preserves relevant features. We hence establish UOT-FM as a principled method for unpaired image translation .
****
Yuheng Jing,Kai Li,Bingyun Liu,Yifan Zang,Haobo Fu,QIANG FU,Junliang Xing,Jian C heng
Towards Offline Opponent Modeling with In-context Learning
Opponent modeling aims at learning the opponent's behaviors, goals, or beliefs t o reduce the uncertainty of the competitive environment and assist decision-maki ng. Existing work has mostly focused on learning opponent models online, which i s impractical and inefficient in practical scenarios. To this end, we formalize an Offline Opponent Modeling (OOM) problem with the objective of utilizing pre-c ollected offline datasets to learn opponent models that characterize the opponen t from the viewpoint of the controlled agent, which aids in adapting to the unkn own fixed policies of the opponent. Drawing on the promises of the Transformers for decision-making, we introduce a general approach, Transformer Against Oppone nt (TAO), for OOM. Essentially, TAO tackles the problem by harnessing the full p otential of the supervised pre-trained Transformers' in-context learning capabil ities. The foundation of TAO lies in three stages: an innovative offline policy embedding learning stage, an offline opponent-aware response policy training sta ge, and a deployment stage for opponent adaptation with in-context learning. The oretical analysis establishes TAO's equivalence to Bayesian posterior sampling i n opponent modeling and guarantees TAO's convergence in opponent policy recognit ion. Extensive experiments and ablation studies on competitive environments with sparse and dense rewards demonstrate the impressive performance of TAO. Our app roach manifests remarkable prowess for fast adaptation, especially in the face o f unseen opponent policies, confirming its in-context learning potency.
****
Shahriar Golchin,Mihai Surdeanu
Time Travel in LLMs: Tracing Data Contamination in Large Language Models
Data contamination, i.e., the presence of test data from downstream tasks in the training data of large language models (LLMs), is a potential major issue in me asuring LLMs' real effectiveness on other tasks. We propose a straightforward ye t effective method for identifying data contamination within LLMs. At its core, our approach starts by identifying potential contamination at the instance level ; using this information, our approach then assesses wider contamination at the partition level. To estimate contamination of individual instances, we employ "g uided instruction:" a prompt consisting of the dataset name, partition type, and the random-length initial segment of a reference instance, asking the LLM to co mplete it. An instance is flagged as contaminated if the LLM's output either exa ctly or nearly matches the latter segment of the reference. To understand if an entire partition is contaminated, we propose two ideas. The first idea marks a d ataset partition as contaminated if the average overlap score with the reference instances (as measured by ROUGE-L or BLEURT) is statistically significantly bet ter with the completions from guided instruction compared to a "general instruct ion" that does not include the dataset and partition name. The second idea marks a dataset partition as contaminated if a classifier based on GPT-4 with few-sho t in-context learning prompt marks multiple generated completions as exact/near-exact matches of the corresponding reference instances. Our best method achieves an accuracy between 92% and 100% in detecting if an LLM is contaminated with se ven datasets, containing train and test/validation partitions, when contrasted w ith manual evaluation by human experts. Further, our findings indicate that GPT-4 is contaminated with AG News, WNLI, and XSum datasets.
****
Shengyao Lu,Keith G Mills,Jiao He,Bang Liu,Di Niu
GOAt: Explaining Graph Neural Networks via Graph Output Attribution
Understanding the decision-making process of Graph Neural Networks (GNNs) is cru

cial to their interpretability. Most existing methods for explaining GNNs typica
lly rely on training auxiliary models, resulting in the explanations remain blac
k-boxed. This paper introduces Graph Output Attribution (GOAt), a novel method t
o attribute graph outputs to input graph features, creating GNN explanations tha
t are faithful, discriminative, as well as stable across similar samples. By exp
anding the GNN as a sum of scalar products involving node features, edge feature
s and activation patterns, we propose an efficient analytical method to compute
contribution of each node or edge feature to each scalar product and aggregate t
he contributions from all scalar products in the expansion form to derive the im
portance of each node and edge. Through extensive experiments on synthetic and r
eal-world data, we show that our method not only outperforms various state-of-th
e-art GNN explainers in terms of the commonly used fidelity metric, but also exh
ibits stronger discriminability, and stability by a remarkable margin.
****
Bang An,Sicheng Zhu,Michael-Andrei Panaitescu-Liess,Chaithanya Kumar Mummadi,Fur
ong Huang
PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts
Vision-language models like CLIP are widely used in zero-shot image classificati
on due to their ability to understand various visual concepts and natural langua
ge descriptions. However, how to fully leverage CLIP's unprecedented human-like
understanding capabilities to achieve better performance is still an open questi
on. This paper draws inspiration from the human visual perception process: when
classifying an object, humans first infer contextual attributes (e.g., backgroun
d and orientation) which help separate the foreground object from the background
, and then classify the object based on this information. Inspired by it, we obs
erve that providing CLIP with contextual attributes improves zero-shot image cla
ssification and mitigates reliance on spurious features. We also observe that CL
IP itself can reasonably infer the attributes from an image. With these observat
ions, we propose a training-free, two-step zero-shot classification method Perce
ptionCLIP. Given an image, it first infers contextual attributes (e.g., backgrou
nd) and then performs object classification conditioning on them. Our experiment
s show that PerceptionCLIP achieves better generalization, group robustness, and
 interpretability.
****
Zhongxia Yan,Cathy Wu
Neural Neighborhood Search for Multi-agent Path Finding
Multi-agent path finding (MAPF) is the combinatorial problem of planning optimal
 collision-avoiding paths for multiple agents, with application to robotics, log
istics, and transportation. Though many recent learning-based works have focused
 on large-scale combinatorial problems by guiding their decomposition into seque
nces of smaller subproblems, the combined spatiotemporal and time-restricted nat
ure of MAPF poses a particular challenge for learning-based guidance of iterativ
e approaches like large neighborhood search (LNS), which is already a state-of-t
he-art approach for MAPF even without learning. We address this challenge of neu
ral-guided LNS for MAPF by designing an architecture which interleaves convoluti
on and attention to efficiently represent MAPF subproblems, enabling practical g
uidance of LNS in benchmark settings. We demonstrate the speedup of our method o
ver existing state-of-the-art LNS-based methods for MAPF as well as the robustne
ss of our method to unseen settings. Our proposed method expands the horizon of
effective deep learning-guided LNS methods into multi-path planning problems, an
d our proposed representation may be more broadly applicable for representing pa
th-wise interactions.
****
Sotirios Panagiotis Chytas,Vishnu Suresh Lokhande,Vikas Singh
Pooling Image Datasets with Multiple Covariate Shift and Imbalance
Small sample sizes are common in many disciplines,
which necessitates pooling roughly similar datasets across
multiple sites/institutions to study weak but relevant
associations between images and disease incidence. Such
data often manifest shifts and imbalances in covariates

(secondary non-imaging data).
These issues are well-studied for classical models, but
the ideas simply do not apply to overparameterized DNN models.
Consequently, recent work has shown how strategies from
fairness and invariant representation learning provides
a meaningful starting point, but the current repertoire
of methods remains limited to accounting for shifts/imbalances in just a couple
of covariates at a time. In this paper, we show how
viewing this problem from the perspective of Category theory
provides a simple and effective solution that completely avoids
elaborate multi-stage training pipelines that would otherwise be
needed. We show the effectiveness of this approach via
extensive experiments on real datasets. Further, we
discuss how our style of formulation offers a unified
perspective on at least 5+ distinct
problem settings in vision, from self-supervised learning
to matching problems in 3D reconstruction.
****
Giung Nam,Byeongho Heo,Juho Lee
Lipsum-FT: Robust Fine-Tuning of Zero-Shot Models Using Random Text Guidance
Large-scale contrastive vision-language pre-trained models provide the zero-shot model achieving competitive performance across a range of image classification tasks without requiring training on downstream data. Recent works have confirmed that while additional fine-tuning of the zero-shot model on the reference data results in enhanced downstream performance, it compromises the model's robustnes s against distribution shifts. Our investigation begins by examining the conditi ons required to achieve the goals of robust fine-tuning, employing descriptions based on feature distortion theory and joint energy-based models. Subsequently, we propose a novel robust fine-tuning algorithm, Lipsum-FT, that effectively uti lizes the language modeling aspect of the vision-language pre-trained models. Ex tensive experiments conducted on distribution shift scenarios in DomainNet and I mageNet confirm the superiority of our proposed Lipsum-FT approach over existing robust fine-tuning methods.
****
Jingcheng Niu,Andrew Liu,Zining Zhu,Gerald Penn
What does the Knowledge Neuron Thesis Have to do with Knowledge?
We reassess the Knowledge Neuron (KN) Thesis: an interpretation of the mechanism underlying the ability of large language models to recall facts from a training corpus. This nascent thesis proposes that facts are recalled from the training corpus through the MLP weights in a manner resembling key-value memory, implying in effect that "knowledge" is stored in the network. Furthermore, by modifying the MLP modules, one can control the language model's generation of factual info rmation. The plausibility of the KN thesis has been demonstrated by the success of KN-inspired model editing methods (Dai et al., 2022; Meng et al., 2022).

We find that this thesis is, at best, an oversimplification. Not only have we fo und that we can edit the expression of certain linguistic phenomena using the sa me model editing methods but, through a more comprehensive evaluation, we have f ound that the KN thesis does not adequately explain the process of factual expre ssion. While it is possible to argue that the MLP weights store complex patterns that are interpretable both syntactically and semantically, these patterns do n ot constitute "knowledge." To gain a more comprehensive understanding of the kno wledge representation process, we must look beyond the MLP weights and explore r ecent models' complex layer structures and attention mechanisms.
****
Yiheng Du,Nithin Chalapathi,Aditi S. Krishnapriyan
Neural Spectral Methods: Self-supervised learning in the spectral domain
We present Neural Spectral Methods, a technique to solve parametric Partial Diff erential Equations (PDEs), grounded in classical spectral methods. Our method us es orthogonal bases to learn PDE solutions as mappings between spectral coeffici

ents, instantiating a spectral-based neural operator. In contrast to current machine learning approaches which enforce PDE constraints by minimizing the numerical quadrature of the residuals in the spatiotemporal domain, we leverage Parseval's identity and introduce a new training strategy through a spectral loss. Our spectral loss enables more efficient differentiation through the neural network, and substantially reduces training complexity. At inference time, the computational cost of our method remains constant, regardless of the spatiotemporal resolution of the domain. Our experimental results demonstrate that our method significantly outperforms previous machine learning approaches in terms of speed and accuracy by one to two orders of magnitude on multiple different problems, including reaction-diffusion, and forced and unforced Navier-Stokes equations. When compared to numerical solvers of the same accuracy, our method demonstrates a $10\times$ increase in performance speed. Our source code is publicly available at https://github.com/ASK-Berkeley/Neural-Spectral-Methods.
****
Hyeonho Jeong,Jong Chul Ye
Ground-A-Video: Zero-shot Grounded Video Editing using Text-to-image Diffusion Models
This paper introduces a novel grounding-guided video-to-video translation framework called Ground-A-Video for multi-attribute video editing.
Recent endeavors in video editing have showcased promising results in single-attribute editing or style transfer tasks, either by training T2V models on text-video data or adopting training-free methods.
However, when confronted with the complexities of multi-attribute editing scenarios, they exhibit shortcomings such as omitting or overlooking intended attribute changes, modifying the wrong elements of the input video, and failing to preserve regions of the input video that should remain intact.
Ground-A-Video attains temporally consistent multi-attribute editing of input videos in a training-free manner without aforementioned shortcomings.
Central to our method is the introduction of cross-frame gated attention which incorporates groundings information into the latent representations in a temporally consistent fashion, along with Modulated Cross-Attention and optical flow guided inverted latents smoothing.
Extensive experiments and applications demonstrate that Ground-A-Video's zero-shot capacity outperforms other baseline methods in terms of edit-accuracy and frame consistency.
Further results and code are available at our project page (http://ground-a-video.github.io)
****
Zijun Wu,Yongkang Wu,Lili Mou
Zero-Shot Continuous Prompt Transfer: Generalizing Task Semantics Across Language Models
Prompt tuning in natural language processing (NLP) has become an increasingly popular method for adapting large language models to specific tasks. However, the transferability of these prompts, especially continuous prompts, between different models remains a challenge. In this work, we propose a zero-shot continuous prompt transfer method, where source prompts are encoded into relative space and the corresponding target prompts are searched for transferring to target models. Experimental results confirm the effectiveness of our method, showing that 'task semantics' in continuous prompts can be generalized across various language models. Moreover, we find that combining 'task semantics' from multiple source models can further enhance the performance of transfer.
****
Marina Zhang,Owen Skipper Vallis,Aysegul Bumin,Tanay Vakharia,Elie Bursztein
RETSim: Resilient and Efficient Text Similarity
This paper introduces RETSim (Resilient and Efficient Text Similarity), a lightweight, multilingual deep learning model trained to produce robust metric embeddings for near-duplicate text retrieval, clustering, and dataset deduplication tasks. We demonstrate that RETSim is significantly more robust and accurate than MinHash and neural text embeddings, achieving new state-of-the-art performance on

dataset deduplication, adversarial text retrieval benchmarks, and spam clusterin
g tasks. Additionally, we introduce the W4NT3D benchmark (Wiki-40B 4dversarial N
ear-T3xt Dataset), enabling the evaluation of models on typo-laden near-duplicat
e text retrieval in a multilingual setting. RETSim and the W4NT3D benchmark are
released under the MIT License at https://github.com/google/unisim.
****
Xinyuan Wang,Chenxi Li,Zhen Wang,Fan Bai,Haotian Luo,Jiayou Zhang,Nebojsa Jojic,
Eric Xing,Zhiting Hu
PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt
 Optimization
Expert-level prompts, carefully engineered by human experts who have a deep unde
rstanding of both large language models (LLMs) and domain knowledge, are the fut
ure of prompting and pivotal to harnessing the full power of advanced LLMs. Disc
overing such prompts with an automated process remains a sought-after and unreso
lved challenge. Existing prompt optimization techniques, though automated throug
h iterative sampling, often fall short in injecting domain knowledge and explori
ng the vast prompt space for complex expert-level prompts efficiently. To addres
s this pressing need and achieve expert-level prompting, we introduce PromptAgen
t, which autonomously discovers prompts equivalent in quality to those handcraft
ed by experts. At its core, PromptAgent views prompt optimization as a strategic
 planning problem and employs a principled planning algorithm (rooted in Monte C
arlo Tree Search) to strategically explore the vast expert-level prompt space. P
romptAgent interacts with the LLM in a human-like trial-and-error manner during
the planning, and injects expert-level knowledge by reflecting on model errors a
nd generating insightful error feedback. This novel formulation allows it to ite
ratively evaluate intermediate prompts, refine them based on errors, simulate fu
ture rewards, and search for high-reward paths leading to expert-level prompts.
We apply PromptAgent to 12 tasks spanning three practical domains: BIG-Bench Har
d (BBH), domain-expert, and general NLU tasks, showing PromptAgent consistently
outperforms strong prompting and prompt optimization baselines by great margins.
 Our qualitative analysis further emphasizes PromptAgent's capability to distill
 insightful errors into expert-level prompts.
****
Xi Victoria Lin,Xilun Chen,Mingda Chen,Weijia Shi,Maria Lomeli,Richard James,Ped
ro Rodriguez,Jacob Kahn,Gergely Szilvasy,Mike Lewis,Luke Zettlemoyer,Wen-tau Yih
RA-DIT: Retrieval-Augmented Dual Instruction Tuning
Retrieval-augmented language models (RALMs) improve performance by accessing lon
g-tail and up-to-date knowledge from external data stores, but are challenging t
o build. Existing approaches require either expensive retrieval-specific modific
ations to LM pre-training or use post-hoc integration of the data store that lea
ds to suboptimal performance. We introduce Retrieval-Augmented Dual Instruction
Tuning (RA-DIT), a lightweight fine-tuning methodology that provides a third opt
ion by retrofitting any LLM with retrieval capabilities. Our approach operates i
n two distinct fine-tuning steps: (1) one updates a pre-trained LM to better use
 retrieved information, while (2) the other updates the retriever to return more
 relevant results, as preferred by the LM. By fine-tuning over tasks that requir
e both knowledge utilization and contextual awareness, we demonstrate that each
stage yields significant performance improvements, and using both leads to addit
ional gains. Our best model, RA-DIT 65B, achieves state-of-the-art performance a
cross a range of knowledge-intensive zero- and few-shot learning benchmarks, sig
nificantly outperforming existing in-context RALM approaches by up to +8.9% in 0
-shot setting and +1.4% in 5-shot setting on average.
****
Renze Lou,Kai Zhang,Jian Xie,Yuxuan Sun,Janice Ahn,Hanzi Xu,Yu Su,Wenpeng Yin
MUFFIN: Curating Multi-Faceted Instructions for Improving Instruction Following
In the realm of large language models (LLMs), enhancing instruction-following ca
pability often involves curating expansive training data. This is achieved throu
gh two primary schemes: i) Scaling-Inputs: Amplifying (input, output) pairs per
task instruction, aiming for better instruction adherence. ii) Scaling Input-Fre
e Tasks: Enlarging tasks, each composed of an (instruction, output) pair (withou

t requiring a separate input anymore). However, LLMs under Scaling-Inputs tend to be overly sensitive to inputs, leading to misinterpretation or non-compliance with instructions. Conversely, Scaling Input-Free Tasks demands a substantial number of tasks but is less effective in instruction following when dealing with instances in Scaling-Inputs. This work introduces MUFFIN, a new scheme of instruction-following dataset curation. Specifically, we automatically Scale Tasks per Input by diversifying these tasks with various input facets. Experimental results across four zero-shot benchmarks, spanning both Scaling-Inputs and Scaling Input-Free Tasks schemes, reveal that LLMs, at various scales, trained on MUFFIN generally demonstrate superior instruction-following capabilities compared to those trained on the two aforementioned schemes.
****
Kevin Clark,Paul Vicol,Kevin Swersky,David J. Fleet
Directly Fine-Tuning Diffusion Models on Differentiable Rewards
We present Direct Reward Fine-Tuning (DRaFT), a simple and effective method for fine-tuning diffusion models to maximize differentiable reward functions, such as scores from human preference models. We first show that it is possible to backpropagate the reward function gradient through the full sampling procedure, and that doing so achieves strong performance on a variety of rewards, outperforming reinforcement learning-based approaches. We then propose more efficient variants of DRaFT: DRaFT-K, which truncates backpropagation to only the last K steps of sampling, and DRaFT-LV, which obtains lower-variance gradient estimates for the case when K=1. We show that our methods work well for a variety of reward functions and can be used to substantially improve the aesthetic quality of images generated by Stable Diffusion 1.4. Finally, we draw connections between our approach and prior work, providing a unifying perspective on the design space of gradient-based fine-tuning algorithms.
****
Mohammad Reza Samsami,Artem Zholus,Janarthanan Rajendran,Sarath Chandar
Mastering Memory Tasks with World Models
Current model-based reinforcement learning (MBRL) agents struggle with long-term dependencies. This limits their ability to effectively solve tasks involving extended time gaps between actions and outcomes, or tasks demanding the recalling of distant observations to inform current actions. To improve temporal coherence, we integrate a new family of state space models (SSMs) in world models of MBRL agents to present a new method, Recall to Imagine (R2I). This integration aims to enhance both long-term memory and long-horizon credit assignment. Through a diverse set of illustrative tasks, we systematically demonstrate that R2I not only establishes a new state-of-the-art for challenging memory and credit assignment RL tasks, such as BSuite and POPGym, but also showcases superhuman performance in the complex memory domain of Memory Maze. At the same time, it upholds comparable performance in classic RL tasks, such as Atari and DMC, suggesting the generality of our method. We also show that R2I is faster than the state-of-the-art MBRL method, DreamerV3, resulting in faster wall-time convergence.
****
Zeqi Xiao,Tai Wang,Jingbo Wang,Jinkun Cao,Wenwei Zhang,Bo Dai,Dahua Lin,Jiangmiao Pang
Unified Human-Scene Interaction via Prompted Chain-of-Contacts
Human-Scene Interaction (HSI) is a vital component of fields like embodied AI and virtual reality. Despite advancements in motion quality and physical plausibility, two pivotal factors, versatile interaction control and the development of a user-friendly interface, require further exploration before the practical application of HSI. This paper presents a unified HSI framework, UniHSI, which supports unified control of diverse interactions through language commands. The framework defines interaction as ``Chain of Contacts (CoC)", representing steps involving human joint-object part pairs. This concept is inspired by the strong correlation between interaction types and corresponding contact regions. Based on the definition, UniHSI constitutes a Large Language Model (LLM) Planner to translate language prompts into task plans in the form of CoC, and a Unified Controller that turns CoC into uniform task execution. To facilitate training and evaluation

, we collect a new dataset named ScenePlan that encompasses thousands of task plans generated by LLMs based on diverse scenarios. Comprehensive experiments demonstrate the effectiveness of our framework in versatile task execution and generalizability to real scanned scenes.
****
Deyao Zhu,Jun Chen,Xiaoqian Shen,Xiang Li,Mohamed Elhoseiny
MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models
The recent GPT-4 has demonstrated extraordinary multi-modal abilities, such as directly generating websites from handwritten text and identifying humorous elements within images. These features are rarely observed in previous vision-language models. However, the technical details behind GPT-4 continue to remain undisclosed.
We believe that the enhanced multi-modal generation capabilities of GPT-4 stem from the utilization of sophisticated large language models (LLM).
To examine this phenomenon, we present MiniGPT-4, which aligns a frozen visual encoder with a frozen advanced LLM, Vicuna, using one projection layer.
Our work, for the first time, uncovers that properly aligning the visual features with an advanced large language model can possess numerous advanced multi-modal abilities demonstrated by GPT-4,
such as detailed image description generation and website creation from hand-drawn drafts.
Furthermore, we also observe other emerging capabilities in MiniGPT-4, including writing stories and poems inspired by given images, teaching users how to cook based on food photos, and so on.
In our experiment, we found that the model trained on short image caption pairs could produce unnatural language outputs (e.g., repetition and fragmentation). To address this problem, we curate a detailed image description dataset in the second stage to finetune the model, which consequently improves the model's generation reliability and overall usability.
****
Bo Peng,Yadan Luo,Yonggang Zhang,Yixuan Li,Zhen Fang
ConjNorm: Tractable Density Estimation for Out-of-Distribution Detection
Post-hoc out-of-distribution (OOD) detection has garnered intensive attention in reliable machine learning. Many efforts have been dedicated to deriving score functions based on logits, distances, or rigorous data distribution assumptions to identify low-scoring OOD samples. Nevertheless, these estimate scores may fail to accurately reflect the true data density or impose impractical constraints. To provide a unified perspective on density-based score design, we propose a novel theoretical framework grounded in Bregman divergence, which extends distribution considerations to encompass an exponential family of distributions. Leveraging the conjugation constraint revealed in our theorem, we introduce a \textsc{ConjNorm} method, reframing density function design as a search for the optimal norm coefficient p against the given dataset. In light of the computational challenges of normalization, we devise an unbiased and analytically tractable estimator of the partition function using the Monte Carlo-based importance sampling technique. Extensive experiments across OOD detection benchmarks empirically demonstrate that our proposed \textsc{ConjNorm} has established a new state-of-the-art in a variety of OOD detection setups, outperforming the current best method by up to 13.25\% and 28.19\% (FPR95) on CIFAR-100 and ImageNet-1K, respectively.
****
Edward Milsom,Ben Anson,Laurence Aitchison
Convolutional Deep Kernel Machines
Standard infinite-width limits of neural networks sacrifice the ability for intermediate layers to learn representations from data. Recent work (“A theory of representation learning gives a deep generalisation of kernel methods”, Yang et al. 2023) modified the Neural Network Gaussian Process (NNGP) limit of Bayesian neural networks so that representation learning is retained. Furthermore, they found that applying this modified limit to a deep Gaussian process gives a practical learning algorithm which they dubbed the “deep kernel machine” (DKM). However,

they only considered the simplest possible setting: regression in small, fully connected networks with e.g. 10 input features. Here, we introduce convolutional deep kernel machines. This required us to develop a novel inter-domain inducing point approximation, as well as introducing and experimentally assessing a number of techniques not previously seen in DKMs, including analogues to batch normalisation, different likelihoods, and different types of top-layer. The resulting model trains in roughly 77 GPU hours, achieving around 99\% test accuracy on MNIST, 72\% on CIFAR-100, and 92.7\% on CIFAR-10, which is SOTA for kernel methods.
****
Shaopeng Fu,Di Wang
Theoretical Analysis of Robust Overfitting for Wide DNNs: An NTK Approach
Adversarial training (AT) is a canonical method for enhancing the robustness of deep neural networks (DNNs). However, recent studies empirically demonstrated that it suffers from robust overfitting, i.e., a long time AT can be detrimental to the robustness of DNNs. This paper presents a theoretical explanation of robust overfitting for DNNs. Specifically, we non-trivially extend the neural tangent kernel (NTK) theory to AT and prove that an adversarially trained wide DNN can be well approximated by a linearized DNN. Moreover, for squared loss, closed-form AT dynamics for the linearized DNN can be derived, which reveals a new AT degeneration phenomenon: a long-term AT will result in a wide DNN degenerates to that obtained without AT and thus cause robust overfitting. Based on our theoretical results, we further design a method namely Adv-NTK, the first AT algorithm for infinite-width DNNs. Experiments on real-world datasets show that Adv-NTK can help infinite-width DNNs enhance comparable robustness to that of their finite-width counterparts, which in turn justifies our theoretical findings. The code is available at https://github.com/fshp971/adv-ntk.
****
Xian Li,Ping Yu,Chunting Zhou,Timo Schick,Omer Levy,Luke Zettlemoyer,Jason E Weston,Mike Lewis
Self-Alignment with Instruction Backtranslation
We present a scalable method to build a high quality instruction following language model by automatically labelling human-written text with corresponding instructions. Our approach, named instruction backtranslation, starts with a language model finetuned on a small amount of seed data, and a given web corpus. The seed model is used to construct training examples by generating instruction prompts for web documents (self-augmentation), and then selecting high quality examples from among these candidates (self-curation). This data is then used to finetune a stronger model. Finetuning LLaMa on two iterations of our approach yields a model that outperforms all other LLaMa-based models on the Alpaca leaderboard not relying on distillation data, demonstrating highly effective self-alignment.
****
Yuxin Zhang,Lirui Zhao,Mingbao Lin,Sun Yunyun,Yiwu Yao,Xingjia Han,Jared Tanner,Shiwei Liu,Rongrong Ji
Dynamic Sparse No Training: Training-Free Fine-tuning for Sparse LLMs
The ever-increasing large language models (LLMs), though opening a potential path for the upcoming artificial general intelligence, sadly drops a daunting obstacle on the way towards their on-device deployment. As one of the most well-established pre-LLMs approaches in reducing model complexity, network pruning appears to lag behind in the era of LLMs, due mostly to its costly fine-tuning (or re-training) necessity under the massive volumes of model parameter and training data. To close this industry-academia gap, we introduce Dynamic Sparse No Training ($\texttt{DSNT}$), a training-free fine-tuning approach that slightly updates sparse LLMs without the expensive backpropagation and any weight updates. Inspired by the Dynamic Sparse Training, $\texttt{DSNT}$ minimizes the reconstruction error between the dense and sparse LLMs, in the fashion of performing iterative weight pruning-and-growing on top of sparse LLMs. To accomplish this purpose, $\texttt{DSNT}$ particularly takes into account the anticipated reduction in reconstruction error for pruning and growing, as well as the variance w.r.t. different input data for growing each weight. This practice can be executed efficiently in

linear time since its obviates the need of backpropagation for fine-tuning LLMs. Extensive experiments on LLaMA-V1/V2, Vicuna, and OPT across various benchmarks demonstrate the effectiveness of $\texttt{DSNT}$ in enhancing the performance of sparse LLMs, especially at high sparsity levels. For instance, $\texttt{DSNT}$ is able to outperform the state-of-the-art Wanda by 26.79 perplexity at 70% sparsity with LLaMA-7B. Our paper offers fresh insights into how to fine-tune sparse LLMs in an efficient training-free manner and open new venues to scale the great potential of sparsity to LLMs. Codes are available at https://github.com/zyxxmu/DSnoT.
****
Adrián Bazaga,Pietro Lio,Gos Micklem
Unsupervised Pretraining for Fact Verification by Language Model Distillation
Fact verification aims to verify a claim using evidence from a trustworthy knowledge base. To address this challenge, algorithms must produce features for every claim that are both semantically meaningful, and compact enough to find a semantic alignment with the source information. In contrast to previous work, which tackled the alignment problem by learning over annotated corpora of claims and their corresponding labels, we propose SFAVEL ($\underline{S}$elf-supervised $\underline{Fa}$ct $\underline{Ve}$rification via $\underline{L}$anguage Model Distillation), a novel unsupervised pretraining framework that leverages pre-trained language models to distil self-supervised features into high-quality claim-fact alignments without the need for annotations. This is enabled by a novel contrastive loss function that encourages features to attain high-quality claim and evidence alignments whilst preserving the semantic relationships across the corpora. Notably, we present results that achieve a new state-of-the-art on FB15k-237 (+5.3\% Hits@1) and FEVER (+8\% accuracy) with linear evaluation.
****
Sidhika Balachandar,Nikhil Garg,Emma Pierson
Domain constraints improve risk prediction when outcome data is missing
Machine learning models are often trained to predict the outcome resulting from a human decision. For example, if a doctor decides to test a patient for disease, will the patient test positive? A challenge is that historical decision-making determines whether the outcome is observed: we only observe test outcomes for patients doctors historically tested. Untested patients, for whom outcomes are unobserved, may differ from tested patients along observed and unobserved dimensions. We propose a Bayesian model class which captures this setting. The purpose of the model is to accurately estimate risk for both tested and untested patients. Estimating this model is challenging due to the wide range of possibilities for untested patients. To address this, we propose two domain constraints which are plausible in health settings: a prevalence constraint, where the overall disease prevalence is known, and an expertise constraint, where the human decision-maker deviates from purely risk-based decision-making only along a constrained feature set. We show theoretically and on synthetic data that domain constraints improve parameter inference. We apply our model to a case study of cancer risk prediction, showing that the model's inferred risk predicts cancer diagnoses, its inferred testing policy captures known public health policies, and it can identify suboptimalities in test allocation. Though our case study is in healthcare, our analysis reveals a general class of domain constraints which can improve model estimation in many settings.
****
Xingchao Liu,Xiwen Zhang,Jianzhu Ma,Jian Peng,qiang liu
InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation
Diffusion models have revolutionized text-to-image generation with its exceptional quality and creativity. However, its multi-step sampling process is known to be slow, often requiring tens of inference steps to obtain satisfactory results. Previous attempts to improve its sampling speed and reduce computational costs through distillation have been unsuccessful in achieving a functional one-step model.
In this paper, we explore a recent method called Rectified Flow, which, thus far

, has only been applied to small datasets. The core of Rectified Flow lies in it
s \emph{reflow} procedure, which straightens the trajectories of probability flo
ws, refines the coupling between noises and images, and facilitates the distilla
tion process with student models. We propose a novel text-conditioned pipeline t
o turn Stable Diffusion (SD) into an ultra-fast one-step model, in which we find
 reflow plays a critical role in improving the assignment between noise and imag
es. Leveraging our new pipeline, we create, to the best of our knowledge, the fi
rst one-step diffusion-based text-to-image generator with SD-level image quality
, achieving an FID (Fréchet Inception Distance) of 23.3 on MS COCO 2017-5k, su
rpassing the previous state-of-the-art technique, progressive distillation, by a
 significant margin (37.2 $\rightarrow$ 23.3 in FID). By utilizing an expand
ed network with 1.7B parameters, we further improve the FID to 22.4. We call o
ur one-step models \emph{InstaFlow}. On MS COCO 2014-30k, InstaFlow yields an FI
D of 13.1 in just 0.09 second, the best in ≤ 0.1 second regime, outperf
orming the recent StyleGAN-T (13.9 in 0.1 second). Notably, the training of
InstaFlow only costs 199 A100 GPU days. Codes and pre-trained models are availab
le at \url{github.com/gnobitab/InstaFlow}.
****
Zhuoyan Xu,Zhenmei Shi,Junyi Wei,Fangzhou Mu,Yin Li,Yingyu Liang
Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning
Foundation models have emerged as a powerful tool for many AI problems. Despite
the tremendous success of foundation models, effective adaptation to new tasks,
particularly those with limited labels, remains an open question and lacks theor
etical understanding.
  An emerging solution with recent success in vision and NLP involves finetuning
 a foundation model on a selection of relevant tasks, before its adaptation to a
 target task with limited labeled samples. In this paper, we study the theoretic
al justification of this multitask finetuning approach.
Our theoretical analysis reveals that with a diverse set of related tasks, this
multitask finetuning leads to reduced error in the target task, in comparison to
 directly adapting the same pretrained model. We quantify the relationship betwe
en finetuning tasks and target tasks by diversity and consistency metrics, and f
urther propose a practical task selection algorithm.
  We substantiate our theoretical claims with extensive empirical evidence.
Further, we present results affirming our task selection algorithm adeptly choos
es related finetuning tasks, providing advantages to the model performance on ta
rget tasks.
  We believe our study shed new light on the effective adaptation of foundation
models to new tasks that lack abundant labels.
  Our code is available at https://github.com/OliverXUZY/Foudation-Model_Multita
sk.
****
Tomoya Murata,Kenta Niwa,Takumi Fukami,Iifan Tyou
Simple Minimax Optimal Byzantine Robust Algorithm for Nonconvex Objectives with
Uniform Gradient Heterogeneity
In this study, we consider nonconvex federated learning problems with the existe
nce of Byzantine workers. We propose a new simple Byzantine robust algorithm cal
led Momentum Screening. The algorithm is adaptive to the Byzantine fraction, i.e
., all its hyperparameters do not depend on the number of Byzantine workers. We
show that our method achieves the best optimization error of $O(\delta^2\zeta_\m
athrm{max}^2)$ for nonconvex smooth local objectives satisfying $\zeta_\mathrm{m
ax}$-uniform gradient heterogeneity condition under δ-Byzantine fraction,
 which can be better than the best known error rate of $O(\delta\zeta_\mathrm{me
an}^2)$ for local objectives satisfying ζ_mean-mean heterogeneity
 condition when $\delta \leq (\zeta_\mathrm{max}/\zeta_\mathrm{mean})^2$. Furthe
rmore, we derive an algorithm independent lower bound for local objectives satis
fying ζ_max-uniform gradient heterogeneity condition and show the
 minimax optimality of our proposed method on this class. In numerical experimen
ts, we validate the superiority of our method over the existing robust aggregati
on algorithms and verify our theoretical results.

****
Jiacheng Guo,Minshuo Chen,Huan Wang,Caiming Xiong,Mengdi Wang,Yu Bai
Sample-Efficient Learning of POMDPs with Multiple Observations In Hindsight
This paper studies the sample-efficiency of learning in Partially Observable Markov Decision Processes (POMDPs), a challenging problem in reinforcement learning that is known to be exponentially hard in the worst-case. Motivated by real-world settings such as loading in game playing, we propose an enhanced feedback model called ``multiple observations in hindsight'', where after each episode of interaction with the POMDP, the learner may collect multiple additional observations emitted from the encountered latent states, but may not observe the latent states themselves. We show that sample-efficient learning under this feedback model is possible for two new subclasses of POMDPs: \emph{multi-observation revealing POMDPs} and \emph{distinguishable POMDPs}. Both subclasses generalize and substantially relax \emph{revealing POMDPs}---a widely studied subclass for which sample-efficient learning is possible under standard trajectory feedback. Notably, distinguishable POMDPs only require the emission distributions from different latent states to be \emph{different} instead of \emph{linearly independent} as required in revealing POMDPs.
****
Bo Zhang,Xinyu Cai,Jiakang Yuan,Donglin Yang,Jianfei Guo,Xiangchao Yan,Renqiu Xia,Botian Shi,Min Dou,Tao Chen,Si Liu,Junchi Yan,Yu Qiao
ReSimAD: Zero-Shot 3D Domain Transfer for Autonomous Driving with Source Reconstruction and Target Simulation
Domain shifts such as sensor type changes and geographical situation variations are prevalent in Autonomous Driving (AD), which poses a challenge since AD model relying on the previous domain knowledge can be hardly directly deployed to a new domain without additional costs. In this paper, we provide a new perspective and approach of alleviating the domain shifts, by proposing a Reconstruction-Simulation-Perception (ReSimAD) scheme. Specifically, the implicit reconstruction process is based on the knowledge from the previous old domain, aiming to convert the domain-related knowledge into domain-invariant representations, e.g., 3D scene-level meshes. Besides, the point clouds simulation process of multiple new domains is conditioned on the above reconstructed 3D meshes, where the target-domain-like simulation samples can be obtained, thus reducing the cost of collecting and annotating new-domain data for the subsequent perception process. For experiments, we consider different cross-domain situations such as Waymo-to-KITTI, Waymo-to-nuScenes, etc, to verify the zero-shot target-domain perception using ReSimAD. Results demonstrate that our method is beneficial to boost the domain generalization ability, even promising for 3D pre-training. Code and simulated points are available at: https://github.com/PJLab-ADG/3DTrans
****
Shiqian Li,Kewen Wu,Chi Zhang,Yixin Zhu
I-PHYRE: Interactive Physical Reasoning
Current evaluation protocols predominantly assess physical reasoning in stationary scenes, creating a gap in evaluating agents' abilities to interact with dynamic events. While contemporary methods allow agents to modify initial scene configurations and observe consequences, they lack the capability to interact with events in real time. To address this, we introduce I-PHYRE, a framework that challenges agents to simultaneously exhibit intuitive physical reasoning, multi-step planning, and in-situ intervention. Here, intuitive physical reasoning refers to a quick, approximate understanding of physics to address complex problems; multi-step denotes the need for extensive sequence planning in I-PHYRE, considering each intervention can significantly alter subsequent choices; and in-situ implies the necessity for timely object manipulation within a scene, where minor timing deviations can result in task failure. We formulate four game splits to scrutinize agents' learning and generalization of essential principles of interactive physical reasoning, fostering learning through interaction with representative scenarios. Our exploration involves three planning strategies and examines several supervised and reinforcement agents' zero-shot generalization proficiency on I-PHYRE. The outcomes highlight a notable gap between existing learning algorithm

s and human performance, emphasizing the imperative for more research in enhancing agents with interactive physical reasoning capabilities. The environment and baselines will be made publicly available.
****
Yukun Huang,Jianan Wang,Yukai Shi,Boshi Tang,Xianbiao Qi,Lei Zhang
DreamTime: An Improved Optimization Strategy for Diffusion-Guided 3D Generation
Text-to-image diffusion models pre-trained on billions of image-text pairs have recently enabled 3D content creation by optimizing a randomly initialized differentiable 3D representation with score distillation. However, the optimization process suffers slow convergence and the resultant 3D models often exhibit two limitations: (a) quality concerns such as missing attributes and distorted shape and texture; (b) extremely low diversity comparing to text-guided image synthesis. In this paper, we show that the conflict between the 3D optimization process and uniform timestep sampling in score distillation is the main reason for these limitations. To resolve this conflict, we propose to prioritize timestep sampling with monotonically non-increasing functions, which aligns the 3D optimization process with the sampling process of diffusion model. Extensive experiments show that our simple redesign significantly improves 3D content creation with faster convergence, better quality and diversity.
****
Alihan Hüyük,Qiyao Wei,Alicia Curth,Mihaela van der Schaar
Defining Expertise: Applications to Treatment Effect Estimation
Decision-makers are often experts of their domain and take actions based on their domain knowledge. Doctors, for instance, may prescribe treatments by predicting the likely outcome of each available treatment. Actions of an expert thus naturally encode part of their domain knowledge, and can help make inferences within the same domain: Knowing doctors try to prescribe the best treatment for their patients, we can tell treatments prescribed more frequently are likely to be more effective. Yet in machine learning, the fact that most decision-makers are experts is often overlooked, and “expertise” is seldom leveraged as an inductive bias. This is especially true for the literature on treatment effect estimation, where often the only assumption made about actions is that of overlap. In this paper, we argue that expertise—particularly the type of expertise the decision-makers of a domain are likely to have—can be informative in designing and selecting methods for treatment effect estimation. We formally define two types of expertise, predictive and prognostic, and demonstrate empirically that: (i) the prominent type of expertise in a domain significantly influences the performance of different methods in treatment effect estimation, and (ii) it is possible to predict the type of expertise present in a dataset, which can provide a quantitative basis for model selection.
****
Morteza Mardani,Jiaming Song,Jan Kautz,Arash Vahdat
A Variational Perspective on Solving Inverse Problems with Diffusion Models
Diffusion models have emerged as a key pillar of foundation models in visual domains. One of their critical applications is to universally solve different downstream inverse tasks via a single diffusion prior without re-training for each task. Most inverse tasks can be formulated as inferring a posterior distribution over data (e.g., a full image) given a measurement (e.g., a masked image). This is however challenging in diffusion models since the nonlinear and iterative nature of the diffusion process renders the posterior intractable. To cope with this challenge, we propose a variational approach that by design seeks to approximate the true posterior distribution. We show that our approach naturally leads to regularization by denoising diffusion process (RED-diff) where denoisers at different timesteps concurrently impose different structural constraints over the image. To gauge the contribution of denoisers from different timesteps, we propose a weighting mechanism based on signal-to-noise-ratio (SNR). Our approach provides a new variational perspective for solving inverse problems with diffusion models, allowing us to formulate sampling as stochastic optimization, where one can simply apply off-the-shelf solvers with lightweight iterates. Our experiments for image restoration tasks such as inpainting and superresolution demonstrate th

e strengths of our method compared with state-of-the-art sampling-based diffusion models.
****
Thiziri Nait Saada,Alireza Naderi,Jared Tanner
Beyond IID weights: sparse and low-rank deep Neural Networks are also Gaussian Processes
The infinitely wide neural network has been proven a useful and manageable mathematical model that enables the understanding of many phenomena appearing in deep learning. One example is the convergence of random deep networks to Gaussian processes that enables a rigorous analysis of the way the choice of activation function and network weights impacts the training dynamics. In this paper, we extend the seminal proof of Matthews et al., 2018 to a larger class of initial weight distributions (which we call pseudo-iid), including the established cases of iid and orthogonal weights, as well as the emerging low-rank and structured sparse settings celebrated for their computational speed-up benefits. We show that fully-connected and convolutional networks initialised with pseudo-iid distributions are all effectively equivalent up to their variance. Using our results, one can identify the Edge of Chaos for a broader class of neural networks and tune them at criticality in order to enhance their training. Moreover, they enable the posterior distribution of Bayesian Neural Networks to be tractable across these various initialization schemes.
****
Sara Klein,Simon Weissmann,Leif Döring
Beyond Stationarity: Convergence Analysis of Stochastic Softmax Policy Gradient Methods
Markov Decision Processes (MDPs) are a formal framework for modeling and solving sequential decision-making problems. In finite time horizons such problems are relevant for instance for optimal stopping or specific supply chain problems, but also in the training of large language models. In contrast to infinite horizon MDPs optimal policies are not stationary, policies must be learned for every single epoch. In practice all parameters are often trained simultaneously, ignoring the inherent structure suggested by dynamic programming. This paper introduces a combination of dynamic programming and policy gradient called dynamical policy gradient, where the parameters are trained backwards in time.

For the tabular softmax parametrisation we carry out the convergence analysis for simultaneous and dynamic policy gradient towards global optima, both in the exact and sampled gradient settings without regularisation. It turns out that the use of dynamic policy gradient training much better exploits the structure of finite-time problems which is reflected in improved convergence bounds.
****
Lukas Struppek,Dominik Hintersdorf,Kristian Kersting
Be Careful What You Smooth For: Label Smoothing Can Be a Privacy Shield but Also a Catalyst for Model Inversion Attacks
Label smoothing – using softened labels instead of hard ones – is a widely adopted regularization method for deep learning, showing diverse benefits such as enhanced generalization and calibration. Its implications for preserving model privacy, however, have remained unexplored. To fill this gap, we investigate the impact of label smoothing on model inversion attacks (MIAs), which aim to generate class-representative samples by exploiting the knowledge encoded in a classifier, thereby inferring sensitive information about its training data. Through extensive analyses, we uncover that traditional label smoothing fosters MIAs, thereby increasing a model's privacy leakage. Even more, we reveal that smoothing with negative factors counters this trend, impeding the extraction of class-related information and leading to privacy preservation, beating state-of-the-art defenses. This establishes a practical and powerful novel way for enhancing model resilience against MIAs.
****
Man Yao,JiaKui Hu,Tianxiang Hu,Yifan Xu,Zhaokun Zhou,Yonghong Tian,Bo XU,Guoqi Li

Spike-driven Transformer V2: Meta Spiking Neural Network Architecture Inspiring the Design of Next-generation Neuromorphic Chips
Neuromorphic computing, which exploits Spiking Neural Networks (SNNs) on neuromorphic chips, is a promising energy-efficient alternative to traditional AI. CNN-based SNNs are the current mainstream of neuromorphic computing. By contrast, no neuromorphic chips are designed especially for Transformer-based SNNs, which have just emerged, and their performance is only on par with CNN-based SNNs, offering no distinct advantage. In this work, we propose a general Transformer-based SNN architecture, termed as ``Meta-SpikeFormer", whose goals are: (1) *Lower-power*, supports the spike-driven paradigm that there is only sparse addition in the network; (2) *Versatility*, handles various vision tasks; (3) *High-performance*, shows overwhelming performance advantages over CNN-based SNNs; (4) *Meta-architecture*, provides inspiration for future next-generation Transformer-based neuromorphic chip designs. Specifically, we extend the Spike-driven Transformer in \citet{yao2023spike} into a meta architecture, and explore the impact of structure, spike-driven self-attention, and skip connection on its performance. On ImageNet-1K, Meta-SpikeFormer achieves 80.0\% top-1 accuracy (55M), surpassing the current state-of-the-art (SOTA) SNN baselines (66M) by 3.7\%. This is the first direct training SNN backbone that can simultaneously supports classification, detection, and segmentation, obtaining SOTA results in SNNs. Finally, we discuss the inspiration of the meta SNN architecture for neuromorphic chip design.
****
Edouard YVINEC,Arnaud Dapogny,Kevin Bailly
Network Memory Footprint Compression Through Jointly Learnable Codebooks and Mappings
The massive interest in deep neural networks (DNNs) for both computer vision and natural language processing has been sparked by the growth in computational power. However, this led to an increase in the memory footprint, to a point where it can be challenging to simply load a model on commodity devices such as mobile phones. To address this limitation, quantization is a favored solution as it maps high precision tensors to a low precision, memory efficient format. In terms of memory footprint reduction, its most effective variants are based on codebooks. These methods, however, suffer from two limitations. First, they either define a single codebook for each tensor, or use a memory-expensive mapping to multiple codebooks. Second, gradient descent optimization of the mapping favors jumps toward extreme values, hence not defining a proximal search. In this work, we propose to address these two limitations. First, we initially group similarly distributed neurons and leverage the re-ordered structure to either apply different scale factors to the different groups, or map weights that fall in these groups to several codebooks, without any mapping overhead. Second, stemming from this initialization, we propose a joint learning of the codebook and weight mappings that bears similarities with recent gradient-based post-training quantization techniques. Third, drawing estimation from straight-through estimation techniques, we introduce a novel gradient update definition to enable a proximal search of the codebooks and their mappings. The proposed jointly learnable codebooks and mappings (JLCM) method allows a very efficient approximation of any DNN: as such, a Llama 7B can be compressed down to 2Go and loaded on 5-year-old smartphones.
****
Zohar Rimon,Tom Jurgenson,Orr Krupnik,Gilad Adler,Aviv Tamar
MAMBA: an Effective World Model Approach for Meta-Reinforcement Learning
Meta-reinforcement learning (meta-RL) is a promising framework for tackling challenging domains requiring efficient exploration. Existing meta-RL algorithms are characterized by low sample efficiency, and mostly focus on low-dimensional task distributions. In parallel, model-based RL methods have been successful in solving partially observable MDPs, of which meta-RL is a special case.
In this work, we leverage this success and propose a new model-based approach to meta-RL, based on elements from existing state-of-the-art model-based and meta-RL methods. We demonstrate the effectiveness of our approach on common meta-RL benchmark domains, attaining greater return with better sample efficiency (up to $15\times$) while requiring very little hyperparameter tuning. In addition, we v

alidate our approach on a slate of more challenging, higher-dimensional domains, taking a step towards real-world generalizing agents.
****
Neta Shaul,Juan Perez,Ricky T. Q. Chen,Ali Thabet,Albert Pumarola,Yaron Lipman
Bespoke Solvers for Generative Flow Models
Diffusion or flow-based models are powerful generative paradigms that are notoriously hard to sample as samples are defined as solutions to high-dimensional Ordinary or Stochastic Differential Equations (ODEs/SDEs) which require a large Number of Function Evaluations (NFE) to approximate well. Existing methods to alleviate the costly sampling process include model distillation and designing dedicated ODE solvers. However, distillation is costly to train and sometimes can deteriorate quality, while dedicated solvers still require relatively large NFE to produce high quality samples. In this paper we introduce ``Bespoke solvers'', a novel framework for constructing custom ODE solvers tailored to the ODE of a given pre-trained flow model. Our approach optimizes an order consistent and parameter-efficient solver (e.g., with 80 learnable parameters), is trained for roughly 1\% of the GPU time required for training the pre-trained model, and significantly improves approximation and generation quality compared to dedicated solvers. For example, a Bespoke solver for a CIFAR10 model produces samples with Fréchet Inception Distance (FID) of 2.73 with 10 NFE, and gets to 1\% of the Ground Truth (GT) FID (2.59) for this model with only 20 NFE. On the more challenging ImageNet-64$\times$64, Bespoke samples at 2.2 FID with 10 NFE, and gets within 2\% of GT FID (1.71) with 20 NFE.
****
Soumyadeep Pal,Yuguang Yao,Ren Wang,Bingquan Shen,Sijia Liu
Backdoor Secrets Unveiled: Identifying Backdoor Data with Optimized Scaled Prediction Consistency
Modern machine learning (ML) systems demand substantial training data, often resorting to external sources. Nevertheless, this practice renders them vulnerable to backdoor poisoning attacks. Prior backdoor defense strategies have primarily focused on the identification of backdoored models or poisoned data characteristics, typically operating under the assumption of access to clean data. In this work, we delve into a relatively underexplored challenge: the automatic identification of backdoor data within a poisoned dataset, all under realistic conditions, *i.e.*, without the need for additional clean data or without manually defining a threshold for backdoor detection. We draw an inspiration from the scaled prediction consistency (SPC) technique, which exploits the prediction invariance of poisoned data to an input scaling factor. Based on this, we pose the backdoor data identification problem as a hierarchical data splitting optimization problem, leveraging a novel SPC-based loss function as the primary optimization objective. Our innovation unfolds in several key aspects. First, we revisit the vanilla SPC method, unveiling its limitations in addressing the proposed backdoor identification problem. Subsequently, we develop a bi-level optimization-based approach to precisely identify backdoor data by minimizing the advanced SPC loss. Finally, we demonstrate the efficacy of our proposal against a spectrum of backdoor attacks, encompassing basic label-corrupted attacks as well as more sophisticated clean-label attacks, evaluated across various benchmark datasets. Experiment results show that our approach often surpasses the performance of current baselines in identifying backdoor data points, resulting in about 4\%-36\% improvement in average AUROC. Codes are available at https://github.com/OPTML-Group/BackdoorMSPC.
****
Xuxi Chen,Yu Yang,Zhangyang Wang,Baharan Mirzasoleiman
Data Distillation Can Be Like Vodka: Distilling More Times For Better Quality
Dataset distillation aims to minimize the time and memory needed for training deep networks on large datasets, by creating a small set of synthetic images that has a similar generalization performance to that of the full dataset. However, current dataset distillation techniques fall short, showing a notable performance gap compared to training on the original data. In this work, we are the first to argue that the use of only one synthetic subset for distillation may not yield

optimal generalization performance. This is because the training dynamics of de
ep networks drastically changes during training. Therefore, multiple synthetic s
ubsets are required to capture the dynamics of training in different stages. To
address this issue, we propose Progressive Dataset Distillation (PDD). PDD synth
esizes multiple small sets of synthetic images, each conditioned on the previous
 sets, and trains the model on the cumulative union of these subsets without req
uiring additional training time. Our extensive experiments show that PDD can eff
ectively improve the performance of existing dataset distillation methods by up
to 4.3%. In addition, our method for the first time enables generating considera
bly larger synthetic datasets. Our codes are available at https://github.com/VIT
A-Group/ProgressiveDD.
****
Xiaohuan Pei,Yanxi Li,Minjing Dong,Chang Xu
Neural Architecture Retrieval
With the increasing number of new neural architecture designs and substantial ex
isting neural architectures, it becomes difficult for the researchers to situate
 their contributions compared with existing neural architectures or establish th
e connections between their designs and other relevant ones. To discover similar
 neural architectures in an efficient and automatic manner, we define a new prob
lem Neural Architecture Retrieval which retrieves a set of existing neural archi
tectures which have similar designs to the query neural architecture. Existing g
raph pre-training strategies cannot address the computational graph in neural ar
chitectures due to the graph size and motifs. To fulfill this potential, we prop
ose to divide the graph into motifs which are used to rebuild the macro graph to
 tackle these issues, and introduce multi-level contrastive learning to achieve
accurate graph representation learning. Extensive evaluations on both human-desi
gned and synthesized neural architectures demonstrate the superiority of our alg
orithm. Such a dataset which contains 12k real-world network architectures, as w
ell as their embedding, is built for neural architecture retrieval.
****
Titas Anciukevi■ius,Fabian Manhardt,Federico Tombari,Paul Henderson
Denoising Diffusion via Image-Based Rendering
Generating 3D scenes is a challenging open problem, which requires synthesizing
plausible content that is fully consistent in 3D space. While recent methods suc
h as neural radiance fields excel at view synthesis and 3D reconstruction, they
cannot synthesize plausible details in unobserved regions since they lack a gene
rative capability. Conversely, existing generative methods are typically not cap
able of reconstructing detailed, large-scale scenes in the wild, as they use lim
ited-capacity 3D scene representations, require aligned camera poses, or rely on
 additional regularizers. In this work, we introduce the first diffusion model a
ble to perform fast, detailed reconstruction and generation of real-world 3D sce
nes. To achieve this, we make three contributions. First, we introduce a new neu
ral scene representation, IB-planes, that can efficiently and accurately represe
nt large 3D scenes, dynamically allocating more capacity as needed to capture de
tails visible in each image. Second, we propose a denoising-diffusion framework
to learn a prior over this novel 3D scene representation, using only 2D images w
ithout the need for any additional supervision signal such as masks or depths.
This supports 3D reconstruction and generation in a unified architecture. Third,
 we develop a principled approach to avoid trivial 3D solutions when integrating
 the image-based rendering with the diffusion model, by dropping out representat
ions of some images. We evaluate the model on several challenging datasets of re
al and synthetic images, and demonstrate superior results on generation, novel v
iew synthesis and 3D reconstruction.
****
Nishant Yadav,Nicholas Monath,Manzil Zaheer,Rob Fergus,Andrew McCallum
Adaptive Retrieval and Scalable Indexing for k-NN Search with Cross-Encoders
Cross-encoder (CE) models which compute similarity by jointly encoding a query-i
tem pair perform better than using dot-product with embedding-based models (dual
-encoders) at estimating query-item relevance. Existing approaches perform k-NN
search with cross-encoders by approximating the CE similarity with a vector embe

dding space fit either with dual-encoders (DE) or CUR matrix factorization. DE-b ased retrieve-and-rerank approaches suffer from poor recall as DE generalizes po orly to new domains and the test-time retrieval with DE is decoupled from the CE . While CUR-based approaches can be more accurate than the DE-based retrieve-and -rerank approach, such approaches require a prohibitively large number of CE cal ls to compute item embeddings, thus making it impractical for deployment at scal e. In this paper, we address these shortcomings with our proposed sparse-matrix factorization based method that efficiently computes latent query and item repre sentations to approximate CE scores and performs k-NN search with the approximat e CE similarity. In an offline indexing stage, we compute item embeddings by fac torizing a sparse matrix containing query-item CE scores for a set of train quer ies. Our method produces a high-quality approximation while requiring only a fra ction of CE similarity calls as compared to CUR-based methods, and allows for le veraging DE models to initialize the embedding space while avoiding compute- and resource-intensive finetuning of DE via distillation. At test time, we keep ite m embeddings fixed and perform retrieval over multiple rounds, alternating betwe en a) estimating the test query embedding by minimizing error in approximating C E scores of items retrieved thus far, and b) using the updated test query embedd ing for retrieving more items in the next round. Our proposed k-NN search method can achieve up to 5 and 54 improvement in k-NN recall for k=1 and 100 respectiv ely over the widely-used DE-based retrieve-and-rerank approach. Furthermore, our proposed approach to index the items by aligning item embeddings with the CE ac hieves up to 100x and 5x speedup over CUR-based and dual-encoder distillation ba sed approaches respectively while matching or improving k-NN search recall over baselines.
****
Seon-Ho Lee,Nyeong-Ho Shin,Chang-Su Kim
Unsupervised Order Learning
A novel clustering algorithm for orderable data, called unsupervised order learn ing (UOL), is proposed in this paper. First, we develop the ordered k-means to group objects into ordered clusters by reducing the deviation of an object from consecutive clusters. Then, we train a network to construct an embedding space, in which objects are sorted compactly along a chain of line segments, determine d by the cluster centroids. We alternate the clustering and the network training until convergence. Moreover, we perform unsupervised rank estimation via a simp le nearest neighbor search in the embedding space. Extensive experiments on vari ous orderable datasets demonstrate that UOL provides reliable ordered clustering results and decent rank estimation performances with no supervision. The source codes are available at https://github.com/seon92/UOL.
****
Zhiyu Zhu,Xinyi Wang,Zhibo Jin,Jiayu Zhang,Huaming Chen
Enhancing Transferable Adversarial Attacks on Vision Transformers through Gradie nt Normalization Scaling and High-Frequency Adaptation
Vision Transformers (ViTs) have been widely used in various domains. Similar to Convolutional Neural Networks (CNNs), ViTs are prone to the impacts of adversari al samples, raising security concerns in real-world applications. As one of the most effective black-box attack methods, transferable attacks can generate adver sarial samples on surrogate models to directly attack the target model without a ccessing the parameters. However, due to the distinct internal structures of ViT s and CNNs, adversarial samples constructed by traditional transferable attack m ethods may not be applicable to ViTs. Therefore, it is imperative to propose mor e effective transferability attack methods to unveil latent vulnerabilities in V iTs. Existing methods have found that applying gradient regularization to extrem e gradients across different functional regions in the transformer structure can enhance sample transferability. However, in practice, substantial gradient disp arities exist even within the same functional region across different layers. Fu rthermore, we find that mild gradients therein are the main culprits behind redu ced transferability. In this paper, we introduce a novel Gradient Normalization Scaling method for fine-grained gradient editing to enhance the transferability of adversarial attacks on ViTs. More importantly, we highlight that ViTs, unlike

traditional CNNs, exhibit distinct attention regions in the frequency domain. Leveraging this insight, we delve into exploring the frequency domain to further enhance the algorithm's transferability. Through extensive experimentation on various ViT variants and traditional CNN models, we substantiate that the new approach achieves state-of-the-art performance, with an average performance improvement of 33.54\% and 42.05\% on ViT and CNN models, respectively. Our code is available at: https://github.com/LMBTough/GNS-HFA.
****
Hang Yin,Zihao Wang,Yangqiu Song
Rethinking Complex Queries on Knowledge Graphs with Neural Link Predictors
Reasoning on knowledge graphs is a challenging task because it utilizes observed information to predict the missing one. Particularly, answering complex queries based on first-order logic is one of the crucial tasks to verify learning to reason abilities for generalization and composition.
Recently, the prevailing method is query embedding which learns the embedding of a set of entities and treats logic operations as set operations and has shown great empirical success. Though there has been much research following the same formulation, many of its claims lack a formal and systematic inspection. In this paper, we rethink this formulation and justify many of the previous claims by characterizing the scope of queries investigated previously and precisely identifying the gap between its formulation and its goal, as well as providing complexity analysis for the currently investigated queries. Moreover, we develop a new dataset containing ten new types of queries with features that have never been considered and therefore can provide a thorough investigation of complex queries. Finally, we propose a new neural-symbolic method, Fuzzy Inference with Truth value (FIT), where we equip the neural link predictors with fuzzy logic theory to support end-to-end learning using complex queries with provable reasoning capability. Empirical results show that our method outperforms previous methods significantly in the new dataset and also surpasses previous methods in the existing dataset at the same time.
****
Greg Yang,Dingli Yu,Chen Zhu,Soufiane Hayou
Tensor Programs VI: Feature Learning in Infinite Depth Neural Networks
Empirical studies have consistently demonstrated that increasing the size of neural networks often yields superior performance in practical applications. However, there is a lack of consensus regarding the appropriate scaling strategy, particularly when it comes to increasing the depth of neural networks. In practice, excessively large depths can lead to model performance degradation. In this paper, we introduce Depth-μP, a principled approach for depth scaling, allowing for the training of arbitrarily deep architectures while maximizing feature learning and diversity among nearby layers. Our method involves dividing the contribution of each residual block and the parameter update by the square root of the depth. Through the use of Tensor Programs, we rigorously establish the existence of a limit for infinitely deep neural networks under the proposed scaling scheme. This scaling strategy ensures more stable training for deep neural networks and guarantees the transferability of hyperparameters from shallow to deep models. To substantiate the efficacy of our scaling method, we conduct empirical validation on neural networks with depths up to 2^{10}.
****
Changli Tang,Wenyi Yu,Guangzhi Sun,Xianzhao Chen,Tian Tan,Wei Li,Lu Lu,Zejun MA,Chao Zhang
SALMONN: Towards Generic Hearing Abilities for Large Language Models
Hearing is arguably an essential ability of artificial intelligence (AI) agents in the physical world, which refers to the perception and understanding of general auditory information consisting of at least three types of sounds: speech, audio events, and music. In this paper, we propose SALMONN, a speech audio language music open neural network, built by integrating a pre-trained text-based large language model (LLM) with speech and audio encoders into a single multimodal model. SALMONN enables the LLM to directly process and understand general audio inputs and achieve competitive performances on a number of speech and audio tasks

used in training, such as
automatic speech recognition and translation, auditory-information-based question answering, emotion recognition, speaker verification, and music and audio captioning etc. SALMONN also has a diverse set of emergent abilities unseen in the training, which includes but is not limited to speech translation to untrained languages, speech-based slot filling, spoken-query-based question answering, audio-based storytelling, and speech audio co-reasoning etc. The presence of cross-modal emergent abilities is studied, and a novel few-shot activation tuning approach is proposed to activate such abilities. To our knowledge, SALMONN is the first model of its type and can be regarded as a step towards AI with generic hearing abilities. The source code, model checkpoints and data are available at https://github.com/bytedance/SALMONN.
****
Sara Ghazanfari,Alexandre Araujo,Prashanth Krishnamurthy,Farshad Khorrami,Siddharth Garg
LipSim: A Provably Robust Perceptual Similarity Metric
Recent years have seen growing interest in developing and applying perceptual similarity metrics. Research has shown the superiority of perceptual metrics over pixel-wise metrics in aligning with human perception and serving as a proxy for the human visual system.
On the other hand, as perceptual metrics rely on neural networks, there is a growing concern regarding their resilience, given the established vulnerability of neural networks to adversarial attacks. It is indeed logical to infer that perceptual metrics may inherit both the strengths and shortcomings of neural networks.
In this work, we demonstrate the vulnerability of state-of-the-art perceptual similarity metrics based on an ensemble of ViT-based feature extractors to adversarial attacks. We then propose a framework to train a robust perceptual similarity metric called LipSim (Lipschitz Similarity Metric) with provable guarantees.
By leveraging 1-Lipschitz neural networks as the backbone, LipSim provides guarded areas around each data point and certificates for all perturbations within an ℓ_2 ball. Finally, a comprehensive set of experiments shows the performance of LipSim in terms of natural and certified scores and on the image retrieval application.
****
Kwangjun Ahn,Xiang Cheng,Minhak Song,Chulhee Yun,Ali Jadbabaie,Suvrit Sra
Linear attention is (maybe) all you need (to understand Transformer optimization)
Transformer training is notoriously difficult, requiring a careful design of optimizers and use of various heuristics. We make progress towards understanding the subtleties of training Transformers by carefully studying a simple yet canonical linearized *shallow* Transformer model. Specifically, we train linear Transformers to solve regression tasks, inspired by J. von Oswald et al. (ICML 2023), and K. Ahn et al. (NeurIPS 2023). Most importantly, we observe that our proposed linearized models can reproduce several prominent aspects of Transformer training dynamics. Consequently, the results obtained in this paper suggest that a simple linearized Transformer model could actually be a valuable, realistic abstraction for understanding Transformer optimization.
****
Anand Siththaranjan,Cassidy Laidlaw,Dylan Hadfield-Menell
Understanding Hidden Context in Preference Learning: Consequences for RLHF
In practice, preference learning from human feedback depends on incomplete data with hidden context. Hidden context refers to data that affects the feedback received, but which is not represented in the data used to train a preference model. This captures common issues of data collection, such as having human annotators with varied preferences, cognitive processes that result in seemingly irrational behavior, and combining data labeled according to different criteria. We prove that standard applications of preference learning, including reinforcement learning from human feedback (RLHF), implicitly aggregate over hidden contexts according to a well-known voting rule called *Borda count*. We show this can produce

counter-intuitive results that are very different from other methods which impl icitly aggregate via expected utility. Furthermore, our analysis formalizes the way that preference learning from users with diverse values tacitly implements a social choice function. A key implication of this result is that annotators hav e an incentive to misreport their preferences in order to influence the learned model, leading to vulnerabilities in the deployment of RLHF. As a step towards m itigating these problems, we introduce a class of methods called *distributional preference learning* (DPL). DPL methods estimate a distribution of possible sco re values for each alternative in order to better account for hidden context. Ex perimental results indicate that applying DPL to RLHF for LLM chatbots identifie s hidden context in the data and significantly reduces subsequent jailbreak vuln erability.
****
Animesh Basak Chowdhury,Marco Romanelli,Benjamin Tan,Ramesh Karri,Siddharth Garg
Retrieval-Guided Reinforcement Learning for Boolean Circuit Minimization
Logic synthesis, a pivotal stage in chip design, entails optimizing chip specifi cations encoded in hardware description languages like Verilog into highly effic ient implementations using Boolean logic gates. The process involves a sequentia l application of logic minimization heuristics (``synthesis recipe"), with their arrangement significantly impacting crucial metrics such as area and delay. Add ressing the challenge posed by the broad spectrum of hardware design complexitie s — from variations of past designs (e.g., adders and multipliers) to entirely n ovel configurations (e.g., innovative processor instructions) — requires a nuanc ed 'synthesis recipe' guided by human expertise and intuition. This study conduc ts a thorough examination of learning and search techniques for logic synthesis, unearthing a surprising revelation: pre-trained agents, when confronted with en tirely novel designs, may veer off course, detrimentally affecting the search tr ajectory. We present ABC-RL, a meticulously tuned α parameter that adeptl y adjusts recommendations from pre-trained agents during the search process. Com puted based on similarity scores through nearest neighbor retrieval from the tra ining dataset, ABC-RL yields superior synthesis recipes tailored for a wide arra y of hardware designs. Our findings showcase substantial enhancements in the Qua lity of Result (QoR) of synthesized circuits, boasting improvements of up to 24. 8\% compared to state-of-the-art techniques. Furthermore, ABC-RL achieves an imp ressive up to 9x reduction in runtime (iso-QoR) when compared to current state-o f-the-art methodologies.
****
Giulio Franzese,Mustapha BOUNOUA,Pietro Michiardi
MINDE: Mutual Information Neural Diffusion Estimation
In this work we present a new method for the estimation of Mutual Information (M I) between random variables. Our approach is based on an original interpretation of the Girsanov theorem, which allows us to use score-based diffusion models to estimate the KL divergence between two densities as a difference between their score functions. As a by-product, our method also enables the estimation of the entropy of random variables.
Armed with such building blocks, we present a general recipe to measure MI, whic h unfolds in two directions: one uses conditional diffusion process, whereas the other uses joint diffusion processes that allow simultaneous modelling of two random variables.
Our results, which derive from a thorough experimental protocol over all the var iants of our approach, indicate that our method is more accurate than the main a lternatives from the literature, especially for challenging distributions. Furth ermore, our methods pass MI self-consistency tests, including data processing an d additivity under independence, which instead are a pain-point of existing meth ods
****
Biswadeep Chakraborty,Beomseok Kang,Harshit Kumar,Saibal Mukhopadhyay
Sparse Spiking Neural Network: Exploiting Heterogeneity in Timescales for Prunin g Recurrent SNN
Recurrent Spiking Neural Networks (RSNNs) have emerged as a computationally effi

cient and brain-inspired machine learning model. The design of sparse RSNNs with fewer neurons and synapses helps reduce the computational complexity of RSNNs. Traditionally, sparse SNNs are obtained by first training a dense and complex SNN for a target task and, next, eliminating neurons with low activity (activity-based pruning) while maintaining task performance. In contrast, this paper presents a task-agnostic methodology for designing sparse RSNNs by pruning an untrained (arbitrarily initialized) large model.
We introduce a novel Lyapunov Noise Pruning (LNP) algorithm that uses graph sparsification methods and utilizes Lyapunov exponents to design a stable sparse RSNN from an untrained RSNN. We show that the LNP can leverage diversity in neuronal timescales to design a sparse Heterogeneous RSNN (HRSNN). Further, we show that the same sparse HRSNN model can be trained for different tasks, such as image classification and time-series prediction. The experimental results show that, in spite of being task-agnostic, LNP increases computational efficiency (fewer neurons and synapses) and prediction performance of RSNNs compared to traditional activity-based pruning of trained dense models.
****
Guocheng Qian,Jinjie Mai,Abdullah Hamdi,Jian Ren,Aliaksandr Siarohin,Bing Li,Hsin-Ying Lee,Ivan Skorokhodov,Peter Wonka,Sergey Tulyakov,Bernard Ghanem
Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors
We present ``Magic123'', a two-stage coarse-to-fine approach for high-quality, textured 3D mesh generation from a single image in the wild using *both 2D and 3D priors*. In the first stage, we optimize a neural radiance field to produce a coarse geometry. In the second stage, we adopt a memory-efficient differentiable mesh representation to yield a high-resolution mesh with a visually appealing texture. In both stages, the 3D content is learned through reference-view supervision and novel-view guidance by a joint 2D and 3D diffusion prior. We introduce a trade-off parameter between the 2D and 3D priors to control the details and 3D consistencies of the generation. Magic123 demonstrates a significant improvement over previous image-to-3D techniques, as validated through extensive experiments on diverse synthetic and real-world images.
****
Hyunju Kang,Geonhee Han,Hogun Park
UNR-Explainer: Counterfactual Explanations for Unsupervised Node Representation Learning Models
Node representation learning, such as Graph Neural Networks (GNNs), has become one of the important learning methods in machine learning, and the demand for reliable explanation generation is growing. Despite extensive research on explanation generation for supervised node representation learning, explaining unsupervised models has been less explored. To address this gap, we propose a method for generating counterfactual (CF) explanations in unsupervised node representation learning, aiming to identify the most important subgraphs that cause a significant change in the k-nearest neighbors of a node of interest in the learned embedding space upon perturbation. The k-nearest neighbor-based CF explanation method provides simple, yet pivotal, information for understanding unsupervised downstream tasks, such as top-k link prediction and clustering. Furthermore, we introduce a Monte Carlo Tree Search (MCTS)-based explainability method for generating expressive CF explanations for **U**nsupervised **N**ode **R**epresentation learning methods, which we call **UNR-Explainer**. The proposed method demonstrates improved performance on six datasets for both unsupervised GraphSAGE and DGI.
****
Xufeng Cai,Ahmet Alacaoglu,Jelena Diakonikolas
Variance Reduced Halpern Iteration for Finite-Sum Monotone Inclusions
Machine learning approaches relying on such criteria as adversarial robustness or multi-agent settings have raised the need for solving game-theoretic equilibrium problems. Of particular relevance to these applications are methods targeting finite-sum structure, which generically arises in empirical variants of learning problems in these contexts. Further, methods with computable approximation err

ors are highly desirable, as they provide verifiable exit criteria. Motivated by these applications, we study finite-sum monotone inclusion problems, which model broad classes of equilibrium problems. Our main contributions are variants of the classical Halpern iteration that employ variance reduction to obtain improved complexity guarantees in which n component operators in the finite sum are ``on average'' either cocoercive or Lipschitz continuous and monotone, with parameter L. The resulting oracle complexity of our methods, which provide guarantees for the last iterate and for a (computable) operator norm residual, is $\widetilde{\mathcal{O}}(n + \sqrt{n}L\varepsilon^{-1})$, which improves upon existing methods by a factor up to $\sqrt{n}$. This constitutes the first variance reduction-type result for general finite-sum monotone inclusions and for more specific problems such as convex-concave optimization when operator norm residual is the optimality measure. We further argue that, up to poly-logarithmic factors, this complexity is unimprovable in the monotone Lipschitz setting; i.e., the provided result is near-optimal.
****
Dongming Wu,Jiahao Chang,Fan Jia,Yingfei Liu,Tiancai Wang,Jianbing Shen
TopoMLP: A Simple yet Strong Pipeline for Driving Topology Reasoning
Topology reasoning aims to comprehensively understand road scenes and present drivable routes in autonomous driving. It requires detecting road centerlines (lane) and traffic elements, further reasoning their topology relationship, \textit{i.e.}, lane-lane topology, and lane-traffic topology. In this work, we first present that the topology score relies heavily on detection performance on lane and traffic elements. Therefore, we introduce a powerful 3D lane detector and an improved 2D traffic element detector to extend the upper limit of topology performance. Further, we propose TopoMLP, a simple yet high-performance pipeline for driving topology reasoning. Based on the impressive detection performance, we develop two simple MLP-based heads for topology generation. TopoMLP achieves state-of-the-art performance on OpenLane-V2 dataset, \textit{i.e.}, 41.2\% OLS with ResNet-50 backbone. It is also the 1st solution for 1st OpenLane Topology in Autonomous Driving Challenge. We hope such simple and strong pipeline can provide some new insights to the community. Code is at https://github.com/wudongming97/TopoMLP.
****
Sifan Zhou,Liang Li,Xinyu Zhang,Bo Zhang,Shipeng Bai,Miao Sun,Ziyu Zhao,Xiaobo Lu,Xiangxiang Chu
LiDAR-PTQ: Post-Training Quantization for Point Cloud 3D Object Detection
Due to highly constrained computing power and memory, deploying 3D lidar-based detectors on edge devices equipped in autonomous vehicles and robots poses a crucial challenge. Being a convenient and straightforward model compression approach, Post-Training Quantization (PTQ) has been widely adopted in 2D vision tasks. However, applying it directly to 3D lidar-based tasks inevitably leads to performance degradation. As a remedy, we propose an effective PTQ method called LiDAR-PTQ, which is particularly curated for 3D lidar detection (both SPConv-based and SPConv-free). Our LiDAR-PTQ features three main components, (1) a sparsity-based calibration method to determine the initialization of quantization parameters, (2) an adaptive rounding-to-nearest operation to minimize the layerwise reconstruction error, (3) a Task-guided Global Positive Loss (TGPL) to reduce the disparity between the final predictions before and after quantization. Extensive experiments demonstrate that our LiDAR-PTQ can achieve state-of-the-art quantization performance when applied to CenterPoint (both Pillar-based and Voxel-based). To our knowledge, for the very first time in lidar-based 3D detection tasks, the PTQ INT8 model's accuracy is almost the same as the FP32 model while enjoying 3X inference speedup. Moreover, our LiDAR-PTQ is cost-effective being 6X faster than the quantization-aware training method. The code will be released.
****
Neel Jain,Ping-yeh Chiang,Yuxin Wen,John Kirchenbauer,Hong-Min Chu,Gowthami Somepalli,Brian R. Bartoldson,Bhavya Kailkhura,Avi Schwarzschild,Aniruddha Saha,Micah Goldblum,Jonas Geiping,Tom Goldstein
NEFTune: Noisy Embeddings Improve Instruction Finetuning

We show that language model finetuning can be improved, sometimes dramatically, with a simple augmentation.
NEFTune adds noise to the embedding vectors during training.
Standard finetuning of LLaMA-2-7B using Alpaca achieves 29.79\% on AlpacaEval, which rises to 64.69\% using noisy embeddings. NEFTune also improves over strong baselines on modern instruction datasets.
Models trained with Evol-Instruct see a 10\% improvement, with ShareGPT an 8\% improvement, and with OpenPlatypus an 8\% improvement.
Even powerful models further refined with RLHF such as LLaMA-2-Chat benefit from additional training with NEFTune. Particularly, we see these improvements on the conversational abilities of the instruction model and not on traditional tasks like those on the OpenLLM Leaderboard, where performance is the same.
****
Chongyi Zheng,Ruslan Salakhutdinov,Benjamin Eysenbach
Contrastive Difference Predictive Coding
Predicting and reasoning about the future lie at the heart of many time-series questions. For example, goal-conditioned reinforcement learning can be viewed as learning representations to predict which states are likely to be visited in the future. While prior methods have used contrastive predictive coding to model time series data, learning representations that encode long-term dependencies usually requires large amounts of data. In this paper, we introduce a temporal difference version of contrastive predictive coding that stitches together pieces of different time series data to decrease the amount of data required to learn predictions of future events. We apply this representation learning method to derive an off-policy algorithm for goal-conditioned RL. Experiments demonstrate that, compared with prior RL methods, ours achieves $2 \times$ median improvement in success rates and can better cope with stochastic environments. In tabular settings, we show that our method is about $20\times$ more sample efficient than the successor representation and $1500 \times$ more sample efficient than the standard (Monte Carlo) version of contrastive predictive coding.
****
Guozheng Ma,Lu Li,Sen Zhang,Zixuan Liu,Zhen Wang,Yixin Chen,Li Shen,Xueqian Wang,Dacheng Tao
Revisiting Plasticity in Visual Reinforcement Learning: Data, Modules and Training Stages
Plasticity, the ability of a neural network to evolve with new data, is crucial for high-performance and sample-efficient visual reinforcement learning (VRL). Although methods like resetting and regularization can potentially mitigate plasticity loss, the influences of various components within the VRL framework on the agent's plasticity are still poorly understood. In this work, we conduct a systematic empirical exploration focusing on three primary underexplored facets and derive the following insightful conclusions: (1) data augmentation is essential in maintaining plasticity; (2) the critic's plasticity loss serves as the principal bottleneck impeding efficient training; and (3) without timely intervention to recover critic's plasticity in the early stages, its loss becomes catastrophic. These insights suggest a novel strategy to address the high replay ratio (RR) dilemma, where exacerbated plasticity loss hinders the potential improvements of sample efficiency brought by increased reuse frequency. Rather than setting a static RR for the entire training process, we propose Adaptive RR, which dynamically adjusts the RR based on the critic’s plasticity level. Extensive evaluations indicate that Adaptive RR not only avoids catastrophic plasticity loss in the early stages but also benefits from more frequent reuse in later phases, resulting in superior sample efficiency.
****
Xiang Fu,Tian Xie,Andrew Scott Rosen,Tommi S. Jaakkola,Jake Allen Smith
MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design
Metal-organic frameworks (MOFs) are of immense interest in applications such as gas storage and carbon capture due to their exceptional porosity and tunable chemistry. Their modular nature has enabled the use of template-based methods to generate hypothetical MOFs by combining molecular building blocks in accordance wi

th known network topologies. However, the ability of these methods to identify t
op-performing MOFs is often hindered by the limited diversity of the resulting c
hemical space. In this work, we propose MOFDiff: a coarse-grained (CG) diffusion
 model that generates CG MOF structures through a denoising diffusion process ov
er the coordinates and identities of the building blocks. The all-atom MOF struc
ture is then determined through a novel assembly algorithm. As the diffusion mod
el generates 3D MOF structures by predicting scores in E(3), we employ equivaria
nt graph neural networks that respect the permutational and roto-translational s
ymmetries. We comprehensively evaluate our model's capability to generate valid
and novel MOF structures and its effectiveness in designing outstanding MOF mate
rials for carbon capture applications with molecular simulations.
****
Haoning Wu,Zicheng Zhang,Erli Zhang,Chaofeng Chen,Liang Liao,Annan Wang,Chunyi L
i,Wenxiu Sun,Qiong Yan,Guangtao Zhai,Weisi Lin
Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision
The rapid evolution of Multi-modality Large Language Models (MLLMs) has catalyze
d a shift in computer vision from specialized models to general-purpose foundati
on models. Nevertheless, there is still an inadequacy in assessing the abilities
 of MLLMs on **low-level visual perception and understanding**. To address this
gap, we present **Q-Bench**, a holistic benchmark crafted to systematically eval
uate potential abilities of MLLMs on three realms: low-level visual perception,
low-level visual description, and overall visual quality assessment. **_a)_** To
 evaluate the low-level **_perception_** ability, we construct the **LLVisionQA*
* dataset, consisting of 2,990 diverse-sourced images, each equipped with a huma
n-asked question focusing on its low-level attributes. We then measure the corre
ctness of MLLMs on answering these questions. **_b)_** To examine the **_descrip
tion_** ability of MLLMs on low-level information, we propose the **LLDescribe**
 dataset consisting of long expert-labelled *golden* low-level text descriptions
 on 499 images, and a GPT-involved comparison pipeline between outputs of MLLMs
and the *golden* descriptions. **_c)_** Besides these two tasks, we further meas
ure their visual quality **_assessment_** ability to align with human opinion sc
ores. Specifically, we design a softmax-based strategy that enables MLLMs to pre
dict *quantifiable* quality scores, and evaluate them on various existing image
quality assessment (IQA) datasets. Our evaluation across the three abilities con
firms that MLLMs possess preliminary low-level visual skills. However, these ski
lls are still unstable and relatively imprecise, indicating the need for specifi
c enhancements on MLLMs towards these abilities. We hope that our benchmark can
encourage the research community to delve deeper to discover and enhance these u
ntapped potentials of MLLMs.
****
Shikun Sun,Longhui Wei,Zhicai Wang,Zixuan Wang,Junliang Xing,Jia Jia,Qi Tian
Inner Classifier-Free Guidance and Its Taylor Expansion for Diffusion Models
Classifier-free guidance (CFG) is a pivotal technique for balancing the diversit
y and fidelity of samples in conditional diffusion models. This approach involve
s utilizing a single model to jointly optimize the conditional score predictor a
nd unconditional score predictor, eliminating the need for additional classifier
s. It delivers impressive results and can be employed for continuous and discret
e condition representations. However, when the condition is continuous, it promp
ts the question of whether the trade-off can be further enhanced. Our proposed i
nner classifier-free guidance (ICFG) provides an alternative perspective on the
CFG method when the condition has a specific structure, demonstrating that CFG r
epresents a first-order case of ICFG. Additionally, we offer a second-order impl
ementation, highlighting that even without altering the training policy, our sec
ond-order approach can introduce new valuable information and achieve an improve
d balance between fidelity and diversity for Stable Diffusion.
****
Yuying Ge,Sijie Zhao,Ziyun Zeng,Yixiao Ge,Chen Li,Xintao Wang,Ying Shan
Making LLaMA SEE and Draw with SEED Tokenizer
The great success of Large Language Models (LLMs) has expanded the potential of
multimodality, contributing to the gradual evolution of General Artificial Intel

ligence (AGI). A true AGI agent should not only possess the capability to perform predefined multi-tasks but also exhibit emergent abilities in an open-world context. However, despite the considerable advancements made by recent multimodal LLMs, they still fall short in effectively unifying comprehension and generation tasks, let alone open-world emergent abilities. We contend that the key to overcoming the present impasse lies in enabling text and images to be represented and processed interchangeably within a unified autoregressive Transformer. To this end, we introduce $\textbf{SEED}$, an elaborate image tokenizer that empowers LLMs with the ability to $\textbf{SEE}$ and $\textbf{D}$raw at the same time. We identify two crucial design principles: (1) Image tokens should be independent of 2D physical patch positions and instead be produced with a $\textit{1D causal dependency}$, exhibiting intrinsic interdependence that aligns with the left-to-right autoregressive prediction mechanism in LLMs. (2) Image tokens should capture $\textit{high-level semantics}$ consistent with the degree of semantic abstraction in words, and be optimized for both discriminativeness and reconstruction during the tokenizer training phase. With SEED tokens, LLM is able to perform scalable multimodal autoregression under its original training recipe, i.e., next-word prediction. SEED-LLaMA is therefore produced by large-scale pretraining and instruction tuning on the interleaved textual and visual data, demonstrating impressive performance on a broad range of multimodal comprehension and generation tasks. More importantly, SEED-LLaMA has exhibited compositional emergent abilities such as multi-turn in-context multimodal generation, acting like your AI assistant. The code (training and inference) and models are released in https://github.com/AILab-CVC/SEED.
****
Claudio Battiloro,Indro Spinelli,Lev Telyatnikov,Michael M. Bronstein,Simone Scardapane,Paolo Di Lorenzo
From Latent Graph to Latent Topology Inference: Differentiable Cell Complex Module
Latent Graph Inference (LGI) relaxed the reliance of Graph Neural Networks (GNNs) on a given graph topology by dynamically learning it. However, most of LGI methods assume to have a (noisy, incomplete, improvable, ...) input graph to rewire and can solely learn regular graph topologies. In the wake of the success of Topological Deep Learning (TDL), we study Latent Topology Inference (LTI) for learning higher-order cell complexes (with sparse and not regular topology) describing multi-way interactions between data points. To this aim, we introduce the Differentiable Cell Complex Module (DCM), a novel learnable function that computes cell probabilities in the complex to improve the downstream task. We show how to integrate DCM with cell complex message-passing networks layers and train it in an end-to-end fashion, thanks to a two-step inference procedure that avoids an exhaustive search across all possible cells in the input, thus maintaining scalability. Our model is tested on several homophilic and heterophilic graph datasets and it is shown to outperform other state-of-the-art techniques, offering significant improvements especially in cases where an input graph is not provided.
****
Edwin Zhang,Yujie Lu,Shinda Huang,William Yang Wang,Amy Zhang
Language Control Diffusion: Efficiently Scaling through Space, Time, and Tasks
Training generalist agents is difficult across several axes, requiring us to deal with high-dimensional inputs (space), long horizons (time), and generalization to novel tasks. Recent advances with architectures have allowed for improved scaling along one or two of these axes, but are still computationally prohibitive to use. In this paper, we propose to address all three axes by leveraging Language to Control Diffusion models as a hierarchical planner conditioned on language (LCD). We effectively and efficiently scale diffusion models for planning in extended temporal, state, and task dimensions to tackle long horizon control problems conditioned on natural language instructions, as a step towards generalist agents. Comparing LCD with other state-of-the-art models on the CALVIN language benchmark finds that LCD outperforms other SOTA methods in multi-task success rates, whilst improving inference speed over other comparable diffusion models by 3.3x~15x. We show that LCD can successfully leverage the unique strength of diffu

sion models to produce coherent long range plans while addressing their weakness in generating low-level details and control.
****
Galen Andrew,Peter Kairouz,Sewoong Oh,Alina Oprea,Hugh Brendan McMahan,Vinith Menon Suriyakumar
One-shot Empirical Privacy Estimation for Federated Learning
Privacy estimation techniques for differentially private (DP) algorithms are useful for comparing against analytical bounds, or to empirically measure privacy loss in settings where known analytical bounds are not tight. However, existing privacy auditing techniques usually make strong assumptions on the adversary (e.g., knowledge of intermediate model iterates or the training data distribution), are tailored to specific tasks, model architectures, or DP algorithm, and/or require retraining the model many times (typically on the order of thousands). These shortcomings make deploying such techniques at scale difficult in practice, especially in federated settings where model training can take days or weeks. In this work, we present a novel “one-shot” approach that can systematically address these challenges, allowing efficient auditing or estimation of the privacy loss of a model during the same, single training run used to fit model parameters, and without requiring any a priori knowledge about the model architecture, task, or DP algorithm. We show that our method provides provably correct estimates for the privacy loss under the Gaussian mechanism, and we demonstrate its performance on a well-established FL benchmark dataset under several adversarial threat models.
****
Xilie Xu,Jingfeng Zhang,Mohan Kankanhalli
AutoLoRa: An Automated Robust Fine-Tuning Framework
Robust Fine-Tuning (RFT) is a low-cost strategy to obtain adversarial robustness in downstream applications, without requiring a lot of computational resources and collecting significant amounts of data. This paper uncovers an issue with the existing RFT,
where optimizing both adversarial and natural objectives through the feature extractor (FE) yields significantly divergent gradient directions. This divergence introduces instability in the optimization process, thereby hindering the attainment of adversarial robustness and rendering RFT highly sensitive to hyperparameters. To mitigate this issue, we propose a low-rank (LoRa) branch that disentangles RFT into two distinct components: optimizing natural objectives via the LoRa branch and adversarial objectives via the FE. Besides, we introduce heuristic strategies for automating the scheduling of the learning rate and the scalars of loss terms. Extensive empirical evaluations demonstrate that our proposed automated RFT disentangled via the LoRa branch (AutoLoRa) achieves new state-of-the-art results across a range of downstream tasks. AutoLoRa holds significant practical utility, as it automatically converts a pre-trained FE into an adversarially robust model for downstream tasks without the need for searching hyperparameters. Our source code is available at [the GitHub](https://github.com/GodXuxilie/RobustSSL_Benchmark/tree/main/Finetuning_Methods/AutoLoRa).
****
Mengzhou Xia,Tianyu Gao,Zhiyuan Zeng,Danqi Chen
Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning
The popularity of LLaMA (Touvron et al., 2023a;b) and other recently emerged moderate-sized large language models (LLMs) highlights the potential of building smaller yet powerful LLMs. Regardless, the cost of training such models from scratch on trillions of tokens remains high. In this work, we study structured pruning as an effective means to develop smaller LLMs from pre-trained, larger models. Our approach employs two key techniques: (1) targeted structured pruning, which prunes a larger model to a specified target shape by removing layers, heads, intermediate and hidden dimensions in an end-to-end manner, and (2) dynamic batch loading, which dynamically updates the composition of sampled data in each training batch based on varying losses across different domains. We demonstrate the efficacy of our approach by presenting the Sheared-LLaMA series, pruning the LLaMA2-7B model down to 1.3B and 2.7B parameters. Sheared-LLaMA models outperform st

ate-of-the-art open-source models of equivalent sizes, such as Pythia, INCITE, and OpenLLaMA models, on a wide range of downstream and instruction tuning evaluations, while requiring less than 3% of compute compared to training such models from scratch. This work provides compelling evidence that leveraging existing LLMs with structured pruning is a far more cost-effective approach for building smaller LLMs.
****
Peiran Yu,Junyi Li,Heng Huang
Dropout Enhanced Bilevel Training
Bilevel optimization problems appear in many widely used machine learning tasks. Bilevel optimization models are sensitive to small changes, and bilevel training tasks typically involve limited datasets. Therefore, overfitting is a common challenge in bilevel training tasks. This paper considers the use of dropout to address this problem. We propose a bilevel optimization model that depends on the distribution of dropout masks. We investigate how the dropout rate affects the hypergradient of this model. We propose a dropout bilevel method to solve the dropout bilevel optimization model. Subsequently, we analyze the resulting dropout bilevel method from an optimization perspective. Analyzing the optimization properties of methods with dropout is essential because it provides convergence guarantees for methods using dropout. However, there has been limited investigation in this research direction. We provide the complexity of the resulting dropout bilevel method in terms of reaching an ϵ stationary point of the proposed stochastic bilevel model. Empirically, we demonstrate that overfitting occurs in data cleaning problems, and the method proposed in this work mitigates this issue.
****
Dujian Ding,Ankur Mallick,Chi Wang,Robert Sim,Subhabrata Mukherjee,Victor Rühle, Laks V. S. Lakshmanan,Ahmed Hassan Awadallah
Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing
Large language models (LLMs) excel in most NLP tasks but also require expensive cloud servers for deployment due to their size, while smaller models that can be deployed on lower cost (e.g., edge) devices, tend to lag behind in terms of response quality. Therefore in this work we propose a hybrid inference approach which combines their respective strengths to save cost and maintain quality. Our approach uses a router that assigns queries to the small or large model based on the predicted query difficulty and the desired quality level. The desired quality level can be tuned dynamically at test time to seamlessly trade quality for cost as per the scenario requirements. In experiments our approach allows us to make up to 40% fewer calls to the large model, with no drop in response quality.
****